М.В. Шилина, О.В. Мусатова,
В.В. Ивановский

БИОМЕТРИЯ

Учебно-методический комплекс

2011

УДК 573:001.8
ББК 28в631я73
Ш57

Авторы: доцент кафедры экологии и охраны природы УО «ВГУ им. П.М. Машерова», кандидат биологических наук **М.В. Шилина;** старший преподаватель кафедры экологии и охраны природы УО «ВГУ им. П.М. Машерова» **О.В. Мусатова;** доцент кафедры экологии и охраны природы УО «ВГУ им. П.М. Машерова», кандидат биологических наук **В.В. Ивановский**

Р е ц е н з е н т ы :
профессор кафедры зоологии УО «ВГУ им. П.М. Машерова»,
кандидат биологических наук *С.И. Денисова*; заведующий кафедрой физики УО «ВГТУ»,
доктор технических наук *В.В. Рубаник*

Научный редактор: заведующий кафедрой экологии и охраны природы УО «ВГУ им. П.М. Машерова», кандидат биологических наук, доцент **А.М.Дорофеев**

Учебно-методический комплекс подготовлен в соответствии с типовой учебной программой по курсу «Биометрия» для студентов, обучающихся по биологическим специальностям. Рассматриваются наиболее часто используемые биометрические методы для характеристики свойств эмпирических совокупностей, обработки массовых экспериментальных материалов. Описаны возможности пакетов анализа MS Excel и Statistica для решения задач прикладной статистики.

Учебное издание предназначено для студентов, обучающихся по биологическим специальностям университета, учителям биологии и экологии, а также лицам, ведущим исследования в различных областях биологии.

УДК 573:001.8
ББК 28в631я73

С О Д Е Р Ж А Н И Е

ВВЕДЕНИЕ

Системный подход при моделировании сложных явлений природы – одна из ведущих идей современного естествознания. В реализации этого подхода важное место занимают экспериментальные методы и методы многомерной статистики, которые позволяют исследователям выявить закономерности происходящих природных процессов, раскрыть причинно-следственные связи между элементами живой природы, сделать их доступными описанию точными математическими моделями.

При научных исследованиях биологических явлений наиболее эффективным является метод массовых наблюдений, использование которого предполагает проведение большого числа наблюдений. Собранный материал обрабатывают, анализируют, делают соответствующие выводы и устанавливают те или иные закономерности. Рассмотренный путь называется прямым индуктивным методом, когда от отдельных фактов переходят к общим положениям. Однако точность и достоверность результатов биологических экспериментов, равно как и корректность формулируемых выводов, зависят не только от качеств экспериментальных методик. Свойства самих биологических объектов и явлений сильно варьируют в пределах сообществ. Применение статистического анализа экспериментально полученных данных дополняет и углубляет познания биологических явлений природы, позволяет объективно оценить полученные результаты, нивелировав субъективизм исследователя и методические ошибки при постановке эксперимента, страхует экспериментатора от неточных и необоснованных выводов и заключений в отношении изучаемого явления.

Цель настоящего учебно-методического комплекса, описывающего стандартные и наиболее часто применяемые биометрические методы, – закрепить теоретические знания и приобрести опыт математической обработки данных наблюдения. Издание включает практические расчетные работы, в которых отрабатываются основные математические методы обработки массовых экспериментальных материалов, повторяются ключевые теоретические моменты отдельных разделов курса «Биометрия», а также задания для самостоятельной индивидуальной работы. В УМК дается оценка методам математической статистики с точки зрения их возможностей и границ применения. При выполнении конкретных видов работ студенты могут использовать в качестве матриц экспериментальных данных результаты собственных исследований.

Отдельный раздел учебного издания посвящен возможностям решения задач прикладной статистики средствами Microsoft Excel. В практических работах приводятся примеры применения стандартных функций и пакета анализа данных MS Excel для решения кон-

кретных задач с подробным описанием алгоритма решения. Приложение включает необходимые справочные данные для выполнения лабораторных работ.

Комплекс предназначен для оказания помощи студенту по работе с программой Statistica по проведению статистического анализа данных. В первую очередь они будут полезны студентам-дипломникам, работающим над своими дипломными работами и проектами. Пакет Statistica занимает в мире устойчиво лидирующее положение среди программ статистической обработки данных. В последнее время появились первые подробные пособия (см. список литературы), посвященные работе с этим и другими пакетами. Однако эта литература для массового пользователя не всегда является легкодоступной.

На простых примерах, касающихся различных количественных вариант, показаны возможности пакета по первичной обработке опытных данных и множественному регрессионному анализу. В издании рассматривают всего два статистических модуля из большого количества, имеющихся в программе.

При подготовке учебно-методического комплекса использован опыт других вузов, научная и методическая литература, основной список которой приводится.

Учебное издание подготовлено для студентов биологических специальностей вузов широкого профиля, может быть с успехом использовано студентами научной специальности 1-33 01 01 «Биоэкология» при подготовке курсовых и дипломных работ, а также учителями-предметниками профильных классов и классов с углубленным изучением дисциплин естественнонаучного цикла гимназий, лицеев при планировании учебной исследовательской работы школьников.

Лекция № 1

БИОЛОГИЧЕСКАЯ СТАТИСТИКА И БИОМЕТРИЯ. ЭТАПЫ СТАТИСТИЧЕСКОГО АНАЛИЗА

1. *Биометрия как наука. История формирования и развития биологической статистики.*
2. *Задачи биометрии. Методы и этапы статистического анализа.*
3. *Базовые структуры и понятия статистики.*
4. *Биологические признаки, их классификация.*

1. Знания о природе приобретаются путем наблюдения, сравнения и опыта. Под наблюдением в широком смысле понимается процесс планомерного изучения какого-нибудь явления независимо от того, как он осуществляется – на основе непосредственного восприятия или в эксперименте. При этом как объекты, так и результаты исследования могут быть и единичными, и массовыми (например, прием врачом пациента, содержание в неволе одной особи редкого вида или измерение веса детей одного возраста в школе, тестирование прочности стали разных марок). При научных исследованиях биологических явлений наиболее эффективным является метод массовых наблюдений. Собранный материал обрабатывают, анализируют, делают соответствующие выводы и устанавливают те или иные закономерности. Рассмотренный путь называется *прямым индуктивным методом*, когда от отдельных фактов переходят к общим положениям. Однако точность и достоверность результатов биологических экспериментов, равно как и корректность формулируемых выводов, зависят не только от качеств экспериментальных методик. Свойства самих биологических объектов и явлений сильно варьируют в пределах сообществ (например, течение одной и той же болезни у разных людей совершенно индивидуально, реакция на медицинские препараты и др.). Вместе с тем в реакции биологических объектов, в природных явлениях можно с той или иной степенью точности выявить определенные закономерности (например, предсказать время наступления и продолжительность сезона дождей и др.) В одних случаях такие прогнозы не требуют высокой степени точности. Однако в некоторых областях науки (медицина, биохимия и др.) вероятность ошибки должна быть сведена к минимуму. Решение этих задач возможно с использованием статистических методов. Применение статистического анализа экспериментально полученных данных дополняет и углубляет познание биологических явлений природы, позволяет объективно оценить полученные результаты, нивелировав субъективизм исследователя и методические

ошибки при постановке эксперимента, страхует экспериментатора от неточных и необоснованных выводов и заключений в отношении изучаемого явления.

В широком смысле ***статистика – наука, изучающая методы сбора и интерпретации числовых данных***. Это важная область практической деятельности людей, позволяющая не только собирать, обобщать и анализировать информацию, но и составлять прогнозы развития природных, социальных процессов и явлений, т.е. рассматривать их в динамике. Наполеон Бонапарт называл статистику бюджетом вещей.

Математическая статистика – наука сугубо теоретическая, абстрактная. Она изучает массовые явления безотносительно к специфике составляющих их элементов, например, закономерности числовых рядов. В приложении к биологии эту отрасль науки советский ученый в области общей биологии, генетики, биометрии и селекции животных П.Ф. Рокицкий в 1967 г. предложил называть ***биологической статистикой (или биометрией)*** – эмпирической, конкретной наукой, изучающей опытные данные.

Трактовка термина «биометрия» не всегда была однозначной. Дело в том, что использование математики в биологии началось не сразу. Биология очень долго развивалась на основе лишь качественного анализа событий. Измерение как один из ведущих методов познания природы начал применяться только с первой половины XVII века. В начале XVIII века Реомюр пытался найти математические законы строения пчелиных сот, а за 30 лет до него Борели делал математические расчеты движения животных. Всю совокупность методов измерения живых организмов и происходящих в них процессов с использованием математических подходов в биологии и обозначали понятием «биометрия» (от греч. bíos – жизнь и metréo – измеряю). Термин ввел Фрэнсис Гальтон (1822–1911), который стоял у истоков развития биометрии как науки. Обучаясь в Кембриджском университете, он увлекся естествознанием, метеорологией, антропологией, наследственностью и теорией эволюции. В своей книге, посвященной природной наследственности, изданной в 1889 году, он впервые и употребил слово «biometry». В это же время он разработал основы корреляционного анализа.

Однако в стройную научную дисциплину биометрию превратил математик Карл Пирсон (1857–1936). В 1884 году Пирсон получил кафедру прикладной математики в Лондонском университете, а в 1889 году познакомился с Гальтоном и его работами. Большую роль в жизни Пирсона сыграл зоолог Ф. Велдон. Помогая ему в анализе реальных зоологических данных, Пирсон ввел в 1893 г. понятия среднего квадратического отклонения и коэффициента вариации. Пытаясь математически оформить теорию наследственности Гальтона,

Пирсон разрабатывает основы множественной регрессии, теории сопряженности признаков.

Следующий этап развития биометрии связан с именем великого английского статистика Рональда Фишера (1890–1962). Во время обучения в Кембриджском университете Фишер знакомится с трудами Менделя и Пирсона. После окончания университета Фишер работал статистиком на одном из предприятий, преподавал физику и математику в средней школе, а в 1933 году был приглашен на должность профессора в Лондонский университет. Фишер заложил основы теории планирования эксперимента, предложил ряд эффективных статистических методов (в первую очередь, дисперсионный анализ), естественно вытекающих из своеобразия биологического эксперимента, развил теорию малых выборок, начатую английским ученым Стьюдентом (В. Госсетом). Его заслуга в том, что он впервые показал, что планирование экспериментов и наблюдений и обработка их результатов – две неразрывно связанные задачи статистического анализа. Значительную роль в распространении биометрических идей и методов сыграли русские ученые В.И. Романовский, А.А. Сапегин, Ю.А. Филипченко, С.С. Четвериков и др.

На современном этапе развития биометрии для решения экспериментальных задач совершенствуются методы многомерной статистики, позволяющие одновременно оценить не только влияние нескольких разных факторов, но и взаимодействие между ними. Широкое распространение получили и непараметрические методы, не содержащие предположений о характере распределения случайной величины, но уступающие по эффективности параметрическим методам. В связи с запросами практики интенсивно разрабатываются методы изучения наследуемости, выборочные методы и изучение динамических процессов (временные ряды).

Биометрия имеет некоторые особенности по сравнению с другими науками:

- Выявляет статистические закономерности, действующие только в сфере массовых явлений, не обнаруживающиеся на единичных случаях наблюдений.
- Не оперирует собственными методами, заимствуя последние из области математической статистики и теории вероятностей.
- Использует готовые математические выводы, применяя их к решению биологических задач.
- Оперирует языком формул и уравнений, которые являются своего рода моделями природных процессов и явлений, выраженных в экономичной форме.

2. ***Задачами биометрии (биостатистики)*** являются:

1. Анализ явлений (процессов), характеризующихся наличием в них стохастических (случайных) элементов.

2. Прогноз (предсказание по вероятности) осуществления случайных явлений и процессов.

Биометрия, как и любая область науки, четко структурирована. Комплексный анализ процессов и явлений осуществляется по определенной схеме с использованием наиболее оптимальных средств и методов в каждом конкретном случае. В самом общем виде при обработке результатов экспериментов и наблюдений возникают несколько основных статистических задач: оценка параметров распределения; сравнение параметров разных совокупностей; выявление статистических связей между явлениями, процессами; факторный анализ (оценка влияния факторов на исследуемый признак). Решение этих задач осуществляется с применением разных групп методов, которые выбираются в зависимости от изучаемого явления и конкретного предмета исследования (связи, закономерности или развития).

Выделяют три этапа работы со статистическими данными:

1) ***статистическое наблюдение*** – массовый научно организованный сбор первичной информации об отдельных единицах изучаемого явления. Результат – подготовка информационной базы для статистических обобщений, для формулирования выводов об изучаемом явлении или процессе;

2**)** ***группировка и сводка материала*** – обобщение данных наблюдения для получения оценочных показателей явления. Этап предполагает распределение множества фактов (единиц) на однородные группы и подгруппы, итоговый подсчет по каждой группе и подгруппе и оформление полученных итогов в виде статистической таблицы;

3) ***обработка статистических данных и анализ результатов*** для получения обоснованных выводов о состоянии изучаемого явления и закономерностях его развития. В процессе статистического анализа исследуются структура, динамика и взаимосвязь общественных явлений и процессов, формулируются предположения и гипотезы, осуществляется их статистическая проверка, формулирование выводов и прогнозов.

Все этапы статистического исследования тесно связаны друг с другом и одинаково важны. Недостатки и ошибки, возникающие на каждой стадии, сказываются на всем исследовании в целом. Поэтому правильное использование специальных методов статистической науки на каждом этапе позволяет получить достоверную информацию в результате статистического исследования.

Методы статистического исследования многообразны и динамичны. Очень часто для отдельных прикладных отраслей науки требуются специфические приемы анализа. Наиболее общими, широко применяемыми статистическими методами являются:

Рис. 1. **Структурно-логическая схема дисциплины.**

1) статистическое наблюдение;

2) сводка и группировка данных, построение статистических таблиц;

3) расчет обобщающих показателей (абсолютные, относительные и средние величины);

4) определение статистического распределения (вариационные ряды);

5) выборочный метод;

6) корреляционно-регрессионный анализ;

7) дисперсионный анализ и другие.

В соответствии с основными этапами статистического исследования мы будем придерживаться следующей структурно-логической схемы изучения дисциплины (рис. 1).

3. Для работы со статистическими методами анализа данных удобно пользоваться специальной терминологией, принятой в статистике. Раскроем сущность самых общих базовых понятий, а дополнительные рассмотрим при описании специальных методов.

Любое исследование (наблюдение, эксперимент) в статистике принято обозначать термином *опыт* – соблюдение условий и правил осуществления действия, при которых наблюдается соответствующее явление. Результат (реализация) опыта называют *событием*. Иными словами, событие – это исследуемый признак, параметр. Каждое событие происходит с определенной частотой, или вероятностью, – отношение числа опытов, в которых событие реализовалось, к полному числу произведенных опытов. Числовое (например, количество монет) или смысловое (например, «орел» или «решка») значение признака называют *случайной величиной (вариантой)* – переменной, принимающей значение события с определенной вероятностью. Всякое множество отдельных объектов, отличающихся друг от друга и в то же время сходных в некоторых отношениях, составляет так называемую *генеральную совокупность, или популяцию*. Скажем, популяция рыжих полевок того или иного района, заготовленные в крае беличьи шкурки, собранное зерно и т.п. Число единиц совокупности называют ее объемом и обозначают буквой N. Объем генеральной совокупности – приближающаяся к бесконечности величина. С ней очень проблематично иметь дело. На практике биологи работают с некоторой частью популяции, или совокупности. Например, группа цыплят, на которых ставится опыт по применению антибиотиков. Такая небольшая группа единиц генеральной совокупности называется *выборкой*, или выборочной совокупностью, ее объем обозначают буквой n. Число, вычисленное по выборке (например, среднее значение признака, самое большое или маленькое значение и др.), называют *статистикой*.

Число, вычисленное по генеральной совокупности (генеральное среднее, генеральная дисперсия и др.) – обозначают термином «*параметр*».

При составлении выборки каждый исследователь должен руководствоваться целым рядом условий. Подбор объектов зависит от условий опыта и конкретных биологических задач, хотя в каждом случае существует некоторая вероятность ошибки, связанная как с субъективизмом исследователя, так и несовершенством методик и инструментальных методов. В любом случае *выборка должна быть типичной и объективной*. Отбор элементов должен отвечать принципу рендомизации, т.е. быть случайным, что не означает беспорядочный, а такой, где устраняются субъективные влияния на состав выборочной совокупности.

В каждом конкретном случае при работе с выборкой у отдельных ее единиц мы анализируем конкретные признаки (события), принимающие определенные значения (варианты). Варианты даже в очень небольшой по размеру выборке могут отличаться между собой. Это различие между единицами совокупности называется *вариацией*, или *дисперсией* (т.е. рассеянием). Мы говорим, что «признак варьирует» т.е. он принимает различные значения у разных членов совокупности.

4. Способы обработки экспериментальных данных сильно зависят от того, каков характер различий между вариантами. В зависимости от этого вариация может быть *качественной и количественной*. Примером признаков с качественной вариацией могут служить окраска шерсти, запах вина, цвет лепестков и другие. Такие варианты учитываются чаще всего в альтернативной форме: есть–нет, окрашенный–бесцветный, с запахом или без. В некоторых случаях исследователи принимают в качестве эталонов определенные шкалы значения признака. Например, черная, рыжая, черно-рыжая и т.д. окраски.

В случаях, когда варианты можно охарактеризовать какими-либо числовыми величинами, вариацию называют количественной. Количественная вариация может быть двух типов: *дискретная (прерывная) и непрерывная*. В первом случае значения вариант описываются целыми числами, между которыми нет никаких переходов (1, 2, 5, 8). Например, число детенышей в помете, количество лепестков в цветке. Изучить определенный признак такой совокупности означает осуществить прямой подсчет изучаемых элементов у каждой единицы совокупности. При непрерывной вариации значения вариант не обязательно выражаются целыми, но и дробными числами. В зависимости от требуемой точности измерений, значение вариант может даваться с учетом десятых и даже сотых долей числа. А это означает, что в таких

случаях всегда присутствует непрерывная вариация, т.е. между каждыми целыми значениями возможны все переходы (2,2; 2,6; 2,7; 3,0). Примеры таких признаков – длина колоса злаков, вес рыб, жирность молока, относительная влажность воздуха.

В соответствии с характером вариации все биологические признаки делятся на *счетные* (дискретная вариация) и *мерные* (непрерывная вариация).

Лекция № 2

МЕТОДЫ ГРУППИРОВКИ ДАННЫХ. СТАТИСТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ

1. Статистическая совокупность и ее характеристики.
2. Способы группировки данных. Статистические таблицы.
3. Кривые распределения.
4. Вид статистического распределения.

1. Ранее мы уже говорили, что научное исследование предполагает работу со *статистической совокупностью* – множеством объектов, явлений, сходных по определенным признакам, но различающихся по ряду варьирующих характеристик, которые являются предметом статистического исследования. Эти объекты или явления представляют собой элементы (единицы) статистической совокупности. Так, например, статистической совокупностью будет население, элементами которой являются жители какой-либо страны в определенное время, что служит объединяющей их качественной основой. Однако жители различаются по социальному положению, по полу, возрасту, семейному положению, образованию и другим признакам. Наличие разносторонних и многообразных форм отношений и связей между объектами обусловливает возможность выделения ряда частных статистических совокупностей для одних и тех же объектов. Например, из общей совокупности студентов могут быть выделены частные совокупности сначала по одному (по специальности), затем по другому (по успеваемости) признаку и т.д.

Статистическая совокупность может быть качественно однородной, если наиболее существенный признак для всех ее элементов является общим, и разнородной, если в нее входят разные типы явлений. Совокупность, однородная в одном отношении, может быть разнородной в других. При проведении выборочного наблюдения различают генеральные (в которые входят все единицы статистической совокупности, подвергающейся исследованию) и выборочные совокупности.

2. Содержание понятия совокупности тесно связано с вопросом о классификации и группировках. Сущность метода группировок состоит в расчленении исследуемых единиц совокупностей (фактов, событий, явлений) на части (группы) по соответствующим характерным признакам, например, населения по месту проживания, особей в популяции по полу или возрасту.

Необходимость группировки связана с тем, что обычно полученные данные об объектах (варианты) представляют собой множество расположенных в беспорядке чисел. Просматривая это множество чисел, трудно выявить какую-либо закономерность их варьирования (изменения). Для выявления этих закономерностей следует сгруппировать полученные варианты, представив для удобства их в виде таблицы или ряда распределения. Упорядочивание данных помогает легче их обработать.

Группирование качественных признаков представляется наиболее простым: группы в данном случае представляют собой долю (в экземплярах, в %) особей данного качества от общего количества вариант. Например, % черных норок от их общего числа. Частным случаем качественной вариации является альтернативная, когда групп всего две (1; 0). Результатом группирования качественных признаков является *атрибутивный* ряд распределения (например, распределение по видам труда, полу, профессии, по религиозному признаку, национальной принадлежности и т.д.).

Для количественных признаков составляются статистические таблицы. Наиболее популярной формой таблиц являются вариационные ряды – упорядоченная по величине последовательность выборочных значений наблюдаемой случайной величины. Вариационные ряды, в которых варианты располагают в ряд в произвольном порядке, образуют *простой вариационный ряд*. После процедуры группирования можно получить три формы вариационного ряда: ранжированный ряд, дискретный ряд и интервальный ряд. Если варианты расположить последовательно в неубывающем порядке, вариационный ряд называют *ранжированным* (операция, при которой наблюдаемые значения случайной величины располагают в порядке неубывания, называется ранжированием опытных данных). Ранжирование позволяет легко разделить количественные данные по группам, сразу обнаружить наименьшее и наибольшее значения признака, выделить значения, которые чаще всего повторяются.

Дискретный ряд – это такой вариационный ряд, в основу построения которого положены признаки с прерывным изменением (дискретные признаки). Эти признаки могут принимать только конечное число определенных значений. Дискретный вариационный ряд представляет таблицу, которая состоит из двух граф. В первой графе указывается конкретное значение признака, а во второй – число единиц совокупности с определенным значением признака, или частоту

варианты (f_i, где i – индекс варианты). Сумма частот равна объему исследуемой совокупности: $\sum_{i=1}^{k} f_i = n$ (k – число вариантов значений признака). Очень часто таблица дополняется графой, в которой подсчитываются накопленные частоты S, показывающие, какое количество единиц совокупности имеет значение признака не большее, чем данное значение. В ряде случаев представляет практический интерес относительная частота того или иного варианта, называемая *частостью* (долей этой варианты) – отношение частоты к общей сумме частот всех вариантов (w_i), т.е. $w_i = f_i \big/ \sum f$. Это несложно сделать, если признак варьирует в незначительных пределах.

Для непрерывно варьирующих признаков при больших объемах выборок такие ряды получаются очень растянутыми, плохо обозримыми. Нецелесообразно также построение дискретного ряда для дискретной случайной величины, число возможных значений которой велико. В подобных случаях нельзя обойтись без разбивки ряда на некоторое количество классов с одинаковым интервалом, включающих сразу несколько значений вариант. Например, если размеры колоса у злаков варьируют от 1 до 30, то всю совокупность вариант можно представить в виде 5 классов: от 1 до 6 – I класс, от 7 до 12 – II класс и т.д. Полученная статистическая таблица носит название *интервального вариационного ряда*.

Интервал – это количественное значение, отделяющее одну единицу совокупности от другой. Число интервалов можно определить через логарифм численности выборки по формуле Стэрджеса, округлив полученное значение до целого числа: $k = 1 + 3{,}322 \lg n$, по формуле Брукса и Краузерс: $k = 5 \cdot \lg n$, где k – число интервалов вариационного ряда.

Некоторые исследователи предлагают использовать для этих целей другую формулу: $k = 1{,}441 \ln(n) + 1$.

Величина классового интервала обозначается через *i*; она определяется по разности между значениями максимальной и минимальной вариант, отнесенной к избранному числу классов, т.е. по следующей приближенной формуле: $i = \frac{x_{\max} - x_{\min}}{k}$, где

i – классовый интервал;

x_{max} – максимальная и x_{min} – минимальная варианты выборки;

k – число классов, на которые разбивается выборочная совокупность.

При установлении пределов интервалов к минимальному значению изучаемого признака последовательно прибавляют принятую величину интервала до тех пор, пока не достигнута (или превышена) максимальная величина признака.

3. Для удобства вариационные ряды изображают графически в системе координат – график функции эмпирического распределения признака. Графическое изображение вариационного ряда в общем виде получило название *вариационной кривой*, или *кривой распределения*. Существуют два способа изображения кривой:

1) ***полигон распределения*** – при дискретной вариации, когда каждый класс включает лишь одну варианту. На оси абсцисс наносятся значения вариант, на оси ординат – их частоты (рис. 2).

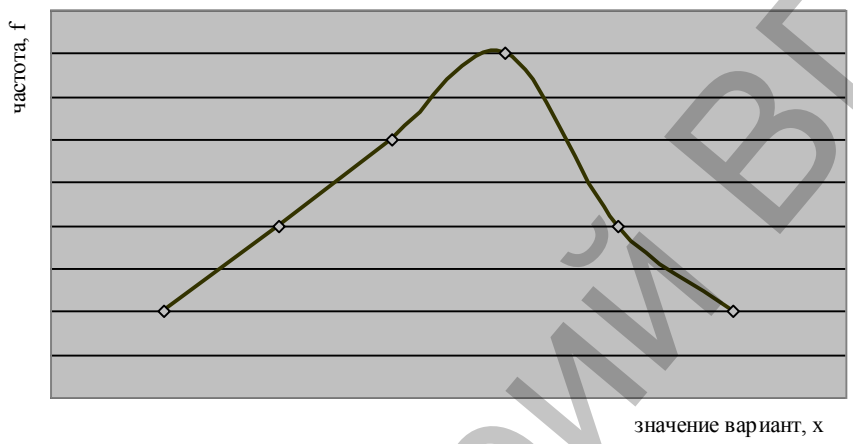


Рис. 2. **Полигон статистического распределения.**

2) ***гистограмма распределения*** – при непрерывной вариации, когда классу соответствует несколько значений вариант, для интервального вариационного ряда. Эмпирическая функция распределения отображается ступенчатой ломаной линией: над каждым интервалом проводится отрезок горизонтальной линии на высоте, пропорциональной накопленной частости в текущем интервале. Частоты вариант выражают столбиками, основания которых соответствуют всем значениям класса (рис. 3).

частоты

Рис. 3. **Частоты вариант.**

Гистограмму можно дополнить полигоном – ломаной линией, отрезки которой соединяют точки с координатами по оси абсцисс, равными серединам интервалов, а по оси ординат – соответствующим частостям.

Существуют и другие виды кривых (*огива, кумулята*), использование которых продиктовано специальными задачами исследования.

Всем вариационным рядам в той или иной мере присущи некоторые общие закономерности, например: большинство вариант располагается в средней части вариационного ряда, на кривой это соответствует пику или вершине; распределение вариант от этого максимума в обе стороны более или менее симметрично; частота вариант постепенно убывает к краям вариационного ряда. Если в вариационном ряду крайние значения имеют наименьшую, а средние – наибольшую частоты, то распределение признака называется *нормальным*.

4. Важным способом «описания» переменной является форма ее распределения. ***Вид статистического распределения*** – определенный тип соответствия между наблюдаемыми значениями измеряемого показателя и их частотами (относительными частотами). Вид статистического распределения может быть произвольным. На практике чаще всего встречаются следующие три вида статистических распределений:

1. Равномерное распределение.
2. Показательное распределение.
3. Нормальное распределение.

О виде статистического распределения можно судить по полигону или гистограмме распределения, которые в каждом случае обладают характерной «внешностью». Гистограмма позволяет «на глаз» оценить нормальность эмпирического распределения. На гистограмму также накладывается кривая нормального распределения. Гистограмма позволяет качественно оценить различные характеристики распределения. Например, на ней можно увидеть, что распределение *бимодально* (кривая имеет 2 пика). Это может быть вызвано, например, тем, что выборка неоднородна, возможно, извлечена из двух разных популяций, каждая из которых более или менее нормальна. В таких ситуациях, чтобы понять природу наблюдаемых переменных, можно попытаться найти качественный способ разделения выборки на две части. Однако окончательное решение о виде статистического распределения можно сделать лишь с помощью статистических критериев.

Каждый из видов распределения подчиняется определенным закономерностям, служит математической моделью эмпирической закономерности. Определение вида статистического распределения имеет большое значение для принятия решения о методах дальнейшего ста-

тистического анализа, так как применение целого ряда статистических критериев требует, чтобы распределение изучаемого показателя было нормальным, и если нормальность распределения не доказана, их применение неправомерно.

Исследование закономерностей распределения признака чаще проводится по выборочной совокупности. Если выборка для проведения исследования является репрезентативной, то полученные данные о статистическом распределении признака дают достаточно точное представление о распределении признака в генеральной совокупности. Почему это важно? Дело в том, что если вы знаете распределение наблюдаемой переменной, то можете предсказать, как в повторных выборках равного объема будет «вести себя» используемая статистика – т.е. каким образом она будет распределена. Следует иметь в виду, что увеличение объема выборки всегда приводит к повышению ее репрезентативности, таким образом, чем больше объем выборки, тем точнее выборочное распределение отражает действительное состояние дел.

Закон нормального распределения лежит в основе многих теорем и методов статистики при:

1) оценке репрезентативности выборки (расчете ошибки выборки и распространении характеристик выборки на генеральную совокупность);

2) измерении степени тесноты связи между признаками и составлении модели регрессии;

3) построении и использовании статистических критериев и др.

Как показывают многочисленные статистические исследования, частоты (частости) эмпирических распределений за редким исключением будут отличаться от значений теоретического распределения. Расхождения между частотами (частостями) эмпирического и теоретического распределения могут быть несущественными и объяснены случайностями выборки и существенными при несоответствии выбранного и эмпирического законов распределения. Для проверки гипотезы о соответствии эмпирического распределения теоретическому закону нормального распределения используются особые статистические показатели – критерии согласия (или критерии соответствия), которые мы рассмотрим позже.

Почему важно нормальное распределение? Нормальное распределение важно по многим причинам. Многие статистики нормально распределены, либо их распределение связано с нормальным и вычисляется на его основе, как, например, t, F или χ^2-распределение. Рассуждая философски, можно сказать, что нормальное распределение представляет собой одну из эмпирически проверенных истин относительно общей природы действительности и его положение может рассматриваться как один из фундаментальных законов природы.

Проблема может возникнуть, когда пытаются применить тесты, основанные на предположении нормальности, к данным, не являющимся нормальными. В этих случаях можно выбрать одно из двух. Во-первых, можно использовать альтернативные «непараметрические» тесты (так называемые «свободно распределенные критерии»). Однако это часто неудобно, потому что обычно эти критерии имеют меньшую мощность и обладают меньшей гибкостью. Как альтернативу, во многих случаях можно все же использовать тесты, основанные на предположении нормальности, если есть уверенность, что объем выборки достаточно велик. Последняя возможность основана на чрезвычайно важном принципе, позволяющем понять популярность тестов, основанных на нормальности. А именно, при возрастании объема выборки, форма выборочного распределения (т.е. распределение выборочной статистики критерия) приближается к нормальной, даже если распределение исследуемых переменных не является таковым. Отметим, что при размере выборки n=30, выборочное распределение «почти» нормально. Этот принцип называется центральной предельной теоремой.

Следует заметить, что большинство природных явлений и процессов распределены нормально, поэтому этому типу распределения мы уделим большое внимание.

Лекция № 3

СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ ДЛЯ ХАРАКТЕРИСТИКИ СОВОКУПНОСТИ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

1. Задачи и методы описательной статистики.
2. Показатели центральной тенденции.
3. Меры разброса данных.

1. На практике часто встречаются задачи, когда использование законов распределения вероятностей не всегда возможно (например, при малых размерах выборки) или не всегда необходимо. В этих случаях используются более простые, хотя и менее информативные характеристики, которые принято называть *числовыми характеристиками выборки*, или *описательными статистиками*. С помощью небольшого набора цифр можно сжато и лаконично отразить наиболее существенные черты распределения определенного вариационного ряда или сравнить разные ряды. ***Основная задача*** описательной ста-

тистики – по статистическому распределению выборки оценить неизвестный показатель генеральной совокупности.

Важнейшими описательными статистиками являются показатели центра распределения и меры разброса данных в выборке. К первым относятся различные средние величины: мода, медиана, средняя арифметическая, средняя взвешенная, геометрическая и др. Ко второй – вариационный размах, среднее абсолютное отклонение, среднее квадратичное отклонение, дисперсия, коэффициенты асимметрии и вариации.

2. Наиболее простой и распространенный способ охарактеризовать некоторую совокупность объектов состоит в том, чтобы описать ее «средние» характеристики. Существует несколько подходов к определению «среднего»:

1. ***Мода*** *(частотное среднее)*. Анализируя любую вариационную кривую, нетрудно заметить, что какая-либо варианта обладает наибольшей частотой, то есть в вариационном ряду встречается большее число раз, чем другие варианты. Такую варианту называют *модой* (M_o). Можно сказать, что статистическое значение термина «мода» близко к бытовому – модальным (модным) является то, что думает, предпочитает или отвечает на вопросы анкеты большинство. Именно в смысле модальности говорят о «среднем» жителе страны или студенте. Например, модальным полом на филологическом факультете университета является женский, на физкультурном – мужской. Для непрерывной переменной говорят о модальном интервале или классе, которому соответствует максимальная частота. Значение моды в данном случае определяется как среднее значение класса.

Распределение может иметь более одной моды. Распределение с двумя модами называется *бимодальным*, с тремя и более – *полимодальным*. Определить би- или полимодальность проще всего на кривой распределения, где четче видны несколько выраженных «пиков». Мода является единственной универсальной мерой центральной тенденции; ее можно использовать для переменных всех типов – номинального, порядкового и количественного, дискретного и непрерывного.

В дискретном ряду мода – варианта с наибольшей частотой, в интервальном ряду определяется по формуле:

$$M_d = x_{md} + i_{md} \times \frac{f_{md} - f_{md-1}}{(f_{md} - f_{md-1}) \cdot (f_{md} - f_{md+1})},$$

где

x_{md} – начало модального интервала;

i_{md} – ширина модального интервала;

f_{md} – частота модального интервала;

f_{md-1}, f_{md+1} – частоты пред- и постмодального интервалов соответственно.

Например, имеется интервальный ряд:

x	500–1000	1000–1500	1500–2000	2000–2500	2500–3000	
f	5	10	15	14	6	Σf=50

$$Md = 1500 + 500 \times \frac{15-10}{(15-10)+(15-14)} = 1917.$$

2. ***Медиана*** *(граница 50%-ного интервала).* Медианой (*Me*) называется значение, которое делит выборку пополам: в обе стороны от медианы в вариационном ряду располагается одинаковое число вариант. Медиана определяется только для порядковых и количественных переменных; к номинальным переменным ее применять нельзя. Для того чтобы найти медиану, необходимо упорядочить значения переменной, учитывая каждое значение столько раз, сколько раз оно встречается в выборке, а затем определить середину полученного ряда.

В дискретном ряду медиана определяется по способу накопления частот. Частоты накапливают до тех пор, пока не будет превышена половина суммы частот. То значение *x*, которое соответствует наибольшей последней накопленной частоте, и есть медиана.

Например,

x	100	200	300	400	500	
f	2	4	8	5	6	Σf=25
Накопленные f	2	2+4=6	6+8=14			

Me=300.

В интервальном ряду вначале определяется медианный интервал по способу накопления частот. Затем медиану определяют по формуле:

$$Me = x_{me} + i_{me} \frac{0{,}5\sum f - \int f_{me-1}}{f_{me}},$$

где

$x_{ме}$ – начало Ме-интервала;

$i_{ме}$ – величина Ме-интервала;

$f_{ме}$, $\int f_{me-1}$ – частота медиального и сумма накопленных частот до предмедиального интервалов соответственно.

Например,

x	100–200	200–300	300–400	400–500	500–600	600–700	700–800	
f	1	3	7	30	19	15	5	Σf=80
Накопленные f	1	4	11	41				

$$Me = 400 + 100 \cdot \frac{40 - 11}{30} = 496{,}76.$$

Мода и медиана используются в настоящее время в биологии довольно ограниченно, характеризуют своего рода типичное в данной совокупности.

3. **Среднее арифметическое** (М) является наиболее часто используемым показателем центра распределения для количественных переменных. Часто говорят «М среднее», подразумевая под этим среднее арифметическое. Среднее арифметическое показывает, каким было бы значение переменной, если бы у всех объектов из выборки оно было бы одинаковым. Среднее арифметическое представляет собой частное от деления суммы всех вариант совокупности на их число, т.е. $M = \frac{x_1 + x_2 + x_3 + \ldots . x_k}{n} = \frac{\Sigma x}{n} = \frac{1}{n}\Sigma x$. Это и есть общая формула средней арифметической, где x_1 x_2, x_3, ..., x_k обозначают варианты, входящие в состав данной совокупности; n – общее число вариант, или объем выборочной совокупности.

При повторяемости отдельных вариант среднюю арифметическую можно представить как сумму произведений отдельных вариант на их частоты, отнесенную к общему числу всех вариант данной совокупности, т.е. как $M = \frac{\Sigma x_i f_i}{\sum f} = \frac{1}{f}\Sigma x_i f_i$. Такая средняя называется *взвешенной* (т.е. каждая варианта «взвешивается» относительно ее доли в общей совокупности). Средняя – число именованное, которая выражается теми же единицами меры или счета, что и характеризуемый ею признак, вычисляется для количественных переменных, оценочных порядковых переменных.

4. Показатели центра распределения характеризуют выборочную совокупность в целом: некоторые общие тенденции, свойственные, в большей или меньшей степени, всем объектам из выборки. Однако объекты в выборке отличаются друг от друга, и часто весьма значительно. При одном и том же значении центра распределения социальные группы могут значительно отличаться друг от друга, т.е. не всегда средние показатели группы эффективно ее характеризуют. Например, сравнивая две группы студентов, имеющих одинаковый средний балл успеваемости, можно было бы предположить, что в качественном отношении это равноценные группы. Но среднее арифметическое успеваемости будет одинаковым и при сильном и при слабом разбросе оценок студентов. Или, например, женский пол будет модальным и в группе из 25 человек, где 13 девушек, и в группе, где все

25 человек девушки. Приведенные примеры указывают на то, что показатели центра распределения ничего не говорят о мере разброса признака, т.е. о степени его вариации. Показателями вариации являются размах вариации, среднее квадратическое отклонение, дисперсия, коэффициенты асимметрии и вариации.

1. *Размах вариации*. Диапазон разброса значений является простейшей мерой степени разброса данных. Лимиты показывают фактические пределы варьирования признака. Разность между лимитами (предельными значениями признака) и является размахом вариации: $d = x_{\max} - x_{\min}$. Размах вариации позволяет судить о выборках с равной средней. Например, если лимиты одной выборки равны 4 и 16, а другой – 3 и 19, то размах вариации в первом случае равен 12, во втором – 16. Отсюда следует, что вариабельность первого признака меньше, чем второго. Вместе с тем этот показатель имеет серьезные недостатки. Прежде всего, он существенно зависит от случайных факторов, которые могут проявляться в появлении «резко выделяющихся» значений (например, очень высокое число потомков у одной из особей по сравнению с остальными, альбиносы среди окрашенных особей и т.п.). Другой недостаток состоит в том, что размах вариации никак не связан с формой распределения.

2. Наиболее подходящей мерой варьирования является *среднее квадратичное отклонение* (стандартное отклонение – s), которое показывает, насколько индивидуальные значения переменной в среднем отклоняются от среднего арифметического:

3. $$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2}{n-1}},$$

где

x_i – значение переменной для объекта с номером i;

M – среднее арифметическое;

n – объем выборки.

Величина n–1 носит название *числа степеней свободы*, под которым подразумевается число свободно варьирующих членов совокупности. Чтобы это понять, представим некую совокупность из трех вариант, сумма которых должна составлять 30. Очевидно, что первые две варианты могут принимать любое значение, тогда как величина третьей будет зависеть от первых двух и может принимать только одно значение, равное разнице между 30 и суммой первых двух вариант. Можно сказать, что первые две варианты свободно и независимо варьируют, а третья не имеет степени свободы, следовательно, в совокупности имеются только 2 степени свободы из 3 (3–1=2). Так, определяя среднее значение показателя по выборке, мы его как бы фикси-

руем и используем в дальнейших расчетах, тем самым уменьшаем число степеней свободы на единицу. Число степеней свободы уменьшается на единицу при фиксации каждого нового параметра выборки. В общем виде при численности совокупности *n* число степеней свободы *df = n-1*. При большом размере выборки разница между n и n-1 невелика и мало отразится на значении s. Но при малых выборках (n<30) введение n-1 лучше оценивает значение стандартного отклонения для генеральной совокупности.

4. В некоторых задачах вместо среднего квадратичного отклонения *(s)* используется параметр сигма *(σ)*, вместо *вариансы(s^2) – дисперсия (σ^2). Варианса* представляет собой квадрат среднего квадратичного отклонения и обозначается s^2. Основная формула для вычисления дисперсии и вариансы: $s^2 = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n-1}$. Дисперсия и СКО полностью определяют друг друга и используются с одинаковым успехом, имеют ту же размерность, что и среднее арифметическое признака. Дисперсия (σ^2) – показатель вариации, а среднее квадратическое отклонение (s^2) – выборочный показатель.

Не все биологические признаки распределяются строго по нормальному закону, когда M=*Mo*=*Me*. Обычно исследователя интересует, насколько точно распределение можно аппроксимировать нормальным. Простые описательные статистики дают об этом некоторую информацию.

Некоторые вариационные ряды обнаруживают асимметрию, т.е. скошенность в ту или иную сторону от центрального момента. Различают правостороннюю (отрицательную) и левостороннюю (положительную) асимметрию. В случаях правосторонней, или отрицательной, асимметрии варианты накапливаются преимущественно в правой части ряда; вершина такого ряда сдвинута вправо. В случае левосторонней асимметрии правая ветвь кривой, начиная от вершины, больше левой.

Пирсон предложил оценивать степень асимметрии по разности между средней арифметической и модой, отнесенной к величине среднего квадратического отклонения: $A_s = \frac{M - Md}{s}$, где A_s – мера скошенности рядов распределения, или коэффициент асимметрии. Чаще, однако, используют формулу $A_s = \frac{\Sigma p(M - x_i)^3}{ns^3}$. Понятно, что при симметричном (нормальном) распределении числитель равен нулю, а значит и коэффициент асимметрии тоже. Если асимметрия существенно отличается от 0, то распределение несимметрично. При асимметрии этот показатель будет иметь либо положительное, либо от-

рицательное значение. Мера косости меньше 0,5 считается малой, от 0,5 до 1 – средней, выше 1 – большой.

Еще одной статистикой, характеризующей форму распределения, является *эксцесс*, показывающий «остроту пика» кривой распределения. Оценить меру эксцесса можно по формуле: $Ex = \frac{\Sigma p(M - x_i)^4}{ns^4}$. Если эксцесс существенно отличен от 0, то распределение имеет или более закругленный пик, чем нормальное, или, напротив, имеет более острый пик (возможно, имеется несколько пиков). Обычно, если эксцесс положителен, то пик заострен, если отрицательный, то пик закруглен. Эксцесс нормального распределения равен 0.

5. *Коэффициент изменчивости, или коэффициент вариации*. Часто на практике приходится сравнивать различные статистические совокупности, признаки которых имеют разную размерность (например, кг и м). В этом случае использование абсолютных показателей (средней, дисперсии и т.п.) очень неудобно. Для сравнения вариации различных признаков, а также степени изменчивости групп организмов разных видов необходим какой-то относительный показатель. Более информативным и удобным при сравнении различных статистических совокупностей является *коэффициент изменчивости*, или *вариации*, так как его величина не зависит от единиц, используемых при измерениях: $C = \frac{S}{M} \times 100\%$.

На основании величины коэффициента изменчивости можно судить о характере и степени варьирования признака (таблица 1).

Таблица 1

Характер изменчивости признаков (по М.Л. Дворецкому, 1971)

Коэффициент изменчивости, С	до 5%	6–10%	11–20%	21–50%	более 50%
Характер изменчивости	слабая	умеренная	значительная	большая	очень большая

Лекция № 4

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И ЕГО ХАРАКТЕРИСТИКА

1. Вероятность и ее свойства.
2. Нормальное распределение и его характеристика. Нормированное отклонение.
3. Уровни значимости и доверительные вероятности. Доверительный интервал.

1. Биологу в отношении каждого происходящего события нельзя с уверенностью предсказать его результат, можно лишь говорить о некоторой возможности значения, которое он приобретает. Скажем, трудно с уверенностью сказать, какое число девочек или мальчиков родится за 1 день в роддоме. Можно лишь предполагать, зная результаты предыдущих дней, частоту рождения детей разных полов. Такое теоретическое значение относительной частоты ожидаемого события называется его *вероятностью*. Иначе говоря, вероятность – это степень уверенности в том, что событие произойдет.

Исходным в понятии вероятности является понятие равновозможности. Если при каждом испытании событие неизбежно наступает, оно называется *достоверным*. Если же в заданных условиях событие произойти не может, его называют *невозможным*. Когда же событие в каждом отдельном испытании может произойти, но может и не произойти, его называют *возможным*, или *случайным*. Согласно классическому определению, ***вероятностью*** называется:

отношение числа благоприятных случаев к числу всех равновозможных случаев: $p = m/n$.

Вероятность, которую можно указать до опыта, называется *априорной*. Например, при метании монеты заранее известно, что она может лечь либо «орлом», либо «решкой» вверх. Здесь только две возможности и вероятность каждой равна 50%, или ½. Другое дело вероятности, которые заранее предугадать невозможно. Например, действие на организм различных лекарственных веществ. Вероятность событий для этого случая может быть установлена только путем многократных испытаний, т.е. после опыта. Такие вероятности называются *эмпирическими*, или *апостериорными*.

Из формулы видно, что вероятность любого события есть число, заключенное между нулем и единицей, т.е. она выражается в долях от единицы. По мере приближения p к единице событие становится все более достоверным. Если p=1, то событие бесспорно наступит. И на-

оборот, то, что обладает малой вероятностью, мало достоверно. В жизни мы всегда считаемся, сознательно или бессознательно, с эмпирическими вероятностями и действуем согласно этим оценкам. При этом мы, как правило, не придаем значения явлениям, обладающим малой вероятностью.

С другой стороны, планируя опыт или испытание, исследователи заранее предполагают получение результата с достаточно высокой вероятностью (или уровнем значимости). После серии проведенных испытаний рассчитывают эмпирическую вероятность события, сравнивают ее с запланированной. После этого решают, достоверен ли результат опыта, можно ли результат использовать для практических целей. Чуть позже мы расшифруем понятие уровня вероятности, или уровня значимости для целей биологической статистики.

2. Мы уже упоминали, что большинство биологических явлений имеют нормальное распределение, признаком которого является накопление большей части вариант в средине вариационного ряда. Кривая распределения носит название нормальной кривой, она характеризуется абсолютной симметричностью и совпадением центральных моментов: $M=M_o=M_e$. Для изучения закономерностей вариации при нормальном распределении пользуются так называемым нормированным отклонением (t). Нормированное отклонение представляет собой отклонение той или иной варианты от среднего арифметического выборки, выраженное в сигмах: $t=\frac{x_i-M}{s}$. Каждая варианта характеризуется определенным *t*, указывающим ее положение в вариационном ряду или на кривой распределения. Например, если варианта имеет значение +1,5, это значит, что она расположена справа от среднего арифметического на расстоянии в 1,5s. Размещение вариант в вариационном ряду характеризуется определенными закономерностями. Дело в том, что в нормальной кривой отклонения от средней арифметической практически охватывают 6s или 6σ: 3 вправо от средней и 3 влево. Математики, исследовав уравнение функции нормального распределения, вычислили вероятности отклонения вариант в соответствии с правилом «трех сигм». Так, 68,3% всех вариант уклоняются от среднего арифметического не более чем на ±1s, 95,5% – не более чем на ±2s (другими словами, при нормальном распределении, нормированные отклонения, меньшие -2 или большие +2, имеют относительную частоту менее 5%), и 99,7% – не более чем на ±3s.

В биологических исследованиях трудно ожидать, что вероятность реализации события будет равна 1, вместе с тем, подтверждение выдвинутой гипотезы требует достаточно высокой степени надежности экспериментальных данных, а значит, и высокой вероятности их

повторения. В биологии приняты несколько уровней достоверных вероятностей, их обычно выражают величинами 95%, 99%, 99,9%. Вероятность 95% – самый низкий уровень достоверности, принятый в биологических исследованиях. Согласно закону нормального распределения он предполагает, что в 95% случаев выборочная средняя не отклонится от средней генеральной совокупности более чем на 2t (1,96s). И только в 5% случаев, считая положительные и отрицательные уклонения, выйдет за эти границы. Это значит, что вероятность ошибки (отличия выборочной средней от генеральной) составляет лишь 0,05 (в обе стороны, в одну – 0,025). С вероятностью же 99% она будет отклоняться от средней не более чем на 2,58s. Вероятность выхода за пределы 2,58s еще меньше – 0,01: 3,29s (99,9%). Это важное правило часто называют *правилом трех сигм*. Три сигмы как бы ограничивают пределы случайного рассеяния внутри вариационного ряда. То, что находится внутри трех сигм, относится к этому ряду; то, что за пределами, вероятнее всего, к этому ряду не принадлежит. Такие варианты называют выпадающими и при анализе их игнорируют.

3. Вероятности (P) 0,95, 0,99 и 0,999 называют *доверительными вероятностями*, они означают долю случаев, когда гипотеза заслуживает доверия. Как уже было сказано, каждый уровень доверительной вероятности характеризуется определенной величиной нормированного отклонения:

P=0,95 t=1,96

P=0,99 t=2,58

P=0,990 t=3,29.

Величина доверительной вероятности при проверке гипотез устанавливается самим исследователем в зависимости от степени точности, которая его удовлетворяет.

Определенным доверительным вероятностям соответствуют так называемые ***уровни значимости*** *(или статистическая значимость результата,* ***p-уровень)***. Статистическая значимость результата представляет собой оцененную меру уверенности в его «истинности» (в смысле «репрезентативности выборки»). Фактически под ними понимают тот процент случаев, когда не подтверждается гипотеза равенства выборочной и генеральной средней, тот процент риска, при котором возможна ошибка в выводах исследователя. Для вероятности 0,95 уровень значимости равен соответственно 0,05, для вероятности 0,99 – 0,01 и для вероятности 0,999 – 0,001. Например, 5%-ный уровень значимости означает, что в 5% случаев возможна случайная ошибка.

Этот показатель, находится в убывающей зависимости от надежности результата. Более высокий p-уровень соответствует более

низкому уровню доверия к найденной в выборке зависимости между переменными. Именно p-уровень представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю популяцию.

Как определить, является ли результат действительно значимым? Не существует никакого способа избежать произвола при принятии решения о том, какой уровень значимости следует действительно считать «значимым». Выбор определенного уровня значимости, выше которого результаты отвергаются как ложные, является достаточно произвольным. На практике окончательное решение обычно зависит от того, был ли результат предсказан априори (т.е. до проведения опыта) или обнаружен апостериорно в результате многих анализов и сравнений, выполненных с множеством данных, а также на традиции, имеющейся в данной области исследований. Обычно во многих областях результат $p \leq 0{,}05$ является приемлемой границей статистической значимости, однако следует помнить, что этот уровень все еще включает довольно большую вероятность ошибки (5%). Результаты, значимые на уровне $p \leq 0{,}01$, обычно рассматриваются как статистически значимые, а результаты с уровнем $p \leq 0{,}005$ или $p \leq 0{,}001$ как высоко значимые. Однако следует понимать, что данная классификация уровней значимости достаточно произвольна и является всего лишь неформальным соглашением, принятым на основе практического опыта в той или иной области исследования. Чаще всего биологов удовлетворяет уровень значимости $p \leq 0{,}05$, однако нередко используется и более высокая статистическая значимость (1%, 0,1%).

Границы, в которых с той или иной вероятностью находится параметр генеральной совокупности, и интервал между ними называются доверительными (рис. 4).

Доверительные интервалы для среднего, например, задают область вокруг среднего, в которой с заданным уровнем доверия содержится «истинное» среднее популяции. Установить генеральное среднее достаточно сложно, можно лишь определить по выборочной совокупности интервал, в пределах которого лежит этот параметр:

$$\mu = M \pm \Delta x.$$

$$\Delta x = m_M \cdot t(p, f),$$

где m_M – ошибка в определении среднего арифметического значения выборочной совокупности, t(p,f) – критическое значение критерия значимости, где p – уровень значимости (обычно принимается значение 95%) и f – число степеней свободы.

$$m_M = s/\sqrt{n}, \text{ т.е. } \mu = M \pm \frac{m_M \cdot t(p, f)}{\sqrt{n}}.$$

Величина $\frac{m_M \cdot t(p, f)}{\sqrt{n}}$ – это и есть доверительный интервал, который используется для оценки воспроизводимости.

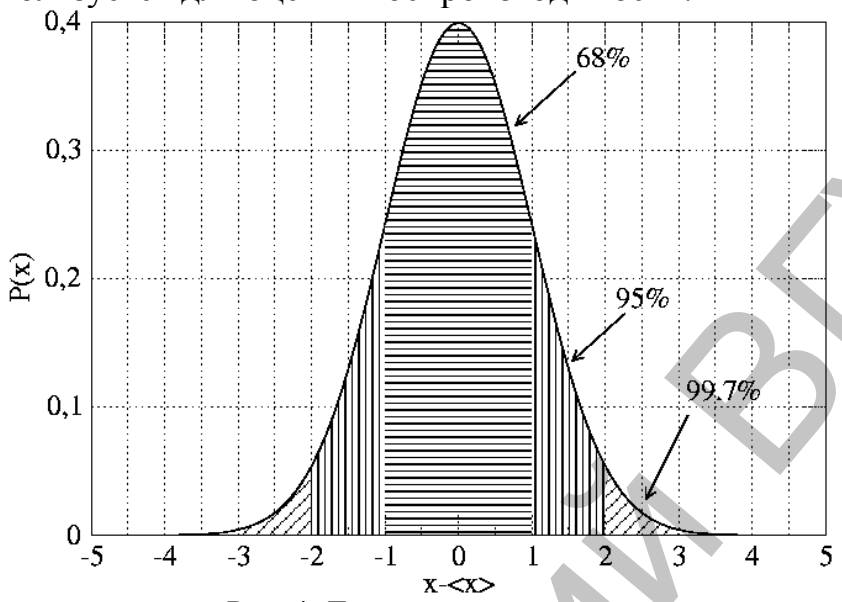


Рис. 4. **Доверительный интервал.**

Можно построить доверительные интервалы для любого p-уровня и любого показателя.

4. При работе с выборочными совокупностями всегда следует учитывать возможности отклонения показателей выборки от параметров генеральной совокупности. Эти отклонения называются *ошибками репрезентативности* и бывают двух видов: случайные и систематические. Первые не зависят от исследователя, а связаны с тем, что любая выборка не может точно представлять совокупность; вторые возникают при несоблюдении правил отбора или методики эксперимента. Они могут быть устранены при выполнении соответствующих условий. Случайные же ошибки всегда остаются и должны учитываться при оценке генеральной совокупности по данным выборки. Ошибки репрезентативности могут возникать при оценке любого параметра. Рассмотрим ошибки основных описательных статистик.

1. *Ошибка отдельно варианты* показывает, насколько индивидуальные значения переменной в среднем отклоняются от среднего арифметического, т.е. ошибка отдельной варианты есть не что иное, как среднее квадратичное отклонение. Ошибки обозначаются буквой m. Для данного случая $m_x = \pm s$.

2. *Ошибка средней (m_M)* является величиной, на которую отличается среднее значение выборочной (опытной) совокупности от среднего значения генеральной совокупности при условии, что распределение изучаемого признака приближается к нормальному. Ос-

новная ошибка среднего рассчитывается по формуле $m_M = \frac{s}{\sqrt{n-1}}$.

Среднее значение необходимо записать с основной ошибкой ($M \pm m_M$), только в этом случае можно судить о точности опыта.

3. *Ошибка среднего квадратичного отклонения* вычисляется по формуле $m_s = \pm s/\sqrt{2n}$.

4. *Ошибка коэффициента вариации:* $m_c = \pm C/\sqrt{2n}$.

После вычисления того или иного статистического показателя необходимо проверить степень его надежности (достоверности) путем деления величины этого показателя на его ошибку. Достоверность среднего значения определяется по формуле $t = \frac{M}{m_м}$. Если значение $t > 4$, то среднее значение показателя является достоверным. Таким показателем можно пользоваться для сопоставления и формулировки корректных выводов. Часто о достоверности показателя судят, если $t \geq 3$. Если же t меньше трех, то по таким показателям нельзя делать категорические заключения или проводить сопоставления.

Важным показателем, характеризующим процент расхождения между выборочной и генеральной средними является *точность опыта (p,%),* или *ошибка наблюдений.* Эта величина характеризует субъективную ошибку исследователя. Ошибка выборки выражается в процентах от соответствующей средней: $p\% = \frac{m_M}{M} 100\%$. Точность опыта показывает, на сколько процентов можно ошибиться, если утверждать, что генеральная средняя равна полученной выборочной средней. В 68 случаях из 100 расхождение между выборочной и генеральной средними не будет превышать однократного значения точности опыта (в ту или иную сторону) при нормальном распределении.

Например, имеются средние M_1=86,1±0,7 см и M_2=17,4±0,2 г. По абсолютной величине ошибок трудно сказать, какая средняя определена более точно, поскольку средние выражены разными единицами меры. Необходимо рассчитать точность опыта: чем меньше будет ее значение, тем точнее определена средняя. Для первой средней точность опыта равна 0,18% и для второй – 1,15%, следовательно, первая средняя более достоверна. Точность считается достаточной, если p% не превышает 3–5%. Для определения ошибки наблюдений можно пользоваться и коэффициентом вариации C: $p\% = C/\sqrt{n}$. Как и для любого статистического параметра для точности опыта вычисляется его ошибка $m_p = \pm p/\sqrt{2n}$.

В некоторых экспериментах требуется очень высокая точность опыта (например, в медико-биологических, токсикометрических и др.), когда ошибка не должна превышать 1%. Рассчитанный по формуле процент ошибки необходимо сопоставить с заданным. Если он не выше заданного, то точность достаточная, а если выше, то точность результата является неудовлетворительной, необходимо увеличить число наблюдений.

Для определения объема выборки с заданной точностью опыта используют формулы: $n = C^2 / p^2$ (если точность указана в процентах) и $n = s^2 / m_M^2$ (если точность дается в абсолютных величинах).

Лекция № 5

СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ. ПРОВЕРКА ГИПОТЕЗ. КРИТЕРИИ ЗНАЧИМОСТИ

Почти во всех случаях выборочного наблюдения параметры генеральной совокупности остаются неизвестными. О них приходится судить по выборочным данным. Кроме того, часто возникает необходимость сравнивать две разные выборочные совокупности между собой, опытную и контрольную группу, и т.п. Для сравнительной оценки величины параметров различных совокупностей и проверки достоверности различий между ними в биометрии используется так называемая *нулевая гипотеза*. Согласно этой гипотезе первоначально принимается, что между показателями групп никакой разницы не существует, т.е. они представляют собой однородный материал. Статистический анализ должен привести или к подтверждению нулевой гипотезы, т.е. признанию того, что группы не отличаются по исследуемым параметрам, либо к отклонению нулевой гипотезы, если доказана достоверность различий. Но так как все статистические параметры и различия между ними характеризуются определенным уровнем значимости, то отбрасывание нулевой гипотезы должно быть связано с определенным уровнем значимости. Для самого низкого уровня значимости в биологии 0,05 достоверность разницы между параметрами не может быть ниже 0,95. Если ниже, то нет оснований отбрасывать нулевую гипотезу.

Для установления достоверности разницы между статистическими параметрами выборочных совокупностей используют несколько приемов.

1. *Достоверность различия средних арифметических выборок*. В этом случае пользуются нормированным отклонением *t*. Его осо-

бенности в применении к малым по объему выборкам обосновал английский математик Госсет (1908 г.), который писал под псевдонимом Стьюдент. Он установил, что распределение значений вариант в малых выборках несколько отличается от нормального и зависит от двух величин: нормированного отклонения и объема выборки. С увеличением объема выборки распределение Стьюдента быстро приближается к нормальному и уже при n≥30 практически не отличается от него. Изученное им распределение вошло в науку как t-распределение, а критерий достоверности называется критерием Стьюдента. Математическое выражение критерия следующее: $t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} = \frac{D}{m_d}$.

Значение этого критерия оценивается по таблицам вероятности Стьюдента (приложение 1) на основании числа степеней свободы для заданного уровня вероятности: P=0,95; P=0,99; P=0,999. Число степеней свободы равно $df = n_1 + n_2 - 2$. В таблицах приведены значения стандартных значений критерия для разного числа степеней свободы и уровня значимости. Если фактическое, рассчитанное по опытным данным, t больше стандартного (табличного) t_{st} для данного уровня вероятности, различие существенное, достоверное и его нельзя объяснить случайными причинами. В таком случае отвергается нулевая гипотеза.

2. *Достоверность различия средних квадратичных отклонений выборок.* Разницу между показателями вариации двух независимых распределений можно оценивать с помощью критерия Фишера, обозначаемого через F и представляющего отношение дисперсий (варианс): $F = s_1^2 / s_2^2$. Числителем всегда берется большая варианса, поэтому критерий F может быть равен 1 или больше ее. Если F=1, это указывает на равенство дисперсий. Когда же такого равенства нет, возникает необходимость оценить, случайно расхождение между дисперсиями или нет. Чем больше величина F, тем значительнее расхождение между дисперсиями, и наоборот, чем ближе значение F к 1, тем меньше расхождение между сравниваемыми показателями вариации. Р.А. Фишер получил значения F-критерия для различных уровней значимости и различного числа степеней свободы, стандартные значения он внес в таблицы. Число степеней свободы равно численности каждой выборки без единицы (n–1). Если фактическое значение критерия Фишера F будет больше стандартного (табличного), то различие дисперсий двух выборок доказано и нулевая гипотеза может быть отвергнута.

3. *Критерий соответствия хи-квадрат*. Количественное изучение биологических явлений обязательно требует создания гипотез, с помощью которых можно объяснить эти явления. Чтобы проверить ту или иную гипотезу, нужно получить путем опыта фактические данные и сопоставить их с теоретически ожидаемыми согласно данной гипотезе. Если фактические и теоретические данные совпадают, то это может служить достаточным основанием для признания данной гипотезы, для признания ее правильности. Степень несоответствия фактических наблюдений теоретически ожидаемым результатам может быть различной. В одних случаях разница между ними очень невелика и может оказаться чисто случайной, в других – она весьма значительна. Возникает необходимость в проверке достоверности разницы между теоретически ожидаемыми результатами и эмпирическими данными. При сравнении наблюдаемых и ожидаемых результатов применяются особые критерии оценки, в частности критерий согласия или соответствия хи-квадрат (χ^2). Критерий предложен Карлом Пирсоном и представляет собой сумму отношений между квадратами разностей эмпирических и вычисленных или ожидаемых частот к ожидаемым частотам: $\chi^2 = \sum \frac{(p - p')^2}{p'}$, где Σ – знак суммирования, p – эмпирическая частота, p’ – ожидаемая или теоретически вычисленная частота.

Использование χ^2-теста необходимо для того, чтобы узнать, подтверждается ли гипотеза экспериментом, т.е. насколько верны условия эксперимента, позволяют ли они с высокой степенью достоверности подтвердить или опровергнуть исходное предположение. Если бы фактические данные полностью совпадали с теоретическими, значение критерия было бы равно нулю. По мере увеличения разницы между этими показателями значение критерия будет возрастать. Критерий может принимать только положительные значения от 0 до ∞.

Каждому значению χ^2 соответствует определенная вероятность его появления (табличные данные). Значение χ^2 в таблице указывают те границы, до которых полученные значения критерия не дают оснований сомневаться в высказанном предположении с определенной степенью вероятности. Значений χ^2, превышающие табличные, будут указывать на несостоятельность нулевой гипотезы, т.е. признание того, что различие между фактическими и теоретически ожидаемыми результатами является достоверным, значимым.

Распределение хи-квадрат зависит от n, вернее от числа степеней свободы, которое определяется по формуле $df = (m - 1)$, где m – число сравниваемых классов.

Лекция № 6

ИЗМЕРЕНИЕ СВЯЗИ. КОРРЕЛЯЦИЯ

1. Понятие корреляции. Функциональная зависимость и корреляция.
2. Коэффициент корреляции и корреляционное отношение.
3. Меры линейности и криволинейности.
4. Корреляция между качественными признаками.

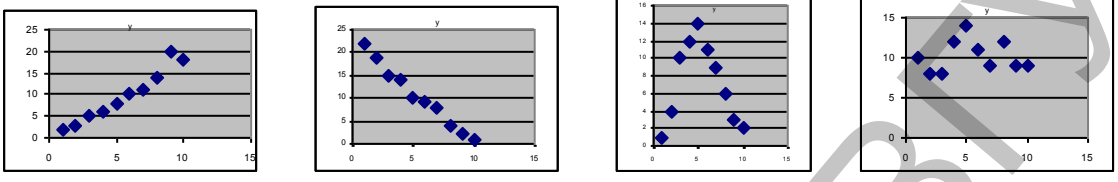
1. Изложенные выше методы анализа биологических данных дают возможность изучать вариацию по каждому отдельному признаку. Вместе с тем, биологические объекты характеризуются многообразием признаков. Например, организмы можно характеризовать возрастом, весом, размерами и т.д. При этом описываемые признаки часто бывают взаимообусловлены. Например, чем старше организм, тем большими размерами он характеризуется. В простейшем случае связь между переменными строго однозначна. Например, вес древесины одного вида полностью определяется ее объемом. Такого рода зависимость называют *функциональной*, когда каждому значению независимой переменной соответствует вполне определенное значение зависимой. Для биологических объектов нередко связь между их характеристиками бывает менее «жесткой»: объекты с одинаковыми значениями одного признака имеют, как правило, разные значения по другим признакам. Такую связь между вариациями разных признаков называют статистической: для такой связи характерно то, что с изменением значения одной переменной меняется распределение другой. Разновидностью статистической связи является *корреляция:* с изменением значения одной переменной меняется среднее значение другой.

По взаимонаправленности связь может быть *прямой* (положительной) – когда с увеличением значений одного признака в общем увеличиваются значения другого, и *обратной* (отрицательной) – когда с увеличением значений одного признака значения другого признака уменьшаются. В случае качественной вариации отрицательная корреляция будет означать, что присутствие одного совпадает с отсутствием другого, а при положительной – присутствие одного признака совпадает с присутствием другого.

По форме связь может быть прямолинейной (линейной), которая аналитически и графически выражается прямой линией, и криволинейной (нелинейной).

Определить, существует ли связь между переменными и является ли она линейной, прямой или обратной, проще всего по *диаграмме рассеяния*, или корреляционной решетке. Это график, отражающий взаимное расположение вариант в корреляционном поле. По оси абсцисс от-

кладывают значения одной переменной, по оси ординат – другой. В графическом поле координат каждой паре значений будет соответствовать одна точка. Их совокупность – *корреляционное поле*, по его форме можно судить о характере связи (рис. 5). Линейная связь является полной, если все точки на диаграмме рассеяния лежат на прямой, сильной или тесной, если облако точек прилегает к прямой достаточно близко; слабой, если облако точек по отношению к прямой широко разбросано.



линейная полож. Линейная отр. Нелинейная отсутствует связь

Рис. 5. **Корреляции между значениями. Х и У.**

2. Основным мерилом связи, существующей между биологическими, признаками, служит коэффициент *корреляции* Пирсона. Он показывает степень приближения корреляционной связи к функциональной (для которой всегда равен единице) и колеблется в пределах от минус (для обратной связи) до плюс (для прямой связи) единицы. Значение коэффициента корреляции, равное нулю или близкое к нулю, говорит лишь об отсутствии прямолинейной связи, но не указывает на наличие или отсутствие криволинейной связи, которая при этом может быть тесной.

Рабочая формула, по которой обычно вычисляется коэффициент корреляции во всех случаях, когда варианты не группируются по классам: $r = \frac{\sum a_x a_y}{n s_x s_y}$.

В числителе этой формулы стоит сумма произведений отклонений вариант от средней арифметической по одному ряду *(X)* на соответствующие отклонения вариант от средней арифметической по другому ряду (У) т.е. $a_x = x_x - M_x$ и $a_y = x_y - M_y$; $\sum a_x a_y = \sum (x_x - M_x)(x_y - M_y)$.

По величине коэффициента корреляции можно установить характер связи (таблица).

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Как и любой статистический параметр коэффициент корреляции – величина, вычисленная по эмпирическим данным выборки, а значит, может содержать случайную ошибку выборочности. Квадратическая ошибка коэффициента корреляции определяется по формуле $m_r = \pm\frac{1-r^2}{\sqrt{n}}$. Она показывает, на какую величину отличается коэффициент корреляции выборки от r генеральной совокупности.

Оценка достоверности коэффициента корреляции осуществляется на основании нулевой гипотезы, согласно которой между исследуемыми случайными величинами корреляция отсутствует. Мерой достоверности служит критерий t-Стьюдента: $t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$, для этих целей можно также пользоваться формулой $t = r/m_r$. Если полученное значение t больше либо равно табличному для данного числа степеней свободы и заданного уровня вероятности, то есть основания отвергнуть нулевую гипотезу и признать существование связи между переменными. Число степеней свободы равно разнице между размером выборки (n) и числом взаимозависимых переменных: k = n – 2.

Для измерения криволинейной зависимости между переменными величинами коэффициент корреляции непригоден. В таких случаях используется другой показатель, предложенный Пирсоном и называемый *корреляционным отношением*. Его принято обозначать греческой буквой η («эта»). В отличие от коэффициента корреляции, который четко определяет взаимосвязь между одной независимой и одной зависимой переменными, корреляционное отношение описывает такую связь двусторонне. Например, при наличии линейной связи между ростом и весом человека значения роста – независимая переменная, веса – зависимая. Коэффициент корреляции определяет степень зависимости веса от роста. Корреляционное отношение при нелинейной связи описывает как зависимость веса от роста, так и, наоборот, роста от веса, поэтому выражается не одним, а двумя показателями $\eta_{y/x}$ и $\eta_{x/y}$. Они определяются по следующим формулам:

$$\eta_{y/x} = \sqrt{s_{yx}^2 / s_y^2} \text{ и } \eta_{x/y} = \sqrt{s_{xy}^2 / s_x^2},$$

где

$s_{yx}^2 = \frac{\sum (M_x - M)^2}{n}$, где M_x – среднее значение переменных y, соответствующих каждому значению x, M – общая средняя переменной y.

$s_y^2 = \frac{\sum (y_i - M)^2}{n}$ – общая дисперсия выборки.

Соответственно $s_{xy}^2 = \frac{\sum (M_y - M)^2}{n}$ и $s_x^2 = \frac{\sum (x_i - M)^2}{n}$.

Корреляционное отношение – величина относительная; этот показатель принимает значения от 0 до 1. Чем сильнее связь между признаками, тем выше значение η. При отсутствии корреляции коэффициент равен 0. Показатели корреляционного отношения для обоих признаков обычно не равны между собой, исключая строго линейную связь.

Ошибка корреляционного отношения и его достоверность определяются по формулам: $m_\eta = \pm\frac{1-\eta^2}{\sqrt{n}}$ и $t = \eta\sqrt{\frac{n-2}{1-\eta^2}}$ или $t = \eta / m_\eta$. Число степеней свободы df = n – 2.

3. Поскольку коэффициент корреляции характеризует только линейную связь, а корреляционное отношение – любую форму связи, то при строго линейной зависимости между переменными X и У должно соблюдаться равенство, во-первых, между коэффициентами корреляционного отношения $\eta_{y/x} = \eta_{x/y}$, а во-вторых, между корреляционным отношением и коэффициентом корреляции η=r. При наличии нелинейной связи эти равенства не соблюдаются. Следовательно, по разности между этими показателями можно судить о форме корреляционной зависимости между варьирующими признаками. В качестве показателя линейности связи, обозначаемого греческой буквой γ, используется разность между квадратами корреляционного отношения и коэффициента корреляции, т.е. $\gamma = \eta^2 - r^2$. Выборочная ошибка этого показателя определяется по следующей приближенной формуле: $m_\gamma = \pm\sqrt{\frac{\gamma}{\sum n}}$, где n – сумма численностей (для малых выборок n берется числом степеней свободы). Критерием достоверности показателя служит его отношение к своей ошибке: $t_\gamma = \gamma / m_\gamma$. При t<3 корреляция между признаками оценивается практически прямолинейной. В более ответственных случаях корреляция принимается прямолинейной при t<2 (в этом случае вероятность правильности криволинейности составляет 0,95).

Показатель криволинейности определяется по формуле: $Kp = \frac{\eta^2 - r^2}{1 - r^2}$. При Кр=0 имеем строго прямолинейную зависимость двух признаков. При Кр=1 криволинейный характер связи достигает максимального предела.

4. На практике приходится определять степень связи не только между количественными, но и между качественными признаками, а также между теми и другими. Для этого используют различные показатели меры связи.

1. Для определения связи между двумя качественными признаками пользуются *коэффициентом ассоциации Дж. Юла, или коэффициентом сходства*. Для этого предварительно составляют четырехклеточную корреляционную решетку для вычисления некоторых параметров. Например, при сравнении урожайности семян ели в течение двух лет получены следующие данные о качестве семян:

2.

Год	Кол-во семян		Всего
	выполненных	пустых	
Урожайный	n_1=68	n_2=32	N_1=100
Неурожайный	n_3=24	n_4=76	N_2=100
Всего	N_3=92	N_4=108	

Требуется установить тесноту связи между качеством семян и обилием урожая.

Коэффициент ассоциации рассчитывается по формуле:

$$r_a = \frac{n_1 n_4 - n_2 n_3}{\sqrt{N_1 N_2 N_3 N_4}},$$

где n_1, n_2, n_3, n_4 – численности альтернативных признаков, расположенные в клетках корреляционной решетки,

$N_{1\text{-}4}$ – соответствующие суммарные численности признаков.

Коэффициент ассоциации выражается в долях от единицы, чем сильнее связь между признаками, тем выше и коэффициент ассоциации. Достоверность коэффициента ассоциации определяется отношением его значения к своей ошибке: $t_{r_a} = r_a / m_{r_a}$, а выборочная ошибка – $m_{r_a} = \pm \frac{1 - r_a^2}{\sqrt{n}}$.

Лекция № 7

РЕГРЕССИОННЫЙ АНАЛИЗ

1. Понятие о регрессии.
2. Построение эмпирических рядов регрессии.
3. Уравнение регрессии.
4. Коэффициенты регрессии.

1. Применение корреляционного анализа дает возможность измерять степень сопряженности между признаками, определять направление и форму существующей между ними связи. Вместе с тем, коэффициент корреляции и корреляционное отношение не дают представления о том, насколько в среднем может измениться варьирующий признак при изменении значения другого, независимого признака. Вместе с тем, при оценке степени взаимосвязи статистических величин важно провести математическое моделирование, т.е. подобрать аналитическое уравнение, которое соответствовало бы природе изучаемого явления с целью предсказания поведения независимой характеристики объекта при изменении зависимого параметра. Динамика взаимной зависимости между переменными величинами получила название *регрессии*, а методика исследования регрессии носит название *регрессионного анализа*.

Функция, позволяющая по величине одного признака (Х) находить средние значения другого, зависимого от Х признака (У), называется *регрессией*.

Термин ввел в биологию Ф. Гальтон, изучавший соотношение между ростом родителей и их детей. Им был установлен так называемый «закон регрессивного наследования», по которому дети очень высоких и очень низких родителей имеют тенденцию отклоняться в своем развитии («регрессировать») в сторону среднего для данной популяции роста. Так вошел в науку этот термин.

Функцией обычно обозначают зависимый признак, а независимый – аргументом. В общем виде взаимосвязь между функцией и аргументом можно выразить формулой $y = f(x)$, то есть признак у есть функция от х. Регрессионный анализ представляет собой процедуру выбора уравнения, наиболее точно отражающего зависимость одного признака от другого, вычисления коэффициентов уравнения и оценку его точности.

Показатели регрессии измеряют отношение между коррелирующими признаками двусторонне, т.е. учитывают изменения Х в зависимости от У и наоборот. Исключением служат лишь так называе-

мые ряды динамики, или временные ряды, показывающие изменение признака во времени. В данном случае аргументом будет не какой-то другой признак, а фактор времени.

2. Регрессия может быть выражена несколькими способами: путем построения так называемых *эмпирических линий регрессии*, путем составления *уравнений регрессии* и, наконец, с помощью вычисления *коэффициента регрессии*. Первые два способа позволяют выразить регрессию графически.

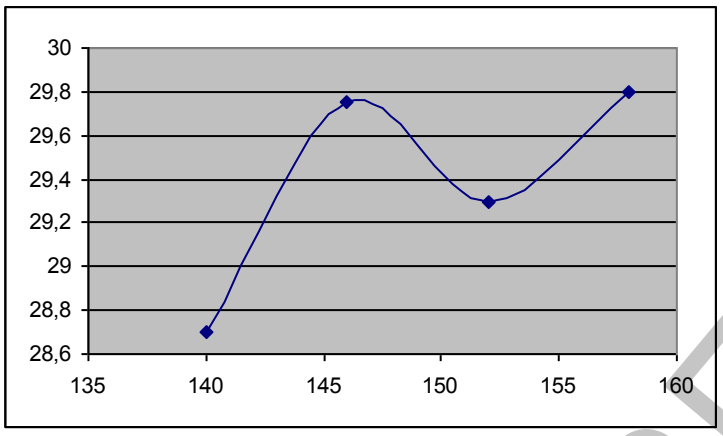
1. Эмпирические линии регрессии строят, пользуясь обычной корреляционной решеткой. Для этого по значениям одного признака рассчитывают групповые средние другого признака, что и дает в результате эмпирический ряд регрессии.

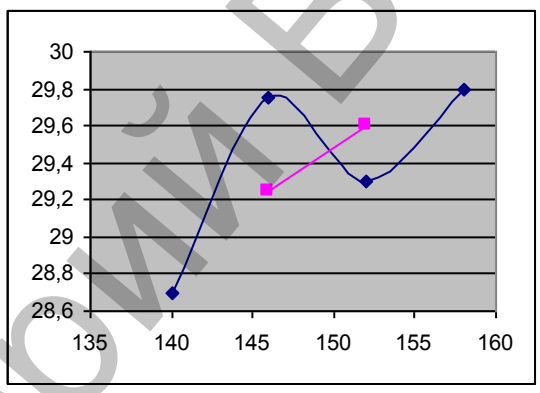
Например, в таблице даны данные о весе детей разного роста.

Рост, см (x)	Вес, кг (y)			Средние по весу для каждого роста (y_x)
	28	29	30	
140	1	2		(28+2·29)/3=28,7
146		1	3	(29+3·30)/4=29,75
152		2	1	29,3
158		1	5	29,8
Средние по росту для каждого веса, (x_y)	140	(140·2+146·1)/3	146	

На основании показателей $\bar{x}$/y и $\bar{y}$/x можно построить на одном графике линии регрессии. На оси x отмечены средние значения классов x, на оси y – средние значения классов y. Уже по одному виду линии регрессии можно определить, какая форма связи имеет место в данном конкретном случае (прямолинейная, параболическая и т.п.).

2. Эмпирическая линия регрессии обычно представляет собой ломаную линию. Это связано с влиянием на признаки, кроме основных, многочисленных второстепенных факторов, нарушающих плавный ход линии регрессии. Чтобы отобразить функциональную зависимость между x и y при полной изоляции действующих на нее причин проводят процедуру выравнивания эмпирических рядов регрессии. Существует несколько способов выравнивания:





a) *графический способ*. На графике на глаз определяются срединные точки линии регрессии, которые затем соединяются при помощи линейки сплошной линией. Недостаток этого способа – субъективное влияние исследователя на результаты выравнивания.

b) *способ скользящей средней*. Заключается в последовательном исчислении средних арифметических из двух или трех соседних значений ряда.

Рост, см 140 146 152 158

Средний вес, кг 28,7 29,75 29,3 29,8

Находим сумму первых трех значений ряда «вес»: 28,7+29,75+29,3 =87,75. Среднее арифметическое этих значений – 87,75/3=**29,25**. Далее суммируем следующие члены ряда и находим соответствующее значение среднего арифметического: **29,6**. Полученные средние значения наносят на график: x – полученные скользящие средние, y – срединные значения y из трех в каждой сумме.

a. Линия регрессии дает представление о характере связи, но не дает возможности точно определить любое значение X по заданному У или наоборот. Для этой цели могут послужить уравнения. Уравнение прямолинейной регрессии в общем виде $y_i-M_y=b\cdot(x_i-M_x)$. Оно выражает определенную зависимость, а именно: вслед за отклонением x от средней происходит и отклонение y от средней, на величину, равную b. При переносе M_y в правую часть равенства получим

$y_i=M_y+b(x_i-M_x)$. Если M_x приравнять к нулю, то M_y будет первоначальным значением y, с которого надо начинать построение линии регрессии при $x_i=0$. Его можно обозначить через a. Уравнение регрессии в таком случае принимает вид обычного уравнения прямой: y=a+bx. Здесь x и y – коррелирующие переменные, a – первоначальное значение y при $x_i=0$ (означает уровень регрессии, т.е. ее подъем от начала координат по оси координат), b – коэффициент пропорциональности, который показывает, на какую величину изменится y, если x изменится на единицу.

Следующий этап регрессионного анализа – определение значений a и b в уравнении. Для этого необходимо решить систему уравнений:

$$\sum y = an + b\sum x$$
$$\sum xy = a\sum x - b\sum x^2 .$$

Значения x, y – эмпирические результаты опыта, n – объем выборки.

Определив значения a и b, подставляем их в уравнение регрессии y=a+bx. Имея конкретные значения x, можно достаточно точно предположить значения y.

Например, зависимость между средним весом новорожденных гамадрилов и весом самок.

Вес самок, x	10	10,8	11,3	10	10,1	Σx=52,2
x^2	100	116,64	127,69	100	102,01	Σx^2=546,34
Вес детенышей, y	0,7	0,73	0,75	0,7	0,65	Σy=3,53
x·y	7	7,884	8,475	7	6,565	Σ x·y=36,924

Подставляем данные в систему уравнений:

3,53=5a + 52,2b

36,924=52,2a – 546,34b

Решаем уравнения, в результате получаем значения в = **-53,8** и a = **562,4.** Полученные величины a и b подставляем в уравнение регрессии y=a+bx: **y= 562,4-53,8x.**

Данное уравнение характеризует прямолинейную зависимость. При наличии разных форм криволинейной зависимости используют более сложные уравнения регрессии.

b. Значительно легче составить уравнение регрессии, если известно значение коэффициента b. Дело в том, что этот коэффициент представляет собой величину регрессии У по Х или Х по У, то есть *коэффициент регрессии* (R). Коэффициент регрессии может использоваться и самостоятельно как количественная мера регрессии. В силу двусторонности регрессии коэффициентов тоже может быть два: $R_{x/y}$ и $R_{y/x}$.

$$R_{x/y} = r\frac{s_x}{s_y} \quad R_{y/x} = r\frac{s_y}{s_x}.$$

Если не известны значения коэффициентов корреляции, тогда можно воспользоваться формулами:

$Ry/x = \frac{\sum(x_i - M_x)(y_i - M_y)}{\sum(x_i - \overline{x})^2}$ – коэффициент регрессии *Y.X;*

$R_{x/e} = \frac{\sum(x_i - M_x)(y_i - M_y)}{\sum(y_i - \overline{y})^2}$ – коэффициент регрессии *X.Y.*

Коэффициент регрессии сопровождается средней квадратической ошибкой, которая вычисляется то формулам: $m_{byx} = \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}}$ и $m_{bxy} = \frac{s_x}{s_y}\sqrt{\frac{1-r^2}{n-2}}$.

Подобно любому статистическому параметру, коэффициент регрессии всегда определяется на основе выборочной совокупности, а значит, его определение сопряжено со случайной ошибкой выборочности. При оценке достоверности коэффициента используют нулевую гипотезу об отсутствии связи между признаками, а, значит, признание того, что коэффициент регрессии равен нулю. Достоверность определяем по формуле $t = \frac{b}{m_b}$. Полученное фактическое значение t следует сравнить со стандартным для n-2 числа степеней свободы. Если значение фактического больше стандартного, коэффициент достоверен при заданном уровне значимости.

Лекция № 8

ДИСПЕРСИОННЫЙ АНАЛИЗ

1. *Сущность метода и его основные задачи.*
2. *Основные понятия и термины.*
3. *Однофакторный дисперсионный анализ.*
4. *Двухфакторный дисперсионный анализ.*

1. Анализируя биологические объекты, мы всегда сталкиваемся с причинной обусловленностью многих их особенностей. Например, признаки потомков обусловлены наследственными факторами, полученными от родителей, возникновение у людей избыточного веса может быть следствием заболевания или неправильного питания, мигре-

ни и повышение давления у гипертоников – результат погодных аномалий и др. Корреляционный и регрессионный анализы выявляют наличие связи между признаками, а также степень этой связи. Между тем, эти статистические методы не эффективны, когда следует выявить влияние отдельных факторов и оценить их относительную роль. Можно, конечно, использовать метод попарных сравнений средних арифметических этих показателей, но это неудобно, если исследуется большое число показателей. Более совершенным является метод сравнения выборочных дисперсий, вернее их отношений, с критическими значениями критерия F. Этот метод уже описывался ранее, когда речь шла о проверке достоверности различий между исследуемыми выборочными показателями (критерий Фишера). Сам метод, разработанный Фишером в 1925 году, получил название *дисперсионного анализа*. Сущность этого вида анализа заключается в установлении роли отдельных факторов в изменчивости того или иного признака.

Дело в том, что влияние тех или иных факторов на изучаемый признак никогда нельзя выделить в чистом виде. Хотя при проведении опытов и стараются сохранить условия максимально однородными, все же на результаты опыта влияют многочисленные случайные обстоятельства, факторы, не поддающиеся контролю. Еще сложнее обстоят дела, если эксперимент проводится в природе, где на организмы совокупно и одновременно действует огромное количество факторов среды.

Задача дисперсионного анализа – разложить общую изменчивость признака на составные части: те, что определяются конкретными изучаемыми факторами, и те, что вызываются случайными, неконтролируемыми обстоятельствами. Фактически дисперсионный анализ дает ответ на вопрос: влияет ли фактор на изучаемый признак. Дисперсионный анализ можно проводить как на малых, так и на больших выборках, на однородном и биологически разнокачественном материале, например, на особях разного пола, возраста, видовой или расовой принадлежности. Этот метод один из самых эффективных и мощных методов биометрического анализа. Он позволяет решать самые разнообразные задачи.

2. Признаки, изменяющиеся под воздействием тех или иных причин, называются *результативными*. А действующие на эти признаки причины – *факторы*. Например, рост, вес, физическое состояние популяции – это признаки, а нормы питания, воздействие климата, лекарственных препаратов – это категория факторов. Факторы делятся на контролируемые в опыте, или *организованные*, и *неорганизованные*, или неконтролируемые, действие которых на признак не регулируется. Организованные факторы принято обозначать большими буквами латинского алфавита – A, B, C, D, а результативные признаки

через X, Y и т.д. Обычно каждый из факторов представлен некоторым количеством подразделений (групп), называемых *градациями*. Они обозначаются теми же буквами, что и факторы с индексами – A_1, A_2 и т.д. Градации фактора определяются степенью воздействия на признак, например, дозами веществ, видовой принадлежностью.

В пределах каждого уровня градации отдельные переменные могут принимать различные значения, т.е. наблюдается случайная вариация. Например, люди, принимающие одинаковое количество одинаковой пищи, обладают разным весом. Выделение уровней градации всегда сопряжено с некоторыми сложностями. Например, изучая влияние сезонов на какой-то результативный признак, мы можем четко установить градацию этого фактора: весна, лето, зима… С другой стороны, часто мы имеем дело с влиянием факторов, плохо поддающихся градации. Например, число детенышей в помете – как фактор, влияющий на их вес. Объем помета может варьировать в очень широких пределах. Такие факторы и называются случайными, т.е. случайными могут быть разные градации этих факторов. В связи с этим возможны очень разные схемы, или модели, для дисперсионного анализа. Они могут различаться по числу анализируемых факторов для анализа (одно-, двух-, многофакторные и т.д.), по характеру градации внутри фактора (с фиксированными факторами, со случайными, смешанные схемы).

Нулевая гипотеза – важный элемент любого статистического анализа. В случае дисперсионного анализа первоначально принимаемой гипотезой является утверждение о том, что данный фактор А (или В, или С и т.д.) не влияет на результативный признак. Если правильна нулевая гипотеза, s^2_A должна быть равна нулю (то же касается и дисперсий остальных факторов), т.е. вся вариация сводится только к случайной. Для того чтобы отбросить нулевую гипотезу, нужно доказать, что s^2_A достоверно (т.е. с вероятностью не меньшей, чем 0,95, или с уровнем значимости 0,05) отличается от нуля. Достоверность значения s^2_A может быть установлена, как это обычно делают по отношению к любому статистическому параметру, путем деления его на ошибку.

3. Однофакторным называется комплекс, в котором учитывается действие на признак только одного организационного фактора А. Однофакторные комплексы могут состоять из малочисленных и больших групп, в градациях которых может быть равное и неравное количество вариант. Рассмотрим простейшую схему анализа влияния одного фактора, принимающего разные градации, или количественные уровни: 1, 2, .., i, .. , a. Отдельные наблюдения (варианты) разбиваются на группы согласно этим градациям фактора.

Исходные данные лабораторных испытаний влияния фактора погоды на изменение продолжительности систолической остановки

сердца при введении хлорида бария для четырех групп экспериментальных животных

Градации фактора погода	Показатели экспериментальных групп				Среднее групповое значение по каждой градации фактора, $\overline{x_i}$	Квадраты $\overline{x_i}$	
	1	2	3	4			
Тихая погода	13,8	11	13,7	12,1	12,65	160,02	
Ветер и вьюга	16	12,2	15,8	14,3	14,57	212,28	

Общая вариация разлагается на 3 компонента: общее варьирование всех вариант, независимо от того, в какой группе они находятся, или общая вариация; вариация групповых средних (по градациям фактора) вокруг общей средней, или межгрупповая вариация; вариация отдельных вариант внутри групп, или внутригрупповая вариация. Последняя как раз и создается неконтролируемыми факторами (кроме учитываемого фактора А). Задача дисперсионного анализа сравнить две вариации: межгрупповую дисперсию, характеризующую действие контролируемого фактора на результативный признак (*факториальную дисперсию*), и внутригрупповую дисперсию, характеризующую варьирование признака под действием неконтролируемых факторов (*остаточную, или случайную дисперсию*). При этом внутригрупповое варьирование не зависит от варьирования межгруппового. Дисперсия есть сумма квадратов отклонений вариант от средней арифметической. Задача, таким образом, сводится к нахождению двух дисперсий, сравнению их и проверке достоверности выводов о влиянии фактора на признак.

Общепринятые обозначения:

x – варианты дисперсионного комплекса;

x_i – варианты отдельных градаций или групп дисперсионного комплекса;

M – общая средняя арифметическая;

M_i – групповая средняя арифметическая;

N – общее число вариант в данном комплексе;

n_i – число вариант в группах комплекса;

n – число групп (градаций).

В общем виде алгоритм анализа следующий.

1. Для начала находят среднее арифметическое значение всего комплекса, называемое общей средней (M) (13,67) и частные или групповые средние M_i – по градациям фактора A.

2. Находят сумму всех вариант, составляющих этот комплекс: $\sum x$ =13,8+11+13,7+12,1+…=108,9 и квадрат этой суммы – $(\sum x)^2$=108,9^2=11859,2.

3. Определяют сумму квадратов тех же вариант: $\sum x^2$=13,8^2+11^2+13,7^2+12,1^2+…= 1504,42.

4. Возводят в квадрат групповые средние (табл.) и находят сумму их квадратов: $\sum M_i^2$ = 12,65^2 + 14,57^2=126,5+212,2=338,8.

5. Находят общую сумму квадратов отклонений: $D_y = \sum x^2 - \frac{(\sum x)^2}{N}$ =1504,42-11859,2/8=22,02.

6. Межгрупповую сумму квадратов отклонений: $D_x = n\sum M_i{}^2 - \frac{(\sum x)^2}{N}$ = 4·338,8 – 11859,2/8= -127,2.

7. Внутригрупповую или остаточную сумму квадратов отклонений: $D_z = \sum x^2 - n\sum M_i{}^2$ = 1504,42-4·388,8=149,22.

8. Определяют число степеней свободы. Число степеней свободы для общей дисперсии равно k_y=N-1=8-1 = 7. Комплекс имеет 2 градации, число степеней свободы для межгрупповой дисперсии k_x=n-1=2-1=1. Для внутригрупповой дисперсии число степеней свободы равно k_z=(N-1)-(n-1)= 7-1=6.

9. Находят значения дисперсий:

общую всего комплекса $s_y^2 = \frac{D_y}{k_y}$ =22,02/7=3,15;

межгрупповую или факториальную $s_x^2 = \frac{D_x}{k_x}$ =127,2/1=127,2;

внутригрупповую или случайную $s_z^2 = \frac{D_z}{k_z}$ = 149,22/6=24,87

10.Сводят полученные величины в итоговую таблицу дисперсионного анализа.

Источники вариации	Степени свободы	Сумма квадратов отклонений	Средний квадрат (s^2)	$F_ф$	F_{st}	
					P=0,05	P=0,01
Межгрупповая (влияние фактора A)	1	127,2	127,2	>1 (5,11)	5,99	13,75
внутригрупповая	6	149,22	24,87	1		
общая	7	22,02	3,15			

При нулевой гипотезе $s_x^2 = s_z^2$, т.е. F=1. Чтобы отвергнуть нулевую гипотезу, надо доказать, что неравенство сравниваемых дисперсий не случайно, т.е. выходит за пределы обычной ошибки. Граничные значения F приведены в таблице критических значений критерия Фишера (см. таблицу). Достоверным признается такое значение F, которое больше табличного. В нашем случае фактическое значение критерия равно 5,11. Оно меньше табличного, следовательно, мы не можем отвергнуть нулевую гипотезу и вынуждены признать, что влияние фактора погоды на исследуемый признак статистически недостоверно. Т.е. вариация продолжительности систолической остановки сердца при введении хлорида бария в разную погоду случайна, так же как и вариация этого показателя при одной и той же погоде внутри групп исследуемых объектов.

Достоверность различий между средними дисперсионного комплекса. Дисперсионный анализ охватывает сразу весь комплекс наблюдений и позволяет оценить достоверность или случайность различий, наблюдаемых между групповыми средними. В этом заключается его преимущество перед дробным методом анализа выборочных данных, когда достоверность оценивается между отдельно взятыми парами средних показателей. В то же время иногда требуется оценить достоверность различий, наблюдаемых между средними арифметическими дисперсионного комплекса. Это можно сделать с помощью критериев Фишера или Стьюдента.

1. Для оценки различий между двумя средними $M_1 - M_2 = D$ в дисперсионном комплексе критерий Фишера определяется по следующей формуле: $F_D = \frac{D^2}{s_z^2} \times \frac{n_1 \times n_2}{n_1 + n_2}$, где s_z^2 – остаточная дисперсия (внутригрупповая сумма квадратов отклонений). При одинаковых числах вариант и градациях комплекса в сравниваемых группах, т.е. при $n_1=n_2$, эта формула упрощается $F_D = \frac{D^2}{s_z^2} \times \frac{n}{2}$. Числа степеней свободы равны $k_1 = a–1 = 2–1 = 1$ и $k_2 = k_z$.

2. При использовании критерия Стьюдента достоверность разницы между средними дисперсионного комплекса оценивается по отношению разницы между групповыми средними к выборочной ошибке этой разницы. Выборочная ошибка разницы определяется по формуле $m_D = m_M \sqrt{2m_M^2} = 1{,}414 s_z$, где m_M – выборочная ошибка частной средней, определяемая по формуле $m_M = \sqrt{\frac{s_z^2}{n_i}}$. Число степеней свободы определяется как $k=k_z-1$. Полученное значение критерия Стьюдента сравнивается с табличным для данного числа степеней свободы и заданного уровня вероятности. Нулевая гипотеза отвергается, если $t_ф > t_{st}$.

Оценка силы влияния факторов на результативный признак. Дисперсионный анализ позволяет установить не только достоверность, но и силу влияния регулируемых и нерегулируемых в опыте факторов на результативный признак.

Сила влияния фактора определяется как доля факториальной вариации в общем варьировании признака.

В качестве показателя силы влияния используется отношение факториальной суммы квадратов D_x к общей сумме квадратов D_y дисперсионного комплекса: $\eta_x^2 = D_x / D_y$. Критерием достоверности этого выборочного показателя, как и любого другого, служит его отношение к ошибке: $t(F_\phi) = \eta_x^2 / m_\eta$. Ошибку можно определить по формуле $m_\eta = (1-\eta_x^2)\frac{a-1}{N-a}$, где a – число градаций фактора A, N – общий объем дисперсионного комплекса. Нулевая гипотеза отвергается, если $F_ф \geq F_{st}$ для чисел степеней свободы $k_1=a-1$ и $k_2=N-a$.

4. В двухфакторных дисперсионных комплексах на результативный признак влияют два фактора A и B, между которыми существует взаимодействие. Это осложняет дисперсионный анализ. При двухфакторной схеме общая сумма квадратов отклонений разлагается не на 3, как при однофакторном анализе, а на 4 компонента:

а) вариация под влиянием фактора A;

б) вариация под влиянием фактора B;

в) вариация под совместным влиянием двух факторов A и B;

г) случайные отклонения вариант внутри групп от средней.

Обработка таких комплексов ничем принципиально не отличается от описанной выше схемы, хотя немного усложняется техника расчетов.

Исходные данные сводят в таблицу. Например:

Сравниваемые группы (фактор A)	% крахмала в зернах при применении удобрений (фактор B)			Групповая средняя M_i	Σx_B	$(\Sigma x_B)^2$	$(\Sigma x_B)^2/n_B$
	удобрение 1	удобрение 2	удобрение 3				
Сорт 1	2	3	2	**2,3**	7	49	16,3
Сорт 2	3	4	1	**2,7**	8	64	21,3
Сорт 3	2	4	2	**2,7**	6	64	21,3
Σx_A	7	11	5				
$(\Sigma x_A)^2$	49	121	25				
$(\Sigma x_A)^2/n_A$	16,3	40,3	8,3				

1. Все варианты дисперсионного комплекса суммируют, найденную сумму возводят в квадрат и относят к общему числу наблюдений.

$\Sigma x=2+3=2+\ldots=23$. $(\Sigma x)^2=529$, $(\Sigma x)^2/N=58,8$

2. Каждую варианту комплекса возводят в квадрат и находят сумму квадратов:

$\Sigma x^2=2^2+3^2+\ldots=67$.

3. Рассчитывают сумму квадратов отклонений по стандартным формулам.

Общая: $D_y=\sum x^2-\frac{(\sum x)^2}{N}$ $D_y=67-58,8=8,2$

По фактору А – $D_A=\sum\frac{(\sum x_A)^2}{n_A}-\frac{(\sum x)^2}{N}$ $D_A=(16,3+40,3+8,3)-58,8=6,1$.

По фактору В – $D_B=\sum\frac{(\sum x_B)^2}{n_B}-\frac{(\sum x)^2}{N}$ $D_B=(49+64+64)-58,8=118,2$.

Межгрупповая (по сочетанию градаций) $D_x=\sum\frac{(\sum x_i)^2}{n_i}-\frac{(\sum x)^2}{N}$, где Σx_i – сумма вариант по градациям комплекса. Для ее вычисления нужно каждую варианту возвести в квадрат, разделить на число вариант в каждой группе каждой градации. В приведенном примере у нас по 1 варианте в каждой группе каждой градации, следовательно, эта величина равна сумме квадратов всех вариант $(\Sigma x)^2=529$. $D_x=529-58,8=470,2$.

Остаточная, или случайная $D_z=D_y-D_x$. $D_z=8,2-470,2=-462$.

По взаимодействию факторов АВ $D_{AB}=D_x-D_A-D_B$ $D_{AB}=8,2-6,1-118,2=-116,1$.

4. Определяют степени свободы, которые равны:

для общей дисперсии: $k_y=N-1=9-1=8$;

дисперсии по фактору А: $k_a=a-1=3-1=2$, где n – число градаций фактора А;

дисперсии по фактору В: $k_в=в-1=3-1=2$, где в – число градаций фактора В;

по взаимодействию факторов АВ – $k_{AB}=(a-1)(b-1)=2\cdot 2=4$;

остаточной дисперсии $k_z=N-ab=9-9=0$.

5. Определяют дисперсии:

общая : $s_y^2=D_y/k_y$ $s_y^2=8,2/8=1,025$;

по фактору А:– $s_A^2=D_A/k_A$ $s_A^2=6,1/2=3,05$;

по фактору В: $s_B^2=D_B/k_B$ $s_B^2=118,2/2=59,1$;

по взаимодействию факторов АВ – $s_{AB}^2=D_{AB}/k_{AB}$ $s_{AB}^2=116,1/4=29,025$;

остаточная (отражающая действие на результативный признак неконтролируемых (случайных) факторов) $s_Z^2 = D_Z / k_Z$ $s_Z^2 = 462/0\ (1) = 462$.

6. Определяют критерий Фишера по отношению дисперсий факторов А, В и АВ к остаточной дисперсии. Результаты заносят в таблицу.

Источники вариации	Степени свободы	Сумма квадратов отклонений	Дисперсия	$F_ф$	F_{st} (P=0,05)
По фактору А	2	6,1	3,05	0,006	18,1
По фактору В	2	118,2	59,1	0,128	18,1
Совместная АВ	4	116,1	29,025	0,063	7,7
Остаточная	0 (1)	462	462		
Общая	8	8,2	1,025		

Все полученные фактические значения критерия меньше стандартных для уровня значимости 0,05, следовательно, можно сделать вывод о том, что ни сортовая принадлежность, ни применяемое удобрение, ни оба эти фактора вместе не оказывают достоверного влияния на % крахмала в зернах растений.

Лекция № 9

СТАТИСТИЧЕСКИЙ АНАЛИЗ ВАРИАЦИИ ПО КАЧЕСТВЕННЫМ ПРИЗНАКАМ

1. *Группировка вариант качественных признаков. Альтернативная вариация.*
2. *Среднее арифметическое и среднее квадратическое отклонение при альтернативной вариации.*
3. *Выборочная ошибка и расчет оптимального объема выборки.*
4. *Достоверность разницы между параметрами выборочной и генеральной совокупностей.*
5. *Корреляция при качественной вариации.*

1. Исследователю часто приходится иметь дело с совокупностями, в которых объекты различаются по качественным признакам, например, окраска мехового покрова или цветков растений, реагирование или нереагирование организмов на определенные факторы среды и др. Группировка особей таких совокупностей сводится к распределению их по каждой качественной группе и выражению количества

особей каждой группы в виде относительной доли в общем объеме совокупности. Эта доля может быть выражена в процентах или долях от единицы. Таким образом, при изучении качественных признаков мы имеем дело со следующими величинами: 1) абсолютные численности группы – их обозначают символами p_0, p_1 и т.д.; 2) их доли, выраженные в долях единицы или в процентах – q,p,r,s и т.д.

Качественные признаки могут быть в такой же степени подвергнуты статистическому анализу, как и количественные, и по отношению к ним устанавливают те же параметры: наиболее типичное значение признака, степень вариации вокруг него, установление зависимости между разными признаками, оценка достоверности полученных данных и т.д.

Простейшим случаем качественной вариации является альтернативная, когда совокупность состоит только из двух групп: одной, имеющей данный признак, и другой – его не имеющей. Численность первой группы можно обозначить через p_1, другой – через p_0. Тогда доля особей, имеющих признак будет равна $p=p_1/n$, а доля – не имеющих его – $q=p_0/n$ или 1-p. Так как с альтернативной вариацией легче работать, то в ряде случаев целесообразно превращать несколько качественных групп в две альтернативные.

2. Возникает вопрос: можно ли при качественной вариации вычислять статистические показатели, как это делалось для количественных признаков? Для этого следует разобрать, каким будет вариационный ряд. В общем виде варианты при альтернативной изменчивости могут быть представлены в виде двух классов: «0» и «1». Относительная доля особей каждого класса в общей совокупности соответствует средней арифметической при количественной вариации, т.е. M= $p=p_1/n$. Среднее квадратичное отклонение определяется выражением $s_p = \sqrt{pq}$. Так как 1-p=q, то это выражение можно преобразовать: $s_p = \sqrt{p(1-p)}$. Дисперсия в таком случае определяется выражением $s^2=pq=p(1-p)$.

Например, в стаде из 284 коров, подвергнутых туберкулинизации, отрицательную реакцию дало 201 животное, положительную – 83. Подвергнем результаты статистической обработке.

Классы	Частоты
0	201
1	83
	n=284

В этом случае p=83/284=0,29 (29%) q=201/284=0,71 (71%). $s_p = \sqrt{0,29 \cdot 0,71}$ =0,45 (45%).

3. Как и в случае количественной изменчивости, частота качественного признака имеет свою статистическую ошибку, так как она определяется на основе изучения конкретной выборочной совокупности. Значения полученных долей, определенные для ряда выборочных совокупностей, будут колебаться вокруг доли генеральной совокупности по тем же законам, что и для количественных параметров. Мерой этих колебаний является средняя, или статистическая, ошибка. Применительно к качественной вариации она вычисляется по следующей формуле: $m_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$. Для приведенного выше примера статистическая ошибка равна $m_p = \sqrt{\frac{0{,}29 \cdot 0{,}71}{284}} = 0{,}027$ (2,7%). Ошибка будет одна и та же как для доли реагировавших на прививку, так и для доли нереагировавших: $p \pm m_p = 0{,}29 \pm 0{,}027$ и $q \pm m_q = 0{,}71 \pm 0{,}027$. Или можно выразить эти показатели в процентах. Так как величины p и q изменяются в обратном отношении друг к другу, то исследователь может проверить точность своих результатов, исходя из расчета максимально возможной для каждого данного значения n средней ошибки, которая не может превышать величину $\sqrt{\frac{0{,}5 \cdot 0{,}5}{n}} = \sqrt{\frac{0{,}25}{n}}$. Приведенная формула ошибки справедлива для случаев, когда объем выборочной совокупности невелик по сравнению с генеральной. Если же выборочная совокупности составляет значительную долю генеральной, следует ввести в подкоренное выражение поправку на множитель 1-n/N, где N – объем генеральной совокупности. В таком случае формула для ошибки примет вид: $m_p = \sqrt{\frac{pq}{n}(1-\frac{n}{N})}$.

Иногда возникает необходимость определить возможную колеблемость не доли (%), а количеств особей с тем или иным признаком. Тогда в формуле ошибки надо заменить значения p и q абсолютными численностями p_0, p_1, а величину ошибки для перевода долей в абсолютные численности умножить на n: $m_{p_1} = m_{p_0} = \sqrt{\frac{p_1(n-p_1)}{n}}$. В приведенном примере величина ошибки в абсолютном измерении будет равна $m_{p_1} = m_{p_0} = \sqrt{\frac{83(284-83)}{284}} = 7{,}7$. Конкретные числа особей с их ошибками можно записать так: $p_1 \pm m_p = 83 \pm 7{,}7$ и $p_0 \pm m_q = 201 \pm 7{,}7$.

При учете качественной вариации возможен случай, когда в какой-то выборочной совокупности нет ни одного случая с признаком А, т.е. $p_1 = 0$, соответственно и доля (частота) таких особей p=0. Доля и ее

ошибка в таком случае вычисляются методом Ван-дер-Вандена. Доля при этом методе равна $p = \frac{(p_1 + 1) \cdot 100}{n + 2}$, а ошибка $m_p = \sqrt{p(100 - p) / n + 3}$.

Например, среди 30 школьников не оказалось ни одного, давшего положительную реакцию Шика. Так как p_1=0, то $p = \frac{(0 + 1) \cdot 100}{30 + 2}$=3,1%, $m_p = \sqrt{3{,}1 \cdot 96{,}9 / 30 + 3}$=3,2%. Вывод, который можно сделать на основании статистической обработки: несмотря на отрицательный результат, полученный на 30 школьниках, в других выборочных совокупностях можно ожидать положительной реакции с вероятностью около 3,2%.

Составляемые выборочные совокупности – объекты, дающие статистически непредсказуемые результаты, прежде всего с точки зрения ошибок. Чтобы устранить этот фактор, часто бывает необходимо рассчитать оптимальный объем выборки для определенной степени точности и вероятности. В случае качественной вариации пользуются формулой $n = t^2 \left[\frac{p(1 - p)}{\Delta^2} \right]$, где Δ^2 – требуемая точность опыта, а t – величина критерия достоверности для определенного уровня значимости.

Допустим, требуется определить размеры выборки для установления в популяции доли особей женского пола со степенью точности не менее чем 0,02 и с вероятностью 0,95. Для такой вероятности t=2, Δ=0,02, а частота берется максимальная для произведения p(p-1)=0,25. Для этого случая p=0,5. При всех прочих значениях p, произведение будет меньше, чем 0,25. Находим оптимальный размер выборки:

$$n = 2^2 \left[\frac{0{,}25}{0{,}02^2} \right] = 2500.$$

Возможные границы, в пределах которых находится значение доли для генеральной совокупности (P_0), определяются по формуле $p - tm_p < P_0 < p - tm_p$, где t – нормированное отклонение, т.е. величина, в пределах которой лежит значение доли генеральной совокупности, вычисленной по выборочной.

Например, с помощью реакции Шика выясняли иммунитет детей по отношению к дифтерии. Всего проверили 1600 детей, из них у 10% реакция была положительной. Спрашивается: какую долю восприимчивых к дифтерии детей можно ожидать в генеральной совокупности по результатам изучения выборочной? В каких доверительных пределах может находиться это значение в генеральной совокупности?

Для этого определяем стандартную ошибку по формуле $m_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{10 \cdot 90}{1600}}$ =0,75, значения t определяются по стандартным

таблицам для разных уровней значимости и числа степеней свободы. Так, для P=0,05 и k=∞ t=1,96. Тогда доверительные границы доли генеральной совокупности равны 10-1,96·0,75< P_0< 10+1,96·0,75, т.е. доверительные границы для процента детей, дающих положительную реакцию на дифтерию, в генеральной совокупности 8,53 – 11,47.

4. Для определения достоверности разницы между двумя показателями нужно знать разницу между сравниваемыми показателями d, ошибку этой разницы m_d, тогда достоверность определяется отношением $t=d/m_d$.

Для альтернативных признаков, когда сравниваются доли из одной выборочной совокупности, пользуются формулами: $d=p_x-p_y$; $m_d=\sqrt{m_{p_x}^2+m_{p_y}^2}$. Например, из 28 обезьян 16 было заражено вирусом А и 12 вирусом В. В первой группе заболели 4 обезьяны, а во второй – 6. Случайна ли разница в степени заражения?

$P_x=4/16=0,25$ $m_{p_x}=\sqrt{\frac{0,25\cdot 0,75}{16}}=0,11$

$p_y=6/12=0,5$ $m_{p_y}=\sqrt{\frac{0,5\cdot 0,5}{12}}=0,14$

$d=|0,25-0,5|=0,25$ $m_d=\sqrt{0,11^2+0,14^2}=0,18$ $t=0,25/0,18=1,4$. По таблице определяем уровень вероятности, соответствующий полученному t: P=0,84. Разница недостоверна.

Если доли определяются на разных совокупностях, процедура проверки достоверности разницы между разными группами иная. Отличие состоит в том, что при вычислении ошибки разницы сравниваемых показателей нужно исходить из вероятностной доли признака не в изучаемых выборках, а в генеральной совокупности: $m_d=\sqrt{pq(\frac{1}{n_x}+\frac{1}{n_y})}$, где $p=\frac{p_{1x}+p_{1y}}{n_x+n_y}$ (в числителе сумма абсолютных численностей особей с изучаемым признаком в обеих совокупностях). Например, при сравнении двух стад численностью 284 и 50 животных соответственно обнаружили, что в первом стаде 83 особи проявили реакцию на туберкулез, а во втором – 6 особей.

$p_x=83/284=0,29$ $p_y=6/50=0,12$ $p=\frac{83+6}{284+50}=0,27$

$m_d=\sqrt{0,27\cdot 0,73(\frac{1}{284}+\frac{1}{50})}=0,068$ $t=0,17/0,068=2,52$, такое значение t соответствует доверительной вероятности 0,988, что позволяет сделать вывод о достоверности разницы в реакции на туберкулез двух стад животных.

5. Существует несколько способов установления зависимости между качественными признаками. В случае альтернативной вариации выясняется вопрос, встречается ли совпадение присутствия обоих качественных признаков или, наоборот, отсутствие их чаще, чем это должно быть по случайным причинам. Классами 0 и 1 обозначаются либо два разных признака, либо отсутствие и присутствие их. Корреляционная решетка имеет следующий вид:

y	x		Σ
	0	1	
0	a	b	a+b
1	c	d	c+d
Σ	a+c	b+d	

Коэффициент корреляции в этом случае вычисляется по формуле $r = \frac{|ad - bc| - \frac{n}{2}}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$. Например, исследуем наличие связи между окраской шерсти и цветом глаз у 100 кроликов.

Цвет шерсти	Цвет глаз		Σ
	красные	не красные	
белая	29	11	40
окрашенная	1	59	60
Σ	30	70	100

Коэффициент корреляции равен $r = \frac{29 \cdot 59 - 11 \cdot 1}{\sqrt{30 \cdot 70 \cdot 40 \cdot 60}}$ =0,76. Ошибка коэффициента корреляции вычисляется по формуле $m_r = \frac{1 - r^2}{\sqrt{100}}$ =0,047. Достоверность коэффициента корреляции не вызывает сомнений. Изучение взаимосвязи между качественными признаками может также проводиться с помощью критерия соответствия хи-квадрат.

Лекция № 10

ДИСПЕРСИОННЫЙ АНАЛИЗ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

1. *Однофакторные комплексы.*
2. *Двухфакторные равномерные комплексы.*
3. *Двухфакторные неравномерные комплексы.*

Дисперсионный анализ качественных признаков проводится по тем же схемам, которые применяются и в отношении количественных признаков. Разница заключается лишь в том, что количественные признаки характеризуются средними показателями, тогда как качественные признаки выражаются долями, т.е. относительными частотами. Дисперсионные комплексы качественных признаков состоят из отдельных градаций, число которых может быть различным, различными бывают и числа вариант в отдельных градациях комплекса. Поэтому дисперсионные комплексы качественных признаков могут быть не только ортогональными, т.е. равномерными и пропорциональными, но и неортогональными. Они делятся на однофакторные и многофакторные.

1. При однофакторном дисперсионном анализе используются следующие рабочие формулы для определения числа степеней свободы и расчета сумм квадратов отклонений в зависимости от того, в каких единицах меры учитывается результативный признак – в долях от единицы или в процентах.

1. Межгрупповая или факториальная сумма квадратов отклонений: $D_x = \sum mp - \frac{(\sum m)^2}{N}$ или в % $D_x = \sum \frac{(p\%)^2}{100} - \frac{(\sum p\%)^2}{100a}$

2. Общая: $D_y = \sum m - \frac{(\sum m)^2}{N}$ или в % $D_y = \sum p\% - \frac{(\sum p\%)^2}{100a}$

3. Остаточная: $D_z = \sum m - \sum mp$ или $D_z = D_y - D_x$, где

m – число вариант, обладающих данным признаком;

p=m/n – относительная частота или доля вариант, имеющих данный признак;

p% – доля вариант с данным признаком, выраженная в %;

N – общее число наблюдений или объем дисперсионного комплекса;

a – число градаций фактора A;

n – число наблюдений в отдельных градациях фактора A.

Число степеней свободы определяется по следующим формулам:

1) для общей вариации: N – 1;

2) для межгрупповой вариации: a – 1;

3) для остаточной вариации: N – a.

Рассмотрим пример анализа однофакторного дисперсионного комплекса. У детей разного возраста анализировался уровень эритроцитарных аутоантител. Данные, характеризующие изменчивость уровня Т-агглютинина в зависимости от возраста детей, представлены в таблице. Нужно установить, достоверны ли различия, наблюдаемые между здоровыми детьми и реконвалесцентными по титру антител.

Показатели	Возраст в годах					
	2	3	4	5	6-7	8-9
Всего обследовано	9	13	20	9	17	12
Из них здоровых	4	4	6	2	12	8
Реконвалесцентов	5	9	14	7	5	4

1. Чтобы определить суммы квадратов отклонений, нужно рассчитать вспомогательные величины Σm и m^2/n. В данном случае m – это дети, обладающие признаком, т.е. реконвалесценты. Рассчитанные величины заносятся во вспомогательную таблицу.

Показатели	Возраст в годах						Сумма
	2	3	4	5	6-7	8-9	
n	9	13	20	9	17	12	80
m	5	9	14	7	5	4	44
m^2	5^2=25	81	196	49	25	16	-
m^2/n	2,78	6,23	9,8	5,45	1,47	1,33	27,06

2. Определяют суммы квадратов отклонений:

общую: $D_y = \sum m - \frac{(\sum m)^2}{N}$ = 44 – 44^2/80=19,8;

факториальную: $D_x = \sum mp - \frac{(\sum m)^2}{N} = D_x = \sum m\frac{m}{n} - \frac{(\sum m)^2}{N}$ =2,9;

Остаточную: $D_z = D_y - D_x$ =19,8-2,9=16,9;

4) определяем число степеней свободы: $к_y$ = N – 1 = 80 – 1 = 79; $к_x$ = a – 1 = 6 – 1 = 5; $к_z$ = N – a = 80 – 6 = 74;

5) определяем значения дисперсий:

$s_x^2 = \frac{D_x}{k_x}$ =2,9/5=0,58 $s_y^2 = \frac{D_y}{k_y}$ =19,8/79=0,25 $s_z^2 = \frac{D_z}{k_z}$ =16,9/74=0,23;

6) определяем значение критерия достоверности Фишера $F = \frac{s_x^2}{s_z^2}$=0,58/0,23=2,5. Стандартное значение этого показателя для данного числа степеней свободы и уровня значимости 0,05 равно 2,3. Следовательно, нулевая гипотеза отвергается, можно с уверенностью 95% утверждать, что разница между здоровыми и реконвалесцентными детьми по уровню Т-агглютинина достоверна;

7) определяем, влияет ли миксовирусная инфекция на уровень Т-агглютинина. Для этого находим отношение факториальной и общей суммы квадратов отклонений: $\eta_x^2 = \frac{D_x}{D_y}$ = 2,9/19,8=0,146. Достоверность этого статистического показателя проверяется делением его значения на статистическую ошибку. Ошибка определяется по формуле $m\eta_x^2 = (1-\eta_x^2)\frac{a-1}{N-a}$ = (1-0,146)·((6-1)/(80-6))=0,058 (5,8%). Критерий достоверности $F_ф$= 0,146/0,058=2,5. Для числа степеней свободы 5 и 74 и уровня значимости 0,05 стандартное значение критерия равно 2,3. Следовательно, можно утверждать, что инфекция влияет на уровень Т-антител, хотя и не очень значительно (5,8%).

Другой пример. По данным некоторых исследований, между разными видами синиц существуют различия по их двигательной активности. Результаты исследований представлены в таблице.

Показатель активности	Виды синиц				Сумма прыжков
	большая	лазоревка	московка	длиннохвостая	
Среднее число прыжков за час	2527	3690	5465	2401	14083
p%	(14083/2527)·100=18	26	39	17	100
p^2	324	676	1521	289	-
$p^2/100$	3,24	6,76	15,21	2,89	28,1

Необходимо установить, достоверно ли отличаются виды синиц по двигательной активности. Выразим абсолютные показатели в процентах и подвергнем их дисперсионному анализу. Результаты внесем в таблицу.

1) Определяем сумму квадратов отклонений:

общую – $D_y = \sum p\% - \frac{(\sum p\%)^2}{100a}$=100-$100^2$/4·100=75;

факториальную – $D_x = \sum \frac{(p\%)^2}{100} - \frac{(\sum p\%)^2}{100a}$ = 28,1-25=3,1;

остаточную – $D_z = D_y - D_x$ =75-3,1=71,9.

2) определяем число степеней свободы: $к_y$=N-1=4·100%-1=399; $к_x$=a-1=4-1=3; $к_z$=N-a=400-4=396.

3) рассчитываем дисперсии:

$s_x^2 = \frac{D_x}{k_x}$ =3,1/3=1,03 $s_y^2 = \frac{D_y}{k_y}$ =75/399=0,19 $s_z^2 = \frac{D_z}{k_z}$ =71,9/396=0,18.

4) Полученные данные заносим в таблицу и рассчитываем критерий Фишера по отношению факториальной и остаточной дисперсий.

Вариация	Степени свободы	Сумма квадратов отклонений	Дисперсия	$F_ф$	F_{st} (0,05)
Факториальная	3	3,1	1,03	5,7	2,6
Остаточная	396	71,9	0,18	1	-
Общая	399	75	0,19	-	-

Так как $F_ф > F_{st}$, нулевая гипотеза отвергается, разница в двигательной активности синиц разных видов признается статистически достоверной. Остается определить, велика ли степень влияния видовой принадлежности синиц на их двигательную активность. $\eta_x^2 = \frac{D_x}{D_y}$ =3,1/75=0,043 (4,3%). Ошибка показателя $m\eta_x^2 = (1-\eta_x^2)\frac{a-1}{N-a}$ = = (1 – 0,043)·((4 – 1)/(400 – 4)) = 0,007. Критерий достоверности $F_ф$ = = 0,043/0,007 = 6,17 >F_{st} =3,8. Следовательно, можно считать, что видовая принадлежность достоверно влияет на активность синиц, хотя и незначительно (4,3%).

2. При анализе двухфакторных равномерных комплексов используются следующие рабочие формулы для определения сумм квадратов отклонения и числа степеней свободы:

общая: $D_y = \sum m - \frac{(\sum m)^2}{N}$; факториальная (по сочетанию градаций факторов): $D_x = \sum m \frac{m}{n} - \frac{(\sum m)^2}{N}$; остаточная: $D_z = \sum m - \sum \frac{m^2}{n}$; факториальная A – $D_A = \sum \frac{(\sum m_A)^2}{n_A} - \frac{(\sum m)^2}{N}$; факториальная B – $D_B = \sum \frac{(\sum m_B)^2}{n_B} - \frac{(\sum m)^2}{N}$ и взаимодействия AB – $D_{AB} = D_X - D_A - D_B$.

Число степеней свободы и дисперсии определяются общим для количественных и качественных признаков при двухфакторном анализе способом.

Рассмотрим пример. При выборочном обследовании учащихся старших классов городских и сельских школ были обнаружены различные аномалии зрения, которые распределились следующим образом.

Показатели	A_1 – мальчики		A_2 – девочки		Сумма
	B_1-городские	B_2-сельские	B_1-городские	B_2-сельские	
Всего обследовано	25	25	25	25	N=100
Число с аномалиями (m)	3	2	8	2	Σm=15
p=m/n	0,12	0,08	0,32	0,08	-
m^2	9	4	64	4	-
m^2/n	0,36	0,16	2,56	0,16	Σ (m^2/n)=3,24

Судя по этим данным, доля аномалий выше у городских детей. Чтобы утверждать это с достаточной степенью достоверности, необходим дисперсионный анализ. В данном случае мы имеем дело с двумя факторами A и B. Находим суммы квадратов отклонений:

общая: $D_y = \sum m - \frac{(\sum m)^2}{N}$ = 15-15^2/100=12,75,

факториальная (по сочетанию градаций факторов): $D_x = \sum m \frac{m}{n} - \frac{(\sum m)^2}{N}$ =3,24-2,25=0,99,

остаточная: $D_z = \sum m - \sum \frac{m^2}{n}$ =15-3,24=11,76.

Для определения сумм квадратов отклонений по факторам необходимо рассчитать вспомогательные величины $\sum \frac{(\sum m_A)^2}{n_A}$ и $\sum \frac{(\sum m_B)^2}{n_B}$.

Для расчета можно использовать вспомогательную таблицу.

Градация факторов	n_i	Σm_i	$(\Sigma m_i)^2$	$(\Sigma m_i)^2/n_i$	$\Sigma m_i/n_i$=p
A_1	50	5	25	0,5	0,1
A_2	50	10	100	2	0,2
Σ по A	100	-	-	$\sum \frac{(\sum m_A)^2}{n_A}$ =2,5	-
B_1	50	11	121	2,45	0,22
B_2	50	4	16	0,32	0,08
Σ по B	100	-	-	$\sum \frac{(\sum m_B)^2}{n_B}$ =2,77	-

Находим суммы квадратов отклонений по факторам A и B и AB:

$D_A = \sum \frac{(\sum m_A)^2}{n_A} - \frac{(\sum m)^2}{N}$ =2,5-2,25=0,25; $D_B = \sum \frac{(\sum m_B)^2}{n_B} - \frac{(\sum m)^2}{N}$ =

= 2,77 – 2,25 = 0,52;

$D_{AB} = D_X - D_A - D_B$ = 0,99 – 0,25 – 0,52 = 0,22.

Определяем числа степеней свободы: $к_y$=N-1=1000-1=99; $к_x$=ab-1=2×2-1=3; $к_z$=N-ab=100-4=96; $к_A$=a-1=2-1=1; $к_B$=b-1=2-1=1; $к_{AB}$= $к_A$× $к_B$=1.

Находим значения дисперсий:

общая – $s_y^2 = D_y / k_y$ s_y^2 =12,75/99=0,12,

по фактору А – $s_A^2 = D_A / k_A$ s_A^2 =0,25/1=0,25,

по фактору В – $s_B^2 = D_B / k_B$ s_B^2 =0,52/1=0,52,

по взаимодействию факторов АВ – $s_{AB}^2 = D_{AB} / k_{AB}$ s_{AB}^2 =0,22/1=0,22,

остаточная (отражающая действие на результативный признак неконтролируемых (случайных) факторов) $s_Z^2 = D_Z / k_Z$ s_Z^2 =11,76/96=0,12,

по градациям фактора $s_x^2 = D_x / k_x$ =0,99/3=0,33.

Сводим полученные данные в таблицу и рассчитываем критерии достоверности.

Вариация	Степени свободы	Сумма квадратов отклонений	Дисперсия	$F_ф$	F_{st} (0,05)
Факториальная	3	0,99	0,33	2,8	2,7
По фактору А	1	0,25	0,25	2,1	3,9
По фактору В	1	0,52	0,52	4,3	3,9
Совместно АВ	1	0,22	0,22	1,8	3,9
Остаточная	96	11,76	0,12	1	-
Общая	99	12,75	0,12	-	-

Из таблицы видно, что с вероятностью 0,95 нулевая гипотеза отвергается только в отношении фактора В и по взаимодействию градаций факторов. Следовательно, можно утверждать, что разница по числу аномалий зрения между учениками сельских и городских школ не случайна. Разница же между этим признаком среди мальчиков и девочек достоверно не подтверждена. Определим силу влияния указанных факторов на признак.

$\eta_B^2 = \frac{D_B}{D_y}$ =0,52/12,75=0,04 (4%) $\eta_x^2 = \frac{D_x}{D_y}$ =0,99/12,75=0,08 (8%).

Ошибки репрезентативности этих показателей $m\eta_B^2 = (1-\eta_B^2)\frac{k_b}{k_z}$ = = (1 – 0,04)1/96 = 0,01 и $m\eta_x^2 = (1-\eta_x^2)\frac{k_x}{k_z}$ = (1 – 0,08)3/96 = 0,03. Критерии достоверности для уровня значимости 0,05: F_B= 0,04/0,01 = 4,0 > F_{st}=3,9 и F_A=0,08/0,03=2,7= F_{st}=2,7. Из этого следует, что хотя сила влияния и слабая (4% и 8% соответственно), но достоверная.

3. В неравномерных и непропорциональных комплексах не выполняется равенство $D_x=D_A + D_B + D_{AB}$. В связи с этим при вычислении сумм квадратов отклонений следует учитывать величину поправки $e=D_x/D'_x$.

Рассмотрим пример. Испытывалось влияние чистопородного и смешанного осеменения на оплодотворяемость крольчих. Опыт проводился на двух разнопородных группах животных в разных вариантах. Результаты приведены в таблице.

Варианты осеменения А	Число осеменений	Количество окролов в группах В	
		первая	вторая
Чистое осеменение 1	15	3	2
Чистое осеменение 2	9	7	8
Смешанное осеменение 1	10	7	5
Смешанное осеменение 2	11	7	6

Нужно выяснить эффективность оплодотворения животных разными дозами эякулята (фактор В) в зависимости от породных свойств (фактор А), число градаций фактора А – 2, число градаций фактора В – 4. Сформируем таблицу для расчета вспомогательных данных.

Показатели	A_1				A_2				Сумма
	B_1	B_2	B_3	B_4	B_1	B_2	B_3	B_4	
n	15	9	10	11	15	9	10	11	N=90
m	3	7	7	7	2	8	5	6	Σm=46
m^2	9	49	49	49	4	64	25	36	
m^2/n	0,6	5,44	4,9	4,45	0,27	7,11	2,5	3,27	28,54
m/n=p	0,2	0,78	0,7	0,64	0,13	0,9	0,5	0,55	4,4
p^2	0,04	0,61	0,49	0,41	0,02	0,81	0,25	0,3	2,93

Определяем суммы квадратов отклонений:

общая: $D_y = \sum m - \frac{(\sum m)^2}{N}$ =46-46^2/90=22,5,

по сочетанию градаций: $D_x = \sum \frac{m^2}{n} - \frac{(\sum m)^2}{N}$ =28,54-23,51=5,03,

остаточную: $D_z = D_y - D_x$ =22,5-5,0=17,5.

Далее необходимо определить неисправленные суммы квадратов D'_x, D'_A, D'_B, и D'_{AB}, а затем, найдя величину поправочного коэффициента, исправить их на это значение. Неисправленные суммы квадратов для качественных признаков определяются по следующим формулам:

$$D'_x = N\left(\frac{\sum p^2}{ab} - M^2\right) \quad D'_A = N\left(\frac{\sum M_A{}^2}{a} - M^2\right) \quad D'_B = N\left(\frac{\sum M_B{}^2}{b} - M^2\right),$$

где

N – общее число наблюдений всего комплекса; p=m/n; a – число градаций фактора A; b – число градаций фактора B; $M = \frac{\sum p}{ab}$; $M_A = \frac{\sum p_A}{b}$; $M_B = \frac{\sum p_B}{a}$.

Из таблицы вспомогательных данных видно, что Σp=4,4; Σp^2=2,934 a=2 и b=4. Определяем среднее арифметическое из суммы долей по градациям факторов A и B:

$M = \frac{\sum p}{ab}$=4,4/2×4=0,55 M^2=0,3, рассчитаем частные средние по факторам A и B отдельно и внесем во вспомогательную таблицу.

Градации факторов	Число градаций	Σp	$\Sigma p/i = \overline{x_i}$	M_i^2
A_1	b=4	2,32 (сумма всех p по A_1)	0,58 (2,32/4)	0,3364
A_2	b=4	2,08	0,52	0,2704
Σ по A				$\sum M_A^2$=0,61
B_1	a=2	0,33 (B_1 по A_1+ B_2 по A_2)	0,165	0,0272
B_2	a=2	1,68	0,84	0,7056
B_3	a=2	1,2	0,6	0,3600
B_4	a=2	1,19	0,595	0,3540
Σ по B				$\sum M_B^2$=1,45

Рассчитав вспомогательные величины, находим неисправленные суммы квадратов отклонений по факторам А и В и их взаимодействию АВ:

$$D'_A = N\left(\frac{\sum M_A^{\,2}}{a} - M^2\right) = 90(0{,}61/2 - 0{,}3) = 0{,}45,$$

$$D'_B = N\left(\frac{\sum M_B^{\,2}}{b} - M^2\right) = 90(1{,}45/4 - 0{,}3) = 5{,}45,$$

$$D'_x = N\left(\frac{\sum p^2}{ab} - \overline{x^2}\right) = 90(2{,}93/8 - 0{,}3) = 6{,}3,$$

$$D'_{AB} = D'_x - D'_A - D'_B = 6{,}3 - 0{,}45 - 5{,}4 = 0{,}45.$$

Находим величину поправочного коэффициента: $e = D_x/D'_x = 5{,}03/6{,}3 = 0{,}8$, исправляем факториальные суммы квадратов отклонений: $D_A = 0{,}45 \times 0{,}8 = 0{,}36$; $D_B = 4{,}36$; $D_{AB} = 0{,}36$.

Определяем числа степеней свободы: $к_y = N-1 = 90-1 = 89$; $к_x = ab-1 = 2\times4-1 = 7$; $к_z = N-ab = 90-8 = 82$; $к_A = a-1 = 2-1 = 1$; $к_B = b-1 = 4-1 = 3$; $к_{AB} = к_A \times к_B = 1\times3 = 3$.

Находим средние квадраты отклонений (дисперсии):

по фактору А – $s_A^2 = D_A / k_A$ $s_A^2 = 0{,}36/1 = 0{,}36$,

по фактору В – $s_B^2 = D_B / k_B$ $s_B^2 = 4{,}36/3 = 1{,}45$,

по взаимодействию факторов АВ – $s_{AB}^2 = D_{AB} / k_{AB}$ $s_{AB}^2 = 0{,}36/3 = 0{,}12$,

остаточная (отражающая действие на результативный признак неконтролируемых, случайных, факторов) $s_Z^2 = D_Z / k_Z$ $s_Z^2 = 17{,}5/82 = 0{,}21$

по градациям фактора $s_x^2 = D_x / k_x = 6{,}3/7 = 0{,}9$.

Сводим полученные данные в таблицу дисперсионного анализа:

Вариация	Степени свободы	Сумма квадратов отклонений	Дисперсия	$F_ф$	F_{st} (0,05)
Факториальная	7	6,3	0,9	4,3	2,1
По фактору А	1	0,20,365	0,36	1,7	4,7
По фактору В	3	0,524,36	1,45	6,9	2,7
Совместно АВ	3	0,220,36	0,12	1,0	-
Остаточная	82	11,7617,5	0,21	1	-
Общая	89	12,7522,5			-

Нулевая гипотеза отвергается с вероятностью 0,95 в отношении дозы эякулята (фактор В) и по градациям факторов. Что касается по-

родных свойств животных, то их влияние на оплодотворяемость не подтвердилось.

Определим силу влияния контролируемых факторов:

$$\eta_B^2 = \frac{D_B}{D_y} = 4{,}36/22{,}5 = 0{,}19\ (19\%),\ \eta_x^2 = \frac{D_x}{D_y} = 6{,}3/22{,}5 = 0{,}28\ (28\%).$$

Ошибки репрезентативности этих показателей равняются:

$$m\eta_B^2 = (1-\eta_B^2)\frac{k_b}{k_z} = (1\text{-}0{,}19)\times 3/82 = 0{,}03\ \ m\eta_x^2 = (1-\eta_x^2)\frac{k_x}{k_z} = (1\text{-}0{,}28)\times 7/82 = 0{,}06.$$

Критерии достоверности для уровня значимости 0,05: F_B=0,19/0,03=6,0>F_{st}=4,0 и F_A=0,28/0,06=4,7>F_{st}=2,9 поскольку в обоих случаях $F_ф$>F_{st}, в достоверности найденных показателей сомневаться не приходится.

Лекция № 11

КЛАСТЕРНЫЙ АНАЛИЗ

1. Основная цель кластерного анализа.
2. Области применения и задачи.
3. Выполнение кластерного анализа.

1. Термин *«кластерный анализ»* (впервые ввел Tryon, 1939) в действительности включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как *организовать* наблюдаемые данные в наглядные структуры, т.е. развернуть таксономии. Например, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. **Кластерный анализ предназначен для разбиения совокупности объектов на однородные группы (кластеры или классы).** Существует около 100 разных алгоритмов кластеризации, однако наиболее часто используемые: иерархический кластерный анализ и кластеризация методов k-средних.

Где применяется кластерный анализ? В маркетинге это сегментация конкурентов и потребителей. В менеджменте: разбиение персонала на различные по уровню мотивации группы, классификация поставщиков, выявление схожих производственных ситуаций, при которых возникает брак. В медицине – классификация симптомов, пациен-

тов, препаратов. В социологии – разбиение респондентов на однородные группы. По сути кластерный анализ хорошо зарекомендовал себя во всех сферах жизнедеятельности человека. Прелесть данного метода состоит в том, что он работает даже тогда, когда данных мало и не выполняются требования нормальности распределений случайных величин и другие требования классических методов статистического анализа.

2. Техника кластеризации применяется в самых разнообразных областях. Хартиган дал прекрасный обзор многих опубликованных исследований, содержащих результаты, полученные методами кластерного анализа. Например, в области медицины кластеризация заболеваний, лечения заболеваний или симптомов заболеваний приводит к широко используемым таксономиям. В археологии с помощью кластерного анализа исследователи пытаются установить таксономии каменных орудий, похоронных объектов и т.д. Известны широкие применения кластерного анализа в маркетинговых исследованиях. В общем, всякий раз, когда необходимо классифицировать «горы» информации к пригодным для дальнейшей обработки группам, кластерный анализ оказывается весьма полезным и эффективным.

Задачи кластерного анализа можно объединить в следующие группы:

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

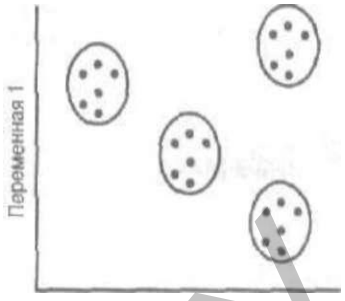
Как правило, при практическом использовании кластерного анализа одновременно решается несколько из указанных задач.

3. Этапы выполнения кластерного анализа представлены на рисунке.

1. Формулировка проблемы. Возможно, самая важная часть формулирования проблемы кластеризации – это выбор переменных, на основе которых проводят кластеризацию. Включение даже одной или двух посторонних (не имеющих отношение к группированию) переменных может исказить результаты кластеризации. Задача состоит в том, чтобы выбранный набор переменных смог описать сходство между объектами с точки зрения признаков, имеющих отношение к данной проблеме исследования.

На рисунке показана идеальная ситуация кластеризации, когда кластеры четко отделены друг от друга на основании различий двух

переменных: ориентация потребителей на чество товара (переменная 1), и чувствительность к цене (переменная 2). Следует отметить, что каждый потребитель попадает в один из кластеров, и перекрывающихся областей нет.



Назначение алгоритма объединения состоит в группировании объектов (например, потребителей) в достаточно большие кластеры, используя некоторую меру сходства или расстояние между объектами.

Рис. **Схема кластерного анализа.**

Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера:

– центр кластера – это среднее геометрическое место точек в пространстве переменных;

– радиус кластера – максимальное расстояние точек от центра кластера.

Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается наложение кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными:

– спорный объект – это объект, который по мере сходства может быть отнесен к нескольким кластерам;

– размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

2. Выбор способа измерения расстояния или меры сходства. Цель кластеризации – группирование схожих объектов. Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о «схожести» объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы). Поэтому для того чтобы оценить, насколько они похожи или непохожи, необходимо использовать некую единицу измерения. Наиболее распространенный метод заключается в том, чтобы в качестве такой меры использовать расстояние между двумя объектами. Объекты с меньшими расстояниями между собой больше похожи, чем объекты с большими расстояниями. Способов определения меры расстояния между кластерами, называемой еще мерой близости, существует несколько. Наиболее распространенный способ – вычисление евклидова расстояния между двумя точками i и j на плоскости, когда известны их координаты X и Y: $D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

Примечание. Чтобы узнать расстояние между двумя точками, надо взять разницу их координат по каждой оси, возвести ее в квадрат, сложить полученные значения для всех осей и извлечь квадратный корень из суммы.

Манхэттенское расстояние (расстояние городских кварталов), также называемое «хэмминговым» или «сити-блок» расстоянием.

Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния Евклида. Однако для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат.

Расстояние Чебышева. Это расстояние стоит использовать, когда необходимо определить два объекта как «различные», если они отличаются по какому-то одному измерению.

Процент несогласия. Это расстояние вычисляется, если данные являются категориальными.

Если переменные измерены в различных единицах, то единица измерения влияет на решение кластеризации. Эта проблема решается при помощи предварительной стандартизации переменных. Стандартизация (*standardization*) или нормирование (*normalization*) приводит

значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных.

Наиболее распространенный способ – деление исходных данных на среднеквадратичное отклонение соответствующих переменных. Наряду со стандартизацией переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов – специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных.

3. Выбор метода кластеризации. Методы кластерного анализа можно разделить на две группы:

– иерархические;

– неиерархические.

Каждая из групп включает множество подходов и алгоритмов.

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (*Agglomerative Nesting, AGNES*). Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров. В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы (*DIvisive ANAlysis, DIANA*). Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде дендрограммы показан на рисунке.

Типичным результатом такой кластеризации является *иерархическое дерево*. Рассмотрим горизонтальную древовидную диаграмму.

Иерархические алгоритмы связаны с построением дендрограмм (от греческого dendron – «дерево»), которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в гра-

фическом виде последовательность объединения (разделения) кластеров. Дендрограмма (*dendrogram*) – древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров. Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

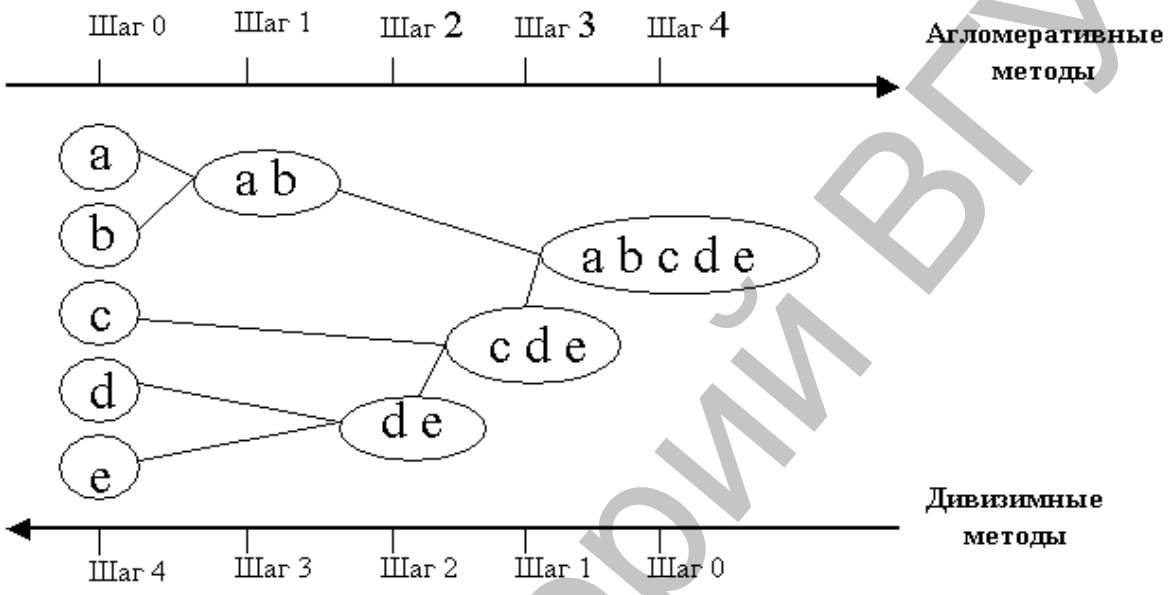


Рис. **Дендрограмма агломеративных и дивизимных методов.**

Существует много способов построения дендограмм. В дендограмме объекты могут располагаться вертикально или горизонтально. Пример вертикальной дендрограммы приведен на рис.

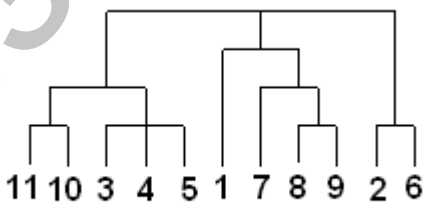


Рис. **Пример дендрограммы.**

Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. Мы видим, что на первом шаге каждое наблюдение представляет один кластер (вертикальная линия), на втором шаге наблюдаем объединение таких наблюдений: 11 и 10; 3, 4 и 5; 8 и 9; 2 и 6. На втором шаге продолжается объединение в кластеры: наблюдения 11, 10, 3, 4, 5 и 7, 8, 9. Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Диаграммы могут быть и горизонтальными, такие диаграммы читают слева направо. Вертикальные линии показывают кластеры, объединяемые вместе. Положение линии относительно шкалы расстояния показывает расстояния, при которых кластеры объединили.

4. Принятие решения о количестве кластеров. Иногда можно априорно определить это число. Однако в большинстве случаев число кластеров определяется в процессе агломерации/разделения множества объектов.

Здесь нет твердых правил, позволяющих быстро принять решение, но можно руководствоваться следующим:

a. при определении количества кластеров руководствуются теоретическими и практическими соображениями. Например, если цель кластеризации – выявление сегментов рынка, то менеджмент может захотеть получить конкретное число кластеров;

b. в иерархической кластеризации в качестве критерия можно использовать расстояния, при которых объединяют кластеры;

c. в неиерархической кластеризации чертят график зависимости отношения суммарной внутригрупповой дисперсии к межгрупповой дисперсии от числа кластеров. Точка, в которой наблюдается изгиб или резкий поворот, указывает на приемлемое количество кластеров. Увеличение числа кластеров за эту точку обычно нерезультативно;

d. относительные размеры кластеров должны быть достаточно выразительными: количество элементов в кластере не может быть очень маленьким (1, 2) или число элементов в кластерах сильно отличным друг от друга.

5. Интерпретация и профилирование кластеров. Интерпретация и профилирование кластеров включает проверку кластерных центроидов. Центроиды представляют средние значения объектов, содержащиеся в кластере по каждой из переменных. Они позволяют описывать каждый кластер, если присвоить ему номер или метку. Часто имеет смысл профилировать кластеры через переменные, которые не явились основанием для кластеризации.

6. Оценка надежности и достоверности. Имея несколько умозаключений, выведенных из кластерного анализа, не следует принимать никакого решения по кластеризации, не выполнив оценку надежности и достоверности этого решения. Формальные процедуры оценки надежности и достоверности решений кластеризации достаточно сложны и не всегда оправданы. Однако следующие процедуры обеспечат адекватную проверку качества кластерного анализа:

– выполняйте кластерный анализ на основании одних и тех же данных, но с использованием различных способов измерения расстояния. Сравните результаты, полученные на основе разных мер расстояния, чтобы определить, насколько совпадают полученные результаты;

– используйте разные методы кластерного анализа и сравните полученные результаты;

– разбейте данные на две равные части случайным образом. Выполните кластерный анализ отдельно для каждой половины. Сравните кластерные центроиды двух подвыборок;

– случайным образом удалите некоторые переменные. Выполните кластерный анализ по сокращенному набору переменных. Сравните результаты с полученными на основе полного набора переменных;

– в неиерархической кластеризации решение может зависеть от порядка случаев в наборе данных. Выполните анализ несколько раз, меняя порядок случаев, до получения стабильного решения.

Лекция № 12

МЕТОДИКА ПОЛЕВОГО ОПЫТА

1. Полевой опыт и его особенности. Основные понятия.
2. Виды полевых опытов. Условия проведения опыта.
3. Основные элементы методики полевого опыта.
4. Планирование полевого эксперимента.

1. В основе любого теоретического или экспериментального исследования лежит общий метод познания – индуктивно-дедуктивный, вскрывающий наиболее общие законы развития природы. Самыми общепринятыми приемами научного исследования являются *наблюдение и эксперимент*.

Наблюдение – это количественная или качественная регистрация интересующих исследователя сторон развития явления, констатация того или иного его состояния, признака или свойства.

Для наблюдения и регистрации тех или иных свойств, или состояний явления применяют разнообразные средства измерений, фиксирования результатов. Например, на метеостанциях регулярно фиксируются температуры воздуха, почвы, влажность воздуха, направление и сила ветра и др. Орнитологи, например, наблюдают за миграциями перелетных птиц, фиксируя время прилета, расстояния, преодолеваемые животными и др. Агротехники наблюдают за качеством почвы, временем наступления весны, всхожестью семян и т.п. Во всех этих случаях наблюдение дает исследователю качественную или количественную характеристику явления, но не вскрывает его сущности. В ряде случаев этого вполне достаточно для установления связи между отдельными признаками, свойствами или явлениями, однако чаще

всего наблюдение не является самостоятельным приемом исследования, а составляет важную часть более сложного метода – эксперимента, который иногда называют активным наблюдением.

Эксперимент, опыт – прием, при котором исследователь искусственно вызывает явления или изменяет условия так, чтобы выяснить сущность явления, происхождение, причинно-следственные связи предметов и явлений.

Опыт – ведущий метод исследования, включающий наблюдения, корреляции, учет измененных условий, учет и обработку результатов исследования. Главная особенность любого опыта – его *воспроизводимость*. Эксперимент – это гипотеза, ищущая проверки фактами, практикой.

Опыт имеет большие преимущества по сравнению с наблюдением:

1) экспериментатор может сам воссоздать нужное ему явление, не дожидаясь, когда оно наступит в природе. Вместе с тем, любой опыт – не бездумное манипулирование природными объектами. Любой эксперимент обязан проистекать из научно обоснованного предположения, возможности.

«Увлекающийся практикой без науки – словно кормчий, вступающий на корабль без компаса».

Леонардо да Винчи

2) в ходе опыта можно расчленять явления (анализ) и вновь объединять их (синтез), создавая надлежащие сопутствующие условия. Однако, планируя эксперимент, следует остерегаться разбрасывать внимание на лишнее, побочное, что отвлекает от решения конкретной проблемы, от поиска наиболее эффективного и логичного пути ее решения.

3) любой опыт можно воспроизвести, особенно если он лабораторный или является модельным, в котором не обязательно наличие объекта исследования. При этом сама возможность воспроизводимости опыта страхует исследователя от излишнего субъективизма, необъективного суждения, возможной ошибки. Неоднократно повторенный эксперимент может с достаточной степенью вероятности подтвердить выдвигаемую гипотезу, равно как и отвергнуть ее.

«Никаким количеством экспериментов нельзя доказать теорию; но достаточно одного эксперимента, чтобы ее опровергнуть».

Альберт Эйнштейн

Для проведения успешного исследования экспериментатору необходимо иметь в виду следующие важные особенности и условия. Даже тщательно спланированный опыт может не привести к ожидаемому результату с достаточной степенью точности.

«Невозможно прямое попадание эксперимента в узко определенную теоретическую мишень».

Имре Лакатос

Иногда можно получить даже обратный, не предполагаемый результат. В связи с этим иногда требуются пробные опыты, поставленные с целью уточнить схему эксперимента, условия проведения опыта, выбора объектов и т.п. С одной стороны, это доставляет определенные трудности, а с другой – эксперимент открывает возможности расширить предположение, гипотезу, область практического применения. Доказательств этому огромное количество. Большое число открытий в химии, физике были совершены побочно, при решении совершенно иных проблем и задач. Во-вторых, следует иметь в виду, что каждый результат можно трактовать двояко.

«Кошка, однажды присевшая на горячую печку, уже никогда не сядет на горячую печку – и хорошо сделает, но уже никогда не сядет и на холодную».

Марк Твен

Поэтому, планируя эксперимент исследователю необходимо учитывать такую возможность. Постановке опыта должен предшествовать предварительный мысленный эксперимент, требующий творческого воображения. Необходимо мысленно представить себе ход эксперимента, убрать все лишнее, спрогнозировать результат, а также возможные решения, если результат получается не тот, который планировался. Это предполагает постоянную мысленную работу над проблемой. Когда Ньютона спросили, как он сделал свои открытия, он ответил: я постоянно думал о них. Кроме того, экспериментатор должен воспитывать в себе привычку критически мыслить, то есть отсутствие чувства непреложности авторитета и догматизма, искать новые пути.

Большинство экспериментов – сравнительные исследования, предполагающие сопоставление эффектов опытных данных (вариант) со стандартными, не подвергавшимися опытному воздействию. Такие стандартные данные называют *контрольными, или контролем*. Контроль может быть пространственным и временным. Пространственный иногда называют абсолютным. Такой контроль закладывается вместе с опытом в одно и то же время при соблюдении одних и тех же условий, кроме испытуемого. Например, опытные данные по влиянию вносимых удобрений на урожайность растений необходимо сравнить с контрольным опытом по выращиванию растений без применения удобрений. Временной контроль является относительным. В этом случае опытные данные сравнивают с подобными, однако, полученными в разные временные периоды (годы, месяцы, дни и т.п.). Например, можно сравнивать число гнездящихся пар хищных птиц в данном году с данными десятилетней давности. Совокупность опытных и контрольных вариантов составляет схему опыта.

Важным элементом любого эксперимента является статистическая обработка опытных данных, позволяющая извлечь максимум ин-

формации из исходных данных, оценить, насколько существенны различия между результатами исследования, составить прогноз развития явления.

К полевому опыту предъявляют ряд требований:

1) типичность, или репрезентативность – соответствие условий его проведения естественным, природным, наблюдаемым. Это требование определяет и выбор объектов – характерных, типичных;

2) принцип единственного различия – при постановке полевых опытов необходимо соблюдать единство всех условий, кроме одного – изучаемого. Например, в опыте с удобрениями единственным различием между вариантами опыта и контроля будут дозы. Все остальные условия должны быть унифицированы: почва, температура, сорт, время посева, режим полива, подкормки и т.п.;

3) учет результатов исследования должен быть систематичным и синхронным. Все результаты фиксируются в опыте и контроле по заранее составленной схеме одновременно;

4) достоверность опыта – логически правильно построенная схема и методика проведения опыта, соответствие их поставленным задачам, правильный выбор объекта и условий проведения опыта. (*Цель – предполагаемый результат*). Этот принцип позволяет максимально устранять ошибки – расхождения между результатами выборочного наблюдения и истинным значением измеряемой величины. При избегании грубых и систематических ошибок, вызванных нарушением основных требований к полевому опыту, случайные ошибки можно учесть при проведении статистической обработки эмпирических данных.

2. Полевые опыты классифицируют по различным признакам. В зависимости от количества изучаемых факторов полевые опыты делят на 2 группы:

1) *однофакторные, или простые*. В данном случае исследуется один простой или сложный (составной) количественный фактор в нескольких градациях (например, влияние на один сорт разных доз удобрений) или сравнивается действие ряда качественных факторов (реакция разных сортов на одну и ту же дозу удобрений).

2) *многофакторные* – опыты, в которых одновременно изучается действие и устанавливается характер и величина взаимодействия двух и более факторов.

В зависимости от масштабов охвата объектов исследования полевые опыты могут быть *единичными* (если закладываются в отдельных пунктах, независимых друг от друга, по различным схемам) или *массовыми или географическими* (если по одной и той же схеме опыты одинакового содержания проводят в различных почвенно-географических условиях).

По длительности проведения опыты разделяют на краткосрочные, многолетние и длительные. К *краткосрочным* относят опыты продолжительностью от 3 до 10 лет. Они могут быть нестационарными и стационарными. Первые повторяют ежегодно по одной и той же схеме на разных участках в течение 3–4 лет. Вторые закладываются на стационарных участках и проводят непрерывно в течение 4–10 лет. К *многолетним* относят полевые опыты продолжительностью 10–50, к *длительным* – более 50 лет. Многолетние и длительные опыты незаменимы при изучении медленно протекающих явлений, например, процессов динамики температуры в глобальном масштабе, процессов миграций организмов. Многолетняя их повторность как бы «спрессовывает время». Часто выделяют так называемые *производственные опыты – комплексное, научно поставленное исследование, которое проводится не посредственно в производственных условиях и отвечает конкретным задачам материального производства, его постоянного развития и совершенствования* (апробация определенного типа севооборота в хозяйстве, внедрение новой технологии в производство, разработка альтернативного метода лечения и т.д.).

3. Под методикой полевого опыта подразумевают совокупность слагающих ее элементов: число вариантов, повторность, систему размещения повторностей, площадь пробных площадок, метод учета результатов, организация опыта во времени.

1. Число вариантов в схеме любого опыта – обычно заранее заданная величина, которая определяется содержанием опыта и его задачами. Число вариант не может повлиять на типичность опыта, но может существенно сказаться на ошибке, так как чем большим количеством вариант представлена выборка, тем ближе она к генеральной совокупности, а характеристики выборки – к характеристикам генеральной совокупности. В статистике существуют способы расчета оптимального количества вариантов с заданной величиной ошибки. Если нет объективной возможности увеличить объем выборки, следует увеличить число самих выборок.

2. Увеличением числа наблюдений (выборок) мы добиваемся повторяемости наших данных, полученных по одной выборке. Такое число одновременно исследуемых выборок называют *повторностью опыта*. Например, для исследования эффективности воздействия лекарственного препарата на здоровье людей мы создаем несколько опытных групп людей определенного количества и контрольную группу. Число опытных групп в данном случае – повторность опыта, а число людей в группе – число вариантов. Иногда испытания следует повторять в течение нескольких лет (например, для проверки эффекта привыкания людей к лекарственному препарату). Число лет испыта-

ний называют *повторностью опыта во времени*. При увеличении повторности заметно снижается ошибка опыта, особенно при увеличении повторности до 4–6-кратной; дальнейшее повышение повторности сопровождается менее значительным уменьшением ошибки. В биологических исследованиях принята 3-кратная повторность. Проведение опытов без повторностей допустимо в предварительных, рекогносцировочных и демонстративных опытах. Количество повторностей во времени зависит от изменчивости факторов в течение периода исследований. Часто опыты по изучению природных явлений приходится повторять в разные времена года, чтобы избежать ошибки, вызванной влиянием разных климатических условий. То же касается и многолетних исследований. Например, опыт заложенный летом одного года с определенными погодными условиями, может не дать воспроизводимого результата летом другого года, с другой погодой.

3. Наиболее часто используемым для организации повторного опыта методом является метод пробных площадок. Например, для изучения природной популяции подорожника большого на мезофильном лугу не обязательно исследовать все особи данного вида на территории всего луга. Для этой цели на площади всего луга закладывают несколько участков-квадратов, площадью от 0,225 до 1 м^2. Расположение этих площадок зависит от рельефа луга (квадраты закладывают в каждом элементе ландшафта), от целей опыта. Различают *сплошное* (когда все повторения объединены территориально) и *разбросанное* (когда повторности расположены хаотично, случайно) размещения организованных повторений (рисунок 1). Для отбора проб почвы часто используют метод «конверта»: с обследуемого элементарного участка взять пять образцов почвы так, что точки отбора проб расположены по углам квадрата со стороной 2–3 м, а также в точке пересечения диагоналей квадрата (рисунок 2).

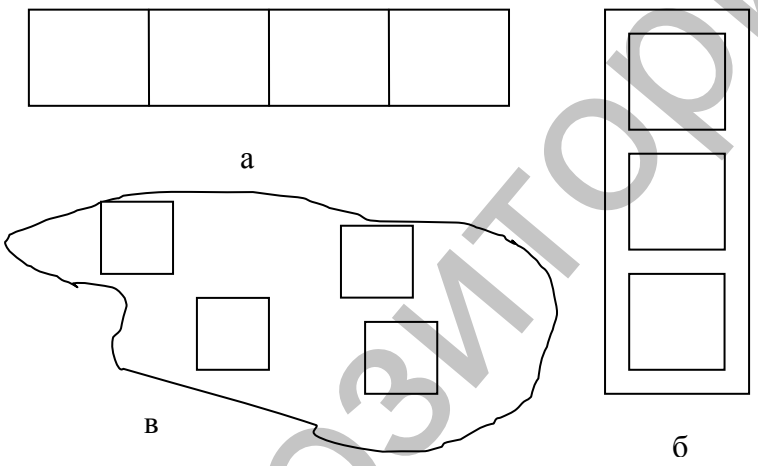


Рис. 1. **Способы размещения повторений:**
а, б – сплошное; в – разбросанное.

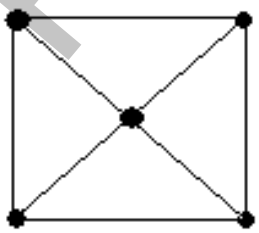


Рис. 2. **Схема взятия почвенных образцов методом «квадрата».**

4. Площадь пробных площадок зависит от типа сообщества, биологии исследуемого вида и целей исследования. Например, в лесу для изучения популяций древесных растений размер пробных площадок должен быть намного большим, чем при изучении травянистых. Это связано с тем, что плотность размещения деревьев меньшая и при маленьком размере учетной площадки не будет соблюдено условие оптимального числа вариант. Метод пробных площадок вполне применим при изучении донных организмов в водоемах, сидячих животных. При изучении популяций животных, ведущих подвижный и скрытный образ жизни, метод пробных площадок не оправдает себя. Наиболее удобен для таких целей маршрутный учет, или метод линейных трансект. Для этого закладываются не квадратные площадки, а маршруты определенной ширины (до 500 м) и длины (до нескольких километров), на всем протяжении которых изучается объект.

5. Размещение опытных и контрольных пробных площадок, а также вариант внутри площадок может осуществляться по определенной схеме. Выделяют три группы методов размещения: *стандартные, систематические и рендомизированные* (случайные).

Стандартные методы характеризуются более частым расположением контроля среди опытных вариант. В этом бывает необходимость, если условия очень неоднородны. Тогда для каждой варианты существует свой контроль. Статистической обработке подвергаются не только опытные, но и контрольные варианты (рисунок 3). Этот метод довольно громоздкий и нерациональный, поэтому используется в опытной работе нечасто.

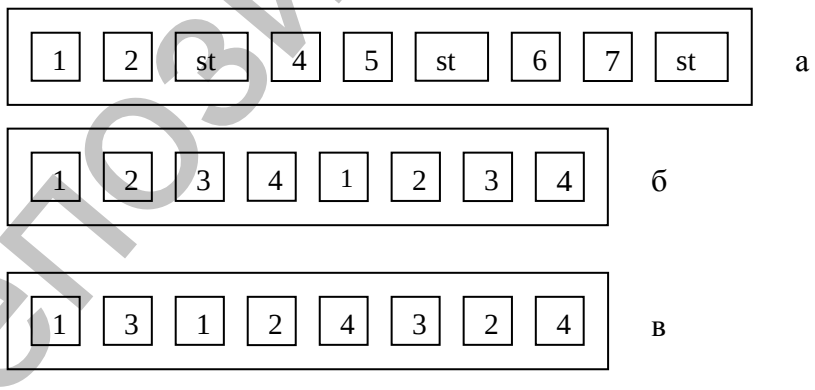


Рис. 3. **Размещение опытных и контрольных пробных площадок и вариант внутри площадок**
а – стандартное; б – систематическое;
в – рендомизированное.

Систематическое размещение вариант и площадок – это такое расположение опыта, когда порядок следования вариант в каждом повторении подчиняется определенной системе. Чаще всего применяют последовательный или шахматный способ размещения.

Описанные методы размещения имеют определенные недостатки: непредвиденные искажения по вариантам, ненадежность в статистической оценке ошибки опыта. В связи с этим большую популярность приобрел метод случайного, или рендомизированного расположения вариантов в эксперименте.

4. Планирование эксперимента включает несколько этапов: 1) подготовительные работы, 2) проведение полевых опытов, наблюдений и учетов; 3) обработка и обобщение полученных данных.

Подготовительный этап предполагает определение задачи и объектов исследования, изучение современного состояния вопроса, выдвижение рабочей гипотезы и ряда конкурирующих гипотез, разработки схемы эксперимента, выбор учетной площади и оптимальной структуры полевого опыта.

Необходимо четко сформулировать цель исследования, построить логическую модель изучаемого явления и правильно выбрать стратегию, которая определяет методы и приемы исследования.

Следующий этап планирования – разработка общей программы и методики исследований. Этот этап требует широкого знакомства с литературой и подготовительными материалами (ведомственные, картографические, нормативные и др.) и выдвижения рабочей гипотезы или ряда конкурирующих гипотез. Рабочая гипотеза служит отправным пунктом для составления схемы будущих опытов и разработки программы исследования. В программе указываются схемы опытов, основные элементы методики и техники эксперимента, наблюдения и учеты. После определения методики комплектуется необходимое оборудование и полевое снаряжение.

Далее, основываясь на сведениях, полученных в результате подготовки к полевым работам, исследователь должен подобрать наиболее типичные для района участки, удобные для полевых наблюдений и удовлетворяющие требованиям к полевому опыту. Для решения этой задачи часто требуются маршрутные рекогносцировочные исследования. Они проводятся для первичного ознакомления с природными условиями и растительным покровом района и носят описательный характер. После этого проводятся детально-маршрутные наблюдения, цель которых – дать более полную характеристику района исследований. Такая характеристика может быть получена путем закладки серии маршрутов с систематической фиксацией растительности и подробным описанием фитоценозов. Результатом таких исследований является составление геоботанической карты выбранного района исследований. При этом особое внимание уделяется описанию местоположения района (область, район, лесорастительная зона, подзона и т.д.), характеризуется преобладающий тип растительности, детально анализируется климат, рельеф и почвы.

Подбор объектов исследования – важная процедура, определяющая во многом успех работы. Объекты должны быть наиболее типичными, их количество определяется программой и реально выполнимым объемом работ. Можно подбирать разные объекты, однако, составляющие один динамический ряд: например, отношение фитоценозов к трофности почв, к увлажнению и т.д. Количество пробных площадей по элементам экологического ряда может быть разным, но должно обеспечить сравнимость исследований. Подбор участка для закладки пробных площадей зависит от возможностей исследователя, целей и объектов. Пробная площадка отбивается квадратной или прямоугольной формы. Стороны площадки ограничиваются визирами – столюами 10–12 см в диаметре. Размер пробной площади для описания лесных фитоценозов обычно составляет от 400 м^2 до 2000 м^2, а размеры пробных площадей для изучения кустарникового фитоценоза – от 200 м^2 до 2000 м^2 в зависимости от видового богатства, густоты и характера размещения кустарников на участке. При изучении травянистой растительности лугов принято закладывать пробные площади размером 100 м^2. Пробные площади при необходимости разбивают на учетные площадки, размер которых колеблется от 1 м^2 до десятков метров. Учетные площадки должны быть идентичны по условиям местообитаний пробным площадям и необходимы для установления ряда признаков ценоза (встречаемости и обилия видов, биологической продуктивности, возрастного и полового состава популяций и др.). Необходимая повторность будущего опыта при установленной площади определяется в основном характером территориальной изменчивости, вариабельностью признаков объектов и заданной величиной ошибки опыта. Большую часть простых однофакторных и маловариантных многофакторных опытов проводят при 4–6-кратной повторности. Многовариантные факториальные опыты проводят в 2–3-кратной повторности с группировкой вариант в блоки.

Второй этап – проведение наблюдений и учетов. В зависимости от задач исследования могут преобладать полевые или лабораторные наблюдения. При этом число наблюдений не должно быть большим. Можно наблюдать за бесчисленным количеством объектов и явлений, в то время как для понимания сущности явлений хватило бы и нескольких из них. Поэтому важным условием наблюдений является *целенаправленность*. Опыт должен сопровождаться не стандартным набором наблюдений, а теми наблюдениями, без которых нельзя изучить явление.

Сроки и периодичность наблюдений и учетов определяются при составлении программы исследований и детерминируются целями и возможностями.

Семинарское занятие № 1

КЛАССИФИКАЦИЯ БИОЛОГИЧЕСКИХ ПРИЗНАКОВ. ГРУППИРОВКА РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ. ПОКАЗАТЕЛИ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

Задание 1. Распределите биологические признаки в группы по характеру вариации: качественные или количественные. Для количественных признаков укажите тип варьирования (дискретный или непрерывный). Результаты внесите в таблицу.

Таблица. Классификация признаков.

Качественные признаки	Количественные признаки	
	счетные	мерные

Признаки: окраска листьев, число лепестков, урожай с единицы земельной площади, вес организмов, запах продуктов, длина колосьев, число зерен в колосьях, уловистость рыбы, число позвонков, рост человека, содержание кальция в сыворотке крови, качество породы, стойкость семян к заморозкам, успеваемость учащихся, форма семян, вкус вина, запах парфюма.

Задание 2. По данным исследования числа зерен в 50 колосьях двухрядного ячменя построить дискретный и интервальный вариационные ряды. Отразить их графически в виде кривых распределения: полигона распределения, гистограммы и кумуляты.

Результаты измерения числа зерен в колосьях ячменя: 21, 27, 17, 20, 22, 12, 24, 13, 20, 19, 22, 16, 22, 21, 16, 23, 16, 21, 24, 18, 11, 22, 15, 23, 21, 10, 15, 18, 15, 21, 14, 15, 9, 18, 22, 15, 17, 19, 17, 18, 17, 18, 24, 19, 16, 17, 15, 25, 16, 17.

Задание 3. Определить средние арифметические по дискретному и интервальному ряду (см. условие задачи 2). Сравнить результаты, сделать выводы.

Задание 4. По данным задания 2 рассчитать непараметрические средние (моду, медиану) в дискретном и интервальном рядах, используя способ накопления частот и формулы:

$$Mo = x_{мо} + i_{мо} \times \frac{f_{мо} - f_{мо-1}}{(f_{мо} - f_{мо-1}) \cdot (f_{мо} - f_{мо+1}}),$$

где $x_{мо}$ – начало модального интервала;

$i_{мо}$ – ширина модального интервала;
$f_{мо}$ – частота модального интервала;
$f_{мо-1}$, $f_{мо+1}$ – частоты пред- и постмодального интервалов соответственно.

$$Me = x_{ме} + i_{ме} \frac{0{,}5\sum f - \int f_{ме-1}}{f_{ме}},$$

где $X_{ме}$ – начало Ме-интервала;
$i_{ме}$ – величина Ме-интервала;
$f_{ме}$, $Sf_{ме-1}$ – частота медиального и сумма накопленных частот предмедиального интервалов соответственно.

Семинарское занятие № 2
ПОКАЗАТЕЛИ ВАРИАЦИИ. ПРИБЛИЖЕННЫЕ ОЦЕНКИ ЗАКОНА РАСПРЕДЕЛЕНИЯ. АСИММЕТРИЯ И ЭКСЦЕСС

Задание 1. Для двух рядов ранжированных значений переменных X и У определите среднее квадратическое отклонение. Сделайте вывод о варьировании значений признака вокруг центра распределения в этих рядах.

X: 10 15 20 25 30 35 40 45 50
У: 10 28 28 30 30 30 32 32 50

Задание 2. Имеются данные о поражаемости клеток при облучении ткани животного организма альфа-частицами. Результаты распределились следующим образом:

Число пораженных клеток, x	0	1	2	3	4	5	6	7
Частота поражений, p	112	168	130	68	32	5	1	1

Определите коэффициент вариации признака и сделайте вывод о степени варьирования числа пораженных клеток под действием облучения.

Задание 3. На пяти опытных делянках был получен следующий урожай (ц/га): 8,3 7,9 9,1 6,8 и 12,1. Оцените в предложенном ряду «выскакивающие» варианты способом нормированного отклонения.

Используя этот способ, сомнительные варианты нормируют относительно средней арифметической. Нулевой гипотезой служит предположение, что выскакивающие варианты принадлежат к той же генеральной совокупности, что и все другие варианты выборки.

Критерием оценки служит нормированное отклонение, которое для малой выборки имеет вид: $t = \frac{x - \bar{x}}{\sigma\sqrt{\frac{n+1}{n}}}$. *Полученное фактическое отклонение следует сравнить со стандартным для принятого уровня доверительной вероятности и числа наблюдений (таблица).*

Таблица

Стандартные значения критерия t для браковки «выскакивающих» вариант (по Л.З. Румшинскому, 1971 г.)

P / n	0,95	0,99	P / n	0,95	0,99
5	3,04	5,04	20	2,14	2,93
6	2,78	4,36	25	2,10	2,85
7	2,62	3,96	30	2,07	2,80
8	2,51	3,71	35	2,06	2,76
9	2,43	3,54	40	2,04	2,74
10	2,37	3,41	45	2,03	2,72
11	2,33	3,31	50	2,03	2,70
12	2,29	3,23	60	2,01	2,68
13	2,26	3,17	70	2,0	2,66
14	2,24	3,12	80	2,0	2,65
15	2,22	3,08	90	1,99	2,64
16	2,20	3,04	100	1,99	2,63
17	2,18	3,01	>100	1,96	2,57
18	2,17	2,98			

Задание 3. Известно, что не все признаки распределяются по нормальному закону: некоторые обнаруживают явную асимметрию или другие случаи отклонения от нормального закона. Поэтому прежде, чем использовать тот или иной критерий оценки генеральных параметров, следует составить представление о законе распределения изучаемого признака. Приближенную оценку закона распределения можно получить при помощи коэффициентов асимметрии и эксцесса.

Принято различать правостороннюю, или отрицательную, асимметрию и положительную, или левостороннюю. В случаях правосторонней, или отрицательной, асимметрии варианты накапливаются преимущественно в правой части ряда; вершина такого ряда сдвинута вправо. В случае левосторонней асимметрии правая ветвь кривой, начиная от вершины, больше левой. При симметричном распределении коэффициент асимметрии равен нулю. Мера косости меньше 0,5 считается малой, от 0,5 до 1 – средней, выше 1 – большой.

Наиболее совершенным показателем асимметрии служит центральный момент третьего порядка, отнесенный к кубу среднего квадратического отклонения: $A_s = \frac{\Sigma p(\bar{x} - x_i)^3}{n\sigma^3}$.

Наряду с симметричными и скошенными распределениями вариационные ряды могут быть высоко- и плосковершинными, многовершинными, или эксцессивными. Это свойство распределения называется *эксцесс*. Величина эксцесса, обозначаемая знаком E_x, измеряется центральным моментом четвертого порядка, отнесенным к среднему квадратическому отклонению в четвертой степени: $E_x = \frac{\Sigma p(\bar{x} - x_i)^4}{n\sigma^4}$.

Для строго симметричных распределений эксцесс равен нулю. При положительном эксцессе показатель E_x – число положительное, а кривая распределения островершинная. При отрицательном эксцессе кривая распределения может иметь две и более вершин, ее называют в таком случае плосковершинной. При Ex≤0,2 эксцесс практически отсутствует. Если же 0,5≤Ex≤1, эксцесс считается заметным, но небольшим. Крайняя степень положительного эксцесса теоретически безгранична. Предельное значение отрицательного эксцесса равно -2, что указывает на наличие двух вариационных рядов, т.е. рядов с самостоятельными центрами распределения, объединенных в одной общей совокупности.

Рассчитайте коэффициенты асимметрии и эксцесса для условия задания 2 и сделайте вывод о характере распределения признака.

Семинарское занятие № 3
ОЦЕНКА ДОСТОВЕРНОСТИ РАЗЛИЧИЙ МЕЖДУ ПРИЗНАКАМИ ВЫБОРОЧНЫХ СОВОКУПНОСТЕЙ. КРИТЕРИИ ДОСТОВЕРНОСТИ

Задание 1. В эксперименте изучалось влияние кобальта на увеличение живого веса кроликов. Опыт проводился на двух группах животных – опытной и контрольной. Обе группы животных содержались на одном и том же кормовом рационе. Опытная группа ежедневно получала в виде водного раствора по 0,06 г хлористого кобальта. За время опыта получены следующие прибавки в весе у животных:

контроль	504	560	600	420	530	490	580	580
опыт	580	692	700	621	640	561	680	630

Сделайте вывод относительно достоверности различий в средней прибавке веса опытных животных по сравнению с контрольными, используя критерий Стьюдента.

Задание 2. Имеются данные о вариабельности длины тела личинок щелкуна, обитающих в посевах озимой ржи и проса. Используя критерий Фишера, установите, есть ли достоверные различия в вариации этого параметра у животных в разных биотопах.

Задание 3. Два молодых человека (Ганс и Генри) одинакового возраста и внешности решили провести исследование сексуальных предпочтений женщин. Для этого они выбрали 6 кафе на открытом воздухе, популярные у женщин, и взяли напрокат два одинаковых велосипеда, один из которых оборудован сиденьем для ребенка. Молодые люди предполагают, что мужчина с велосипедом, у которого имеется сиденье для ребенка, будет более привлекателен для женщин. Эксперимент проводится после обеда в один из дней. Они совершают тур по шести уличным кафе, обозначенным от A до F. В каждом кафе они останавливаются на 10 минут. Стоя перед кафе со своими велосипедами и притворяясь, что говорят друг с другом, молодые люди стараются установить как можно больше зрительных контактов с представителями женского пола, которые сидят в кафе. После каждого кафе они меняются велосипедами. Результаты этого эксперимента представлены в таблице.

	Количество зрительных контактов в кафе от A до F						
	A	B	C	D	E	F	Сумма
Ганс	**12**	10	**14**	7	**17**	12	72
Генри	9	**17**	10	**10**	12	**20**	78
Сумма	21	27	24	17	29	32	150

Выделены и подчеркнуты результаты, полученные мужчиной, который держал велосипед с сиденьем для ребенка.

1. Какое из следующих утверждений соответствует нулевой гипотезе для эксперимента Ганса и Генри?

a. Ганс и Генри имеют одинаковую привлекательность для женщин.

b. Привлекательность мужчины, велосипед которого имеет сиденье для ребенка, такая же, как и у мужчины с велосипедом без сиденья.

c. Все шесть кафе не различаются по тому, какие женщины их посещают.

d. То, что мужчина установит зрительный контакт с женщиной, не является показателем его привлекательности.

e. Привлекательность мужчины, велосипед которого имеет сиденье для ребенка, большая, чем у мужчины с велосипедом без сиденья.

2. Проверьте значимость различий между ситуацией А (мужчина + велосипед с сиденьем) и Б (мужчина + велосипед без сиденья) с помощью t-теста. Для расчетов используйте таблицу.

Уровень значимости, %	Критическое значение t
10	2,02
5	2,57
2,0	3,37
1	4,03
0,1	6,86

Насколько можно быть уверенным, что нулевая гипотеза неверна (разница между ситуациями А и Б значима)?

a. Менее чем 75%
b. Между 75 и 90%
c. Между 90 и 95%
d. Между 95 и 97,5%
e. Между 97,5 и 99%
f. Между 99 и 99,5%
g. Более 99,5%.

3. Ганс и Генри показывают свои результаты руководителю. Он утверждает, что Ганс и Генри сделали большую ошибку, анализируя общее число контактов в расчете на 1 кафе, поскольку эти 6 кафе сильно различаются. Ганс и Генри не согласны с руководителем и хотят доказать свою точку зрения с помощью χ^2-теста. Определите значение χ^2 и по его величине уровень значимости (Р), используя данные таблицы.

(df)	Вероятность (Р) случайности отклонения									
	0,995	0,975	0,9	0,5	0,3	0,25	0,1	0,05	0,025	0,01
1	0,00	0,00	0,02	0,46	1,07	1,32	2,71	3,84	5,02	6,64
2	0,01	0,05	0,21	1,39	2,41	2,77	4,61	5,99	7,38	9,21
3	0,07	0,22	0,58	2,37	3,67	4,11	6,25	7,82	9,35	11,35
4	0,21	0,48	1,06	3,36	4,88	5,39	7,78	9,49	11,14	13,28

Какой из следующих выводов является верным, исходя из результатов χ^2-теста?

a. Кафе различаются, но различия незначительны.
b. Различия между кафе значительны.
c. Результаты сомнительны или могут быть поставлены под вопрос, что-то неправильно сделано в постановке эксперимента.
d. Кафе не различаются, но это не является значимым.
e. Кафе не различаются и это является значимым.

Семинарское занятие № 4

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Задание 1. С помощью корреляционного анализа установите, существует ли прямолинейная зависимость между весом самцов павианов-гамадрилов и количеством гемоглобина в их крови. Оцените достоверность полученных показателей.

Вес самцов, кг (X)	**Кол-во Hb, % (У)**
18	70
17,7	74
19	72
18	80
19	77
22	80
21	89
20	76
30	86

Задание 2. Определите степень криволинейной зависимости между признаками, рассчитав корреляционное отношение $\eta_{x/y}$ для следующих данных:

X: 2 4 6 8 4 6 2 6
У: 4 8 8 7 4 10 6 12
Определите достоверность полученного коэффициента.

Задание 3. Рассчитайте коэффициент ассоциации и сделайте вывод о наличии корреляции между качественными признаками (форма крыльев и окраска тела) в лабораторной культуре мух-дрозофил, если у потомства получены следующие результаты скрещивания:

серые особи с нормальными крыльями – 75;
серые с зачатками крыльев – 16;
черные с нормальными крыльями – 14;
черные с зачаточными крыльями – 68.

Оцените достоверность выборочного коэффициента ассоциации.

Лабораторная работа № 1

БИОЛОГИЧЕСКИЕ ПРИЗНАКИ

Цель работы: освоить навыки выбора и измерения мерных и счетных признаков на различных биологических объектах.

Контрольные вопросы:

5. Биологические признаки, их классификация.

6. Ошибки наблюдений.

7. Понятие выборочной совокупности и генеральной совокупности.

8. Статистическая совокупность и ее свойства.

9. Требования к выборке.

10. Статистические таблицы.

Изучение биологических явлений проводится не по отдельным наблюдениям, которые могут оказаться случайными, нетипичными, неполно выражающими сущность данного явления, а на множестве однородных наблюдений, что дает более полную информацию об изучаемом объекте. Некоторое множество относительно однородных предметов или объектов, объединяемых по тому или иному признаку для совместного изучения, называют статистической совокупностью. Элементы, входящие в состав совокупности, называются ее членами, или вариантами (от лат. varians – изменяющийся). Варианты – это отдельные наблюдения или числовые значения признака. Так, если обозначить признак через X (икс большое), то его значения или варианты будут обозначаться через x (икс малое), т.е. как x_1, x_2, x_3, x_n и т.д. Общее число вариант, входящих в состав данной совокупности, называется ее объемом и обозначается буквой n.

Вместо сплошного обследования генеральной совокупности изучению подвергается обычно какая-то ее часть, получившая название выборочной совокупности, или выборки.

Одной из наиболее распространенных форм группировок выборочных данных служат статистические таблицы. Статистические таблицы имеют иллюстративное значение, показывая какие-то общие итоги, положение отдельных элементов в общей серии наблюдений и т.д. Все это способствует пониманию описываемых явлений, выяснению того существенного, чем они отличаются друг от друга.

Задание

1. На предложенных объектах провести выделение мерных и счетных биологических признаков и произвести их измерение и подсчет.
2. Произвести систематизацию и группировку полученных данных.
3. Оформить статистические таблицы по образцу (табл. 1).

Таблица 1

Результаты измерений признаков биологических объектов

№ п.п.	Признак			
	Высота растения	Масса растения	Количество листьев	Количество соцветий
1.				
2.				
3.				
….				
n				
Среднее значение				

Лабораторная работа № 2
ВАРИАЦИОННЫЙ РЯД

Цель работы: получить навыки построения и графического изображения вариационного ряда статистической совокупности.

Контрольные вопросы:

1. Вариационный ряд значений;
2. Ранжирование (способы и значение);
3. Графическое выражение распределений:
   a. гистограмма;
   b. полигон;
   c. кумулята;
   d. огива.

Помимо статистических таблиц формой первичной группировки выборочных данных служит способ ранжирования, т.е. расположение вариант в определенном порядке – по возрастающим или убывающим значениям признака. При большом числе наблюдений ранжировать выборочную совокупность принято в виде двойного ряда, т.е. с указанием частоты или повторяемости отдельных вариант ранжированного ряда. Такой двойной ряд ранжированных значений признака называется вариационным рядом, или рядом распределения.

Для составления вариационного ряда необходимо установить число интервалов (классов), которое зависит от числа выборки (таблица 1).

Таблица 1

Число интервалов вариационного ряда по П.Ф. Рокицкому

Число наблюдений	25–40	41–60	61–100	101–200	≥201
Число интервалов	5–6	6–8	7–10	8–12	9–15

Число интервалов можно определить через логарифм численности выборки по формуле

$$k = 1 + 3{,}322 \lg n,$$

где k – число интервалов вариационного ряда;

n – численность выборки.

Величина классового интервала обозначается через *i*; она определяется по разности между максимальной и минимальной вариантами, отнесенной к избранному числу классов, т.е. по следующей приближенной формуле: $i = \frac{x_{\max} - x_{\min}}{k}$,

где *i* – классовый интервал, который берется целым числом;

x_{max} – максимальная и x_{min} — минимальная варианты выборки;

k – число классов, на которые разбивается выборочная совокупность.

Вычисленные величины интервалов округляют до целых чисел. При установлении пределов интервалов к минимальному значению изучаемого признака последовательно прибавляют принятую величину интервала до тех пор, пока не достигнута максимальная величина признака. Среднее значение интервала определяется как полусумма предельных значений данного интервала.

После того, как установлен классовый интервал и выборочная совокупность разбита на классы, производится разноска вариант по классам, определяются частоты каждого класса.

Чтобы выразить эту закономерность более наглядно, принято изображать вариационные ряды графически в виде гистограммы, полигона, кумуляты или огивы. Гистограмма получается, если по оси абсцисс отложить границы классов, а по оси ординат – частоты классов вариационного ряда. Таким образом, гистограмма изображает распределение вариант при непрерывном варьировании признака. Прямоугольники соответствуют классам, а их высота — количеству вариант, заключенных в каждом классе.

Если из срединных точек вершин прямоугольников гистограммы опустить перпендикуляры на ось абсцисс, а затем эти точки соединить между собой, получится график прерывистого варьирования, называемый полигоном распределения. Если на оси абсцисс нанести классовые варианты, т.е. срединные значения классов, а на оси ординат – накопленные частоты классов, соединив затем соответствующие точки в системе координат, то получится график, называемый *кумулятой.*

Накопленные частоты классов получаются последовательным суммированием (кумуляцией) всех частот вариационного ряда в направлении от минимальной варианты до конца ряда. Если ряд накопленных частот нанести на ось абсцисс, а срединные значения классов – на ось ординат и построить соответствующий график, то он будет на-

зываться *огивой*. Нетрудно понять, что огива есть не что иное, как кумулята, повернутая на 180^0.

Ниже приводятся примеры построения гистограммы, полигона распределения, кумуляты и огивы на примере измерений длины листа выборки одуванчика лекарственного (рисунок 1). Результаты измерений приводятся в таблице 2.

Таблица 2

Результаты измерений длины листа одуванчика лекарственного

Значения вариант	8,8	9,3	9,8	10,3	10,8	11,3	11,8	12,3	12,8	13,2	13,8
Частота	1	4	6	9	12	14	16	11	7	5	3
Накопленные частоты	1	5	11	20	32	46	62	73	80	85	88

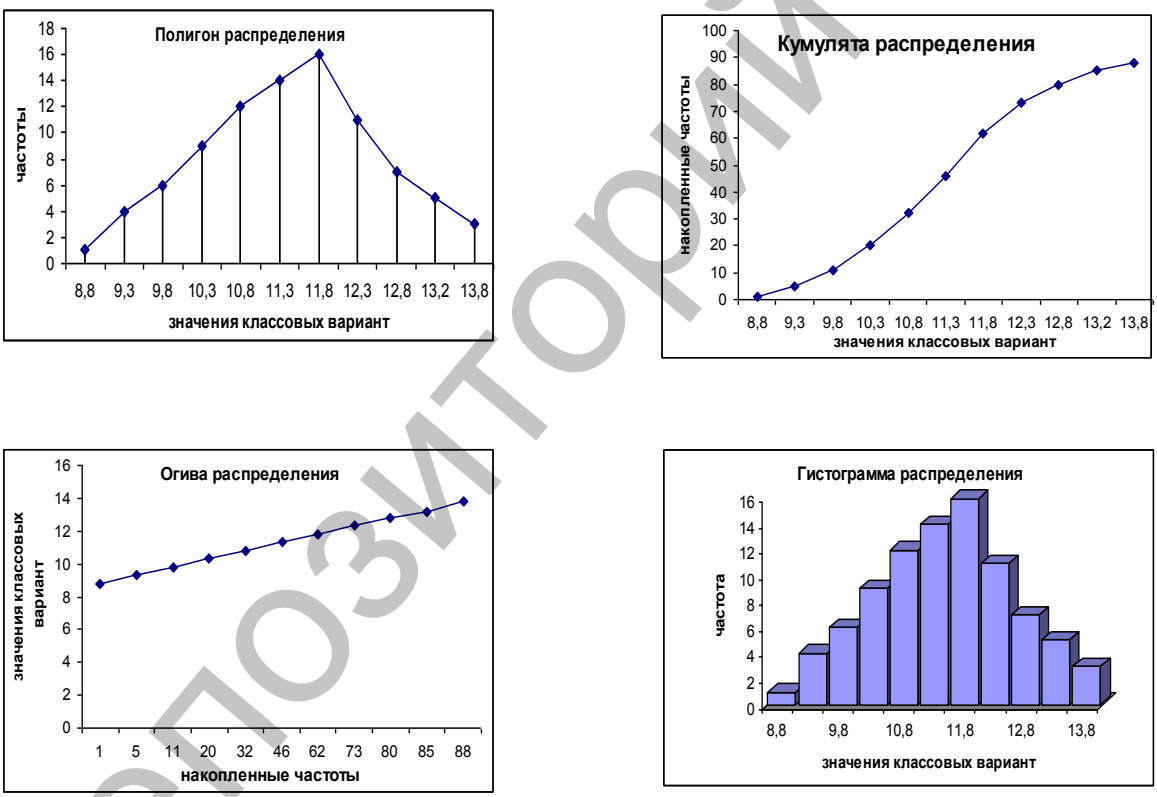


Рис. 1. **Графическое изображение закономерностей распределения признаков.**

Задание

1. Ранжировать данные измерений биологических признаков, полученные на предыдущем занятии.
2. Построить вариационный ряд значений.
3. Изобразить вариационный ряд значений в виде диаграммы, полигона распределения, кумуляты, огивы.

Лабораторная работа № 3

СРЕДНИЕ ВЕЛИЧИНЫ И СПОСОБЫ ИХ ВЫЧИСЛЕНИЯ

Цель работы: получить навыки выбора вычисления средних величин на больших и малых выборках.

Контрольные вопросы

1. Средняя арифметическая и способы ее вычисления.
2. Основные свойства средней арифметической.
3. Непараметрические средние:
   - медиана;
   - мода.
4. Средняя арифметическая в оценке качественных признаков.

Наиболее часто и широко как в практической деятельности человека, так и в научных исследованиях используется средняя величина. Она дает суммарную характеристику любого признака, указывая на то типичное и устойчивое в явлении, что наиболее полно выражает его содержание. Средняя арифметическая, которую принято обозначать через $\bar{x}$, или через М, есть не что иное, как частное от деления суммы всех вариант совокупности на их число, т.е. $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots . x_k}{n} = \frac{\Sigma x}{n} = \frac{1}{n}\Sigma x$ (1).

Это и есть общая формула средней арифметической, где x_1 x_2, x_3, ..., x_k обозначают варианты, входящие в состав данной совокупности; Σ – знак суммирования; n – общее число вариант, или объем выборочной совокупности. Средняя арифметическая – число именованное, она выражается теми же единицами меры или счета, что и характеризуемый ею признак.

При повторяемости отдельных вариант среднюю арифметическую можно представить как сумму произведений отдельных вариант на их частоты, отнесенную к общему числу всех вариант данной совокупности, т.е. как $\bar{x} = \frac{\Sigma xp}{n} = \frac{1}{n}\Sigma xp$. (2)

Существует упрощенный способ, позволяющий быстро и точно определять среднюю величину. Сущность этого способа очень проста: одну из вариант, все равно какую, условно принимают за среднюю арифметическую. Обычно в качестве условной средней берется варианта с большей частотой, хотя это совершенно не обязательно. Условную среднюю обозначим *A*. После выбора условной средней остается найти величину той поправки, которую нужно прибавить или отнять от условной средней, чтобы получить истинное значение средней арифметической данной совокупности. Эта поправка, называемая центральным моментом первого порядка, равна сумме произ-

ведений частот вариационного ряда на отклонения вариант от условной средней (А), отнесенной к числу всех вариант данного ряда.

Формула средней арифметической, вычисляемой по этому способу, принимает следующий вид:

$$\bar{x} = A + \frac{\Sigma pa}{n} \quad (3),$$

где $\bar{x}$ – средняя арифметическая; А – условная средняя; *Σpa* – сумма произведений частот *(p)* на отклонения вариант от условной средней, *a* = *x*-А; n – объем выборки.

Преимущество этого способа более заметно на больших выборках, особенно в тех случаях, когда при вычислении средней арифметической по формуле (2) приходится перемножать многозначные числа.

Медиана – показатель описательного характера – не зависит от параметрических характеристик ряда. Она служит серединой вариационного ряда, который делит на две равные части: в обе стороны от медианы располагается одинаковое число вариант. Срединное значение вычисляется по формуле

$$Me = x_{\min} + i\frac{0{,}5N - \sum n}{n_e},$$

где x_{min} – минимальное значение предела интервала, где находится срединное значение; i – величина интервала; N – численность выборочной совокупности; Σn – суммарная численность до интервала, в котором находится срединное значение; n_e – численность интервала, где находится срединное значение.

Медиану можно определить и графически по кумуляте. Для этого из точки, лежащей на оси ординат и соответствующей 5%-й всей численности выборки, проводится прямая, параллельная оси абсцисс, до пересечения с кумулятой. Перпендикуляр, опущенный из точки пересечения на ось абсцисс, укажет срединное значение признака.

Модой называется наиболее часто встречающаяся величина. В непрерывных вариационных рядах мода находится обычно в том классе, который имеет наибольшее число вариант. Этот класс называется модальным классом. Мода, как и медиана, служит вспомогательной характеристикой вариационного ряда. Правильное вычисление моды возможно в том случае, когда известен закон распределения значений статистической величины. Приближенно моду вычисляют, используя формулу Пирсона:

$$Mo = 3Me - 2M,$$

где Me – медиана ряда распределения; M – среднее значение признака.

В рядах симметричных (нормальных) среднее значение, медиана и мода совпадают. В действительных рядах такое совпадение маловероятно, так как фактическое распределение численностей биологических явлений природы отличается от нормального, имеет ту или иную косость.

Задание

Для данных, полученных при измерении натуральных биологических объектов (занятие 1), определить:

- среднюю арифметическую с использованием всех вариантов вычисления;
- медиану;
- моду.

Лабораторная работа № 4
ПОКАЗАТЕЛИ ВАРИАЦИИ И СПОСОБЫ ИХ ВЫЧИСЛЕНИЯ

Цель работы: получить навыки оценки степени вариации признаков на больших и малых выборках.

Контрольные вопросы

1. Среднее линейное отклонение.
2. Дисперсия или варианса.
3. Среднее квадратическое отклонение.
4. Коэффициент вариации.
5. Ошибка средней арифметической.

Средняя арифметическая служит одной из важнейших характеристик вариационного ряда. Но она ничего не говорит о величине вариации характеризуемого признака. Величина вариации может быть оценена и по разности между максимальной и минимальной вариантами совокупности. Этот показатель получил название *размаха* вариации.

Средним линейным отклонением называют сумму отклонений вариант от средней арифметической, отнесенную к общему числу вариант данной совокупности: $\Delta = \frac{\sum(\bar{x} - x_i)}{n}$.

Среднее квадратичное отклонение, или основное отклонение – важнейшая характеристика вариационного ряда, являющаяся мерой рассеяния ряда распределения, показывает отклонение (для 68 случаев из 100) статистических величин от среднего значения. Определяется эта величина по формуле

$$S_x(\sigma_x) = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}},$$

где Σ – знак суммирования произведений отклонений вариант x_i от их средней x на веса или частоты p этих отклонений в пределах от первого до i-го класса; n – общее число наблюдений, или объем выборки; (n -1) – число степеней свободы.

Дисперсия, или *варианса*, генеральной совокупности обозначается символом σ^2, а дисперсия выборки – S^2. В случае нормального распределения их величины совпадают.

Ошибка средней (m_x) является величиной, на которую отличается среднее значение выборочной (опытной) совокупности от среднего значения генеральной совокупности при условии, что распределение изучаемого признака приближается к нормальному. Основная ошибка среднего рассчитывается по формуле $m_x = \frac{\sigma}{\sqrt{n-1}}$. Среднее значение необходимо записать с основной ошибкой ($X \pm m_x$), только в этом случае можно судить о точности опыта.

Значение σ не всегда достаточно полно характеризует вариабельность рассматриваемой величины, особенно если приходится сравнивать изменчивость разных в количественном и качественном отношении признаков. Более информативным и удобным при сравнении различных статистических совокупностей является *коэффициент изменчивости*, или *вариации*, так как его величина не зависит от единиц, используемых при измерениях.

Коэффициент изменчивости – это основное отклонение, выраженное в процентах от среднего значения, которое рассчитывается по формуле $C = \frac{\sigma}{X} \times 100\%$.

На основании величины коэффициента изменчивости можно судить о характере и степени варьирования признака (таблица 1).

Таблица 1

Характер изменчивости признаков (по М.Л. Дворецкому, 1971)

Коэффициент изменчивости, С	до 5%	6–10%	11–20%	21–50%	более 50%
Характер изменчивости	слабая	умеренная	значительная	большая	очень большая

После вычисления того или иного статистического показателя необходимо проверить степень его надежности (достоверности) путем деления величины этого показателя на его ошибку. Достоверность среднего значения определяется по формуле $t = \frac{X}{m_x}$. Если значение t больше четырех, то среднее значение показателя является достоверным. Таким показателем можно пользоваться для сопоставления и формулировки корректных выводов. Часто о достоверности показателя судят, если $t \geq 3$. Если же t меньше трех, то по таким показателям нельзя делать категорические заключения или проводить сопоставления.

Важным показателем, характеризующим процент расхождения между выборочной и генеральной средними, является *точность опыта (p,%),* или *ошибка наблюдений.* Эта величина характеризует субъективную ошибку исследователя. Ошибка выборки выражается в процентах от соответствующей средней: $p\% = \frac{m_x}{\bar{x}} 100$. Точность опыта показывает, на сколько процентов можно ошибиться, если утверждать, что генеральная средняя равна полученной выборочной средней. В 68 случаях из 100 расхождение между выборочной и генеральной средними не будет превышать однократного значения точности опыта (в ту или иную сторону).

В некоторых экспериментах требуется очень высокая точность опыта (например, в медико-биологических, токсикометрических и др.), когда ошибка не должна превышать 1%. Рассчитанный по формуле процент ошибки необходимо сопоставить с заданным. Если он не выше заданного, то точность достаточная, а если выше, то точность результата является неудовлетворительной, необходимо увеличить число наблюдений.

Для определения объема выборки с заданной точностью опыта используют формулы: $n = C^2 / p^2$ (если точность указана в процентах) и $n = \sigma^2 / m_x^2$ (если точность дается в абсолютных величинах).

Задание

Для данных, полученных при измерении признаков на натуральных объектах, вычислить:

- линейное отклонение;
- дисперсию;
- среднеквадратическое отклонение;
- ошибку средней;
- коэффициент вариации.

Проверить достоверность полученной средней, оценить точность опыта. Рассчитать оптимальный объем выборки для вычисления среднего значения с точностью 1%, 0,2 см (г, экз. и т.п.).

Лабораторная работа № 5

АСИММЕТРИЯ И ЭКСЦЕСС

Цель работы: получить навыки вычисления асимметрии и эксцесса.

Контрольные вопросы

1. Нормальное распределение.
2. Основные свойства нормального распределения.
3. Асимметрия.
4. Причины асимметрии.
5. Измерение асимметрии.
6. Эксцесс и его измерение.

Наряду с практически симметричными распределениями встречаются и скошенные, асимметричные ряды. Аналитически они характеризуются нарушением равенства между модой, медианой и средней арифметической распределения. Графически они выражаются асимметричными кривыми распределения. Принято различать правостороннюю, или отрицательную, асимметрию и положительную, или левостороннюю. В случаях правосторонней, или отрицательной, асимметрии варианты накапливаются преимущественно в правой части ряда; вершина такого ряда сдвинута вправо. В случае левосторонней асимметрии правая ветвь кривой, начиная от вершины, больше левой.

Пирсон предложил оценивать степень асимметрии по разности между средней арифметической и модой, отнесенной к величине среднего квадратического отклонения:

$$A_s = \frac{\bar{x} - M_o}{\sigma},$$

где A_s – мера скошенности рядов распределения, или коэффициент асимметрии.

В качестве показателя асимметрии также может служить утроенная разность между средней арифметической и медианой, отнесенная к величине среднего квадратического отклонения, т.е. $A_s = \frac{3(\bar{x} - M_e)}{\sigma}$.

Величина этого показателя обычно не выходит за пределы –3 и +3, что указывает на отрицательную или положительную асимметрию. При симметричном распределении коэффициент асимметрии равен нулю. Мера косости меньше 0,5 считается малой, от 0,5 до 1 – средней, выше 1 – большой.

Наиболее совершенным показателем асимметрии служит *центральный* момент третьего порядка, отнесенный к кубу среднего квадратического отклонения: $A_s = \sum pa^3 \Big/ n\sigma^3$, или $A_s = \frac{\Sigma p(\bar{x} - x_i)^3}{n\sigma^3}$.

Наряду с симметричными и скошенными распределениями вариационные ряды могут быть высоко- и плосковершинными, многовершинными, или эксцессивными. Это свойство распределения называется *эксцесс*. Величина эксцесса, обозначаемая знаком E_x, измеряется центральным моментом четвертого порядка, отнесенным к среднему квадратическому отклонению в четвертой степени $E_x = \frac{\sum pa^4}{n\sigma^4}$

Для строго симметричных распределений эксцесс равен нулю. При положительном эксцессе показатель E_x – число положительное, а при отрицательном эксцессе – отрицательное. В обоих случаях коэффициент эксцесса – величина отвлеченная, не именованная.

При $E_x \leq 0{,}2$ эксцесс практически отсутствует. Если же $0{,}5 \leq E_x \leq 1$, эксцесс считается заметным, но небольшим. Крайняя степень положительного эксцесса теоретически безгранична. Предельное значение отрицательного эксцесса равно -2, что указывает на наличие двух вариационных рядов, т.е. рядов с самостоятельными центрами распределения, объединенных в одной общей совокупности.

Задание

- Для вариационных рядов значений исследуемых биологических признаков определить величину асимметрии и оценить степень скошенности распределения.
- Определить величину эксцесса, объяснить его причину.

Лабораторная работа № 6
НОРМИРОВАННОЕ ОТКЛОНЕНИЕ И ПОНЯТИЕ НОРМЫ

Цель работы: изучить особенности нормального распределения.

Контрольные вопросы

1. Нормальное распределение.
2. Основные свойства нормального распределения.
3. Статистические границы нормы.
4. Нормированное отклонение.

Наблюдения показывают, что большинство учитываемых признаков у человека, животных и растительных организмов распределяется по нормальному закону. Следовательно, общие статистические

границы нормы дает критерий x±3σ, поскольку все варианты практически симметричного распределения укладываются в эти пределы.

В теоретической и прикладной статистике большое значение имеет нормирование, позволяющее использовать среднее квадратическое отклонение для оценки отдельных вариант по отношению их к средней величине данной совокупности. Такого рода оценка производится по разности между вариантой и средней арифметической, отнесенной к величине среднего квадратического отклонения, т.е. $t = \frac{\bar{x} - x}{\sigma}$. Здесь t называется нормированным отклонением. Этот показатель удобен и прост как при оценке единичных вариант, так и при относительной характеристике сравниваемых друг с другом индивидов.

Разумеется, статистические границы нормы не могут быть очень жесткими: в зависимости от задачи исследования и различных обстоятельств их можно сузить до x ± 0,5σ или, наоборот, расширить до x±1σ. Варианты, которые распределяются между пределами от 0,67σ до 2σ, должны считаться субнормальными. Все же остальные члены, совокупности, выходящие за пределы x ± 2σ, следует отнести к категории аномалий.

Задание:

Для исследуемых качественных и количественных признаков установить степень отклонения от нормы отдельных вариант.

Лабораторная работа № 7
ОШИБКИ РЕПРЕЗЕНТАТИВНОСТИ

Цель работы: получить навыки определения достоверности эмпирических показателей.

Контрольные вопросы:

1. Ошибка отдельно взятой варианты.
2. Ошибка средней арифметической.
3. Ошибка среднего квадратического отклонения.
4. Ошибка коэффициента вариации.
5. Ошибки показателей асимметрии и эксцесса.
6. Оценка достоверности различий между дисперсиями.

Расхождение между величиной средней арифметической (x) выборки и величиной средней арифметической генеральной совокупности (M) принято называть *ошибкой репрезентативности*, т.е. ошибкой, допускаемой не в самом процессе измерительной и вы-

числительной работы, а в результате случайного отбора вариант из генеральной совокупности при образовании выборки.

Если судить о величине статистической ошибки отдельно взятой варианты, то она равна среднему квадратическому отклонению, так как любое эмпирическое распределение, следующее нормальному закону, практически укладывается в пределах плюс-минус трех сигм, т.е. x±3σ. Ошибку репрезентативности поэтому называют средней квадратической ошибкой, или просто средней ошибкой. Будем ее обозначать через m, указывая при этом и характеристику, которую она сопровождает. Таким образом, средняя квадратическая ошибка отдельно взятой варианты выразится в виде $m_x=\pm\sigma$.

Выборочная средняя (x) отклоняется от своего математического ожидания или средней арифметической (M) генеральной (теоретически рассчитанной) совокупности меньше в $\sqrt{n}$ раз по сравнению с отдельными вариантами данного распределения. Отсюда $m_x = \frac{\sigma}{\sqrt{n}}$.

Поскольку весь вариационный ряд нормально распределяющейся случайной величины *X* практически укладывается в пределах между *x*-3σ и *x*-σ, то можно сказать, что генеральная средняя *(M)* таких распределений не выходит за пределы утроенного значения средней ошибки средней арифметической любой выборки, взятой из данной генеральной совокупности, т.е. она всегда заключена между пределами от *x*-3m_x до *x*+3m_x или в пределах x±3m_x. Поэтому утроенное значение средней квадратической ошибки называется предельной ошибкой средней арифметической выборочной совокупности. А выражение *x*±3m_x заключает в себе содержание так называемого «правила утроенной ошибки».

При вычислении ошибки средней арифметической на малых выборках число наблюдений *(n)* берется «числом степеней свободы», и формула принимает следующий вид: $m_x = \frac{\sigma}{\sqrt{n-1}}$. Средняя ошибка среднего квадратического отклонения вычисляется по формуле $m_\sigma = \frac{\sigma}{\sqrt{2n}}$. Средняя ошибка коэффициента вариации (C) определяется по следующей приближенной формуле: $m_C = \frac{C}{\sqrt{2n}} \times \sqrt{1+2\left(\frac{C}{100}\right)^2}$.

Средняя ошибка показателя асимметрии определяется по следующей приближенной формуле: $m_{A_s} = \sqrt{\frac{6}{n}}$, или более точно по формуле $m_{A_s} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$.

Ошибку коэффициента эксцесса можно вычислить по следующим аналогичным формулам: $m_{E_x} = 2\sqrt{\frac{6}{n}}$, или $m_{Ex} = \sqrt{\frac{24}{n}}$.

Как и в случаях сравнения выборочных средних, разность между двумя средними квадратическими отклонениями – σ_1 и σ_2 – оценивается путем нормирования, т.е. по критерию достоверности: $t = \frac{\sigma_1 - \sigma_2}{m_\sigma} = \frac{D_\sigma}{m_\sigma}$, где m_a – средняя квадратическая ошибка указанной разности. Она вычисляется по следующей формуле: $m_\sigma = \sqrt{\frac{\sigma_2^1}{2n_1} + \frac{\sigma_2^2}{2n_2}}$.

Задание. Рассчитать для исследуемых признаков:

- ошибку отдельно взятой варианты;
- ошибку средней арифметической;
- ошибку среднего квадратического отклонения;
- ошибку коэффициента вариации;
- ошибки показателей асимметрии и эксцесса;
- достоверности различий между дисперсиями.

Лабораторная работа № 8
КРИТЕРИЙ ДОСТОВЕРНОСТИ

Цель работы: получить навыки оценки достоверности полученных результатов.

Контрольные вопросы

1. Достоверность различий между выборочными средними.
2. Достоверность различий между двумя дисперсиями.
3. Критерий соответствия между ожидаемыми и наблюдаемыми частотами.

Сравнительный анализ биометрических показателей сводится обычно к оценке степени достоверности наблюдаемых между ними различий. Оценка существенности или достоверности различий, наблюдаемых между двумя выборочными средними x_1 и x_2, производится на основе нормирования, т.е. отнесения разности между средними, которую обозначим через D, к средней квадратической ошибке этой разности:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{m_1^2 + m_2^2}} = \frac{D}{m_d}.$$

Здесь t называется критерием достоверности. Значение этого критерия оценивается по таблицам вероятности Стьюдента (приложение 1) на основании числа степеней свободы для заданного уровня ве-

роятности: P=0,95; P=0,99; P=0,999. Число степеней свободы равно $u = n_1 + n_2 - 2$. Если фактическое t больше стандартного (табличного) t_{st} для данного уровня вероятности, различие существенное, достоверное и его нельзя объяснить случайными причинами.

Разница между показателями вариации двух независимых распределений оценивается с помощью критерия Фишера, обозначаемого через F и представляющего отношение дисперсий (варианс): $F = \sigma_1^2 / \sigma_2^2$. Числителем всегда берется большая варианса. Поэтому критерий F может быть равен 1 или больше ее. Если F=1, это указывает на равенство дисперсий. Когда же такого равенства нет, возникает необходимость оценить, случайно расхождение между дисперсиями или нет. Чем больше величина F, тем значительнее расхождение между дисперсиями и, наоборот, чем ближе значение F к 1, тем меньше расхождение между сравниваемыми показателями вариации. Р.А. Фишер получил значения F-критерия для различных уровней значимости и различного числа степеней свободы (приложение 2). Число степеней свободы равно численности выборки без единицы (n-1). Если фактическое значение критерия Фишера F будет больше стандартного (табличного), то различие дисперсий двух выборок доказано.

При сравнении наблюдаемых и ожидаемых результатов применяются особые критерии оценки, в частности, критерий хи-квадрат (χ^2). Критерий предложен Карлом Пирсоном и представляет собой сумму отношений между квадратами разностей эмпирических и вычисленных или ожидаемых частот к ожидаемым частотам: $\chi^2 = \sum \frac{(p - p')^2}{p'}$, где Σ – знак суммирования, p – эмпирическая частота, p’ – ожидаемая или теоретически вычисленная частота.

Использование χ^2-теста необходимо для того, чтобы узнать, подтверждается ли гипотеза экспериментом, т.е. насколько верны условия эксперимента, позволяют ли они с высокой степенью достоверности подтвердить или опровергнуть исходное предположение. Если бы фактические данные полностью совпадали с теоретическими, значение критерия было бы равно нулю. По мере увеличения разницы между этими показателями значение критерия будет возрастать. Каждому значению χ^2 соответствует определенная вероятность его появления (табличные данные, приложение 3). Значение χ^2 в таблице указывают те границы, до которых полученные значения критерия не дают оснований сомневаться в высказанном предположении с определенной степенью вероятности. Значений χ^2, превышающие табличные, будут указывать на несостоятельность гипотезы, т.е. признание того, что различие между фактическими и теоретически ожидаемыми результатами является достоверным, значимым.

Задание. Для предложенных примеров произвести расчет критериев:

- достоверности;
- Фишера;
- хи-квадрат;

и оценить их величину.

Лабораторная работа № 9
ДИСПЕРСИОННЫЙ АНАЛИЗ

Цель работы: получить навыки изучения статистического влияния одного или нескольких факторов на результативный признак.

В практике нередко возникает необходимость в оценке целых комплексов количественных показателей – необходимость сравнивать между собой одновременно не две, а несколько выборок, объединенных в единый комплекс. Статистический анализ целого комплекса требует особого метода, который был разработан Р.А. Фишером (1925) и получил название дисперсионного анализа.

Ход анализа

Методика дисперсионного анализа сводится к некоторой общей схеме, которую можно свести к следующим 6 пунктам:

1. Собранные данные упорядочиваются, сводятся в таблицу в соответствии с условиями опыта. Затем определяется общая вариация для всего комплекса, равная сумме квадратов отклонений отдельных вариант от общей средней для всего комплекса $(\bar{x}_s)$. Будем обозначать общую вариацию через a_0, ее удобнее определять по следующим аналогичным формулам:

$$a_0 = \sum x^2 - n\bar{x}_s^2 \text{, или } a_0 = \sum x^2 - \frac{1}{x}(\sum x)_s^2 ,$$

где Σx^2 – сумма квадратов вариант, входящих в состав всего комплекса;

$\bar{x}$ – средняя арифметическая комплекса, т.е. общая средняя выборочной совокупности;

n – общее число наблюдений.

2. Определяется межгрупповая вариация (a_1):

$$a_1 = N\sum \bar{x}^2 - \frac{1}{n}(\sum x)^2 \text{ или } a_1 = N\sum \bar{x}^2 - n\bar{x}^2 ,$$

где N – число вариант в группах;

$\bar{x}$ – средняя арифметическая групп;

x – значение отдельной варианты (остальные обозначения см. выше).

3. Определяется внутригрупповая, или остаточная, вариация по разности $a_2 = a_0 - a_1$.

4. Находится значение дисперсий – межгрупповой и внутригрупповой, для чего соответствующие вариации относятся к числу степеней свободы в группах *(n*-1) и к числу степеней свободы внутригрупповой вариации *(n - k)*, т.е. $\sigma_1^2 = \frac{a_1}{N-1}$ и $\sigma_2^2 = \frac{a_2}{n-k}$.

5. Берется отношение дисперсий $F = \frac{\sigma_1^2}{\sigma_2^2}$.

6. Критерий *F* оценивается по таблицам Фишера для соответствующих степеней свободы и взятого уровня значимости. При этом, когда F≤1, следует брать отношение большей вариансы к меньшей, т.е. оценивать $F = \frac{\sigma_2^2}{\sigma_1^2}$ и предельное значение F для большего числа степеней свободы находить по столбцам, а для меньшего – по строкам таблицы. Если найденная величина *F* при заданном уровне значимости и данных числах степеней свободы превышает значение этого показателя, указанное в таблице, то различия, наблюдаемые между групповыми средними, достоверны. В противном случае, т.е. когда критерий *F* меньше своего предельного значения, эти различия носят случайный характер и не могут быть признаны существенными, достоверными.

Задание: провести дисперсионный анализ для предложенных вариантов.

Пример задания для расчета.

Варианты опыта	Урожай в кг по повторностям			Средний урожай, кг
	1	2	3	
Контроль	21,2	28,0	31,2	26,80
1. Удобрения помещались ниже семян на 3–5 см	23,6	22,5	28,0	24,73
2. Удобрения помещались в стороне от семян на 3–5 см	24,0	30,0	29,2	27,73
3. Удобрения помещались выше заделки семян на 3–5 см	29,2	28,0	27,0	28,07

Лабораторная работа № 10

КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Цель работы: получить навыки оценки степени связи между признаками.

Контрольные вопросы

1. Суммарный показатель связи.
2. Функциональная зависимость и корреляция.
3. Коэффициент корреляции.
4. Понятие о регрессии.
5. Построение эмпирических рядов регрессии.
6. Уравнение регрессии.
7. Коэффициенты регрессии.

Отличительной чертой биологических объектов является многообразие признаков, характеризующих каждый из них. Например, организмы можно характеризовать возрастом, весом, размерами и т.д. При этом описываемые признаки часто бывают взаимообусловлены. Например, чем старше организм, тем большими размерами он характеризуется. В простейшем случае связь между переменными строго однозначна. Например, вес древесины одного вида полностью определяется ее объемом. Такого рода зависимость называют *функциональной*, когда каждому значению независимой переменной соответствует только одно значение зависимой. Для биологических объектов редко связь между их характеристиками бывает менее «жесткой»: объекты с одинаковыми значениями одного признака имеют, как правило, разные значения по другим признакам. Такую связь между вариациями разных признаков называют *корреляцией*. По взаимонаправленности связь может быть прямой – когда с увеличением значений одного признака в общем увеличиваются значения другого, и обратной – когда с увеличением значений одного признака значения другого признака уменьшаются.

По форме связь может быть прямолинейной (линейной) и криволинейной (нелинейной). Основным мерилом связи, существующей между биологическими признаками, служит коэффициент *корреляции*. Он показывает степень приближения корреляционной связи к функциональной (для которой всегда равен единице) и колеблется в пределах от минус (для обратной связи) до плюс (для прямой связи) единицы. Значение коэффициента корреляции, равное нулю или близкое к нулю, говорит лишь об отсутствии прямолинейной связи, но не указывает на наличие или отсутствие криволинейной связи, которая при этом может быть тесной.

Рабочая формула, по которой обычно вычисляется коэффициент корреляции во всех случаях, когда варианты не группируются по классам: $r = \frac{\sum a_x a_y}{n\sigma_x \sigma_y}$.

В числителе этой формулы стоит сумма произведений отклонений вариант от средней арифметической по одному ряду *(X)* на соответствующие отклонения вариант от средней арифметической по другому ряду (У), т.е. $a_x = x_x - \bar{x}_x$ и $a_y = x_y - \bar{x}_y$; $\sum a_x a_y = \sum (x_x - \bar{x}_x)(x_y - \bar{x}_y)$.

По величине коэффициента корреляции можно установить характер связи (таблица).

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Квадратическая ошибка коэффициента корреляции $m_r = \pm\frac{1-r^2}{\sqrt{n}}$.

Значение коэффициента корреляции оценивается с помощью критерия достоверности Стьюдента: $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$ или с помощью формулы $t = r/m_r$.

При оценке степени взаимосвязи статистических величин важно провести математическое моделирование, т.е. подобрать аналитическое уравнение, которое соответствовало бы природе изучаемого явления с целью предсказания поведения независимой характеристики объекта при изменении зависимого параметра. Динамика взаимной зависимости между переменными величинами получила название *регрессии*, а методика исследования регрессии носит название *регрессионного анализа*.

Ряды регрессии выражаются не только графически, но и аналитически при помощи следующих уравнений: $Y' = \bar{y} + Ry/x(x - \bar{x})$ – уравнение регрессии *Y* по *X;* $X' = \bar{x} + Rx/y(y - \bar{y})$ – уравнение регрессии *X* по *Y*.

Здесь *Y'* и *X'* – теоретические, т.е. вычисленные по эмпирическим данным, значения регрессии *Y/X* и *X/Y;* $\bar{y}$ и $\bar{x}$ – средние арифметические рядов распределения Y и X; R – коэффициент регрессии, который определяется по следующим аналогичным формулам:

$Ry/x = \frac{\sum a_x a_y}{\sum a_x^2}$ – коэффициент регрессии *Y.X;*

$Rx/y = \frac{\sum a_x a_y}{\sum a_y^2}$ – коэффициент регрессии *X.Y,*

$a_x = x - \bar{x}$ и $a_y = y - \bar{y}$.

Когда известны средние квадратические отклонения варьирующих признаков, т.е. σ_x и σ_y, а также вычислен коэффициент корреляции (r), коэффициенты регрессии определяются по формулам: $Ry/x = r \cdot \frac{\sigma_y}{\sigma_x}$ и $Rx/y = r \cdot \frac{\sigma_x}{\sigma_y}$.

Коэффициент регрессии сопровождается средней квадратической ошибкой, которая вычисляется по формулам: $m_{R_{y/x}} = \frac{\sigma_{y/x}}{\sqrt{\sum a_x^2}}$ и $m_{R_{x/y}} = \frac{\sigma_{x/y}}{\sqrt{\sum a_y^2}}$.

Задание

- Рассчитать коэффициенты корреляции и регрессии для биологических признаков, измеренных на 1 занятии (длина листа и высота растения, длина и ширина листа, высота и вес растения).
- Оценить степень связи между признаками

Лабораторная работа № 11
РЕШЕНИЕ ЗАДАЧ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ СРЕДСТВАМИ MS EXCEL

Цель работы: освоить применение стандартных функций MS Excel для решения задач описательной статистики.

Контрольные вопросы

1. Ввод исходных данных.
2. Вычисление размаха (вариации).
3. Оценка среднего, среднеквадратичного отклонения и дисперсии, асимметрии и эксцесса.
4. Построение таблицы частот и гистограммы.

Пример 1. Проведите анализ данных в рамках описательной статистики с использованием средств **Вставка функций** и **Мастер диаграмм** MS Excel.

1. Запустите MS Excel: **Пуск/Программы/ MS Excel** и переименуйте ярлычок рабочего листа **Лист 1**: двойной щелчок по ярлыч-

ку и напечатайте поверх выделения **Статистика 1:** введите исходные данные и заголовки статистической таблицы по **Образцу 1:** выделите ячейку А1 щелчком мыши/введите текст заголовка и зафиксируйте щелчком по инструменту Enter ✓/расположите заголовок по центру столбцов А-Е – выделите ячейки А1:Е1 и щелкните инструмент **Объединить и поместить в центре**/ аналогично выделяя последовательно ячейки А2 – Е11, введите числа исходных данных таблицы:

x, прирост населения в 50 городах				
27	36	34	46	34
28	29	37	41	43
40	33	50	37	41
32	29	43	34	32
30	43	54	42	47
35	49	49	54	36
36	51	36	24	35
25	33	38	38	36
29	51	32	36	53
30	55	44	46	38

Среднее
Среднеквадратичное отклонение (СТАНДОТКЛОН)
Дисперсия (ДИСПА)
Медиана (МЕДИАНА)
Мода (МОДА)
Асимметрия (СКОС)
Эксцесс (ЭКСЦЕСС)
Наименьшее (МИН)
Наибольшее (МАКС)
Кол-во выборок 50

В ячейках G2 – G14 заголовки строк статистической таблицы и число выборок:

1. Выполните расчеты указанных в заголовках строк статистической таблицы параметров, вставляя при помощи средства **Вставка функций** расчетные формулы. Например, для расчета среднего выделите щелчком мыши ячейку Н2, щелкните инструмент **Вставка функций**/ в окне **Мастер функций** в поле **Категории** щелкните **Статистические,** в поле **Функция** при помощи полосы прокрутки пролистайте список названия функций, найдите и щелкните **СРЗНАЧ** и **ОК**/ в окне вставки функции справа от поля **Число 1** щелкните кнопку сворачивания/ выделите мышью диапазон ячеек А2:Е11 (удерживая

левую кнопку мыши)/ в свернутом окне вставки функции щелкните кнопку разворачивания/ОК; аналогично вставьте остальные формулы.

2. Сформируйте таблицу частот исследуемой величины, выполнив группировку данных и расчеты непосредственным вводом формул и при помощи средства **Вставка функций:**

– вставьте формулу для вычисления минимального числа интервалов группирования при помощи средства **Вставка функций:** выделите ячейку A14 и введите «мин. кол-во интервалов», выделите ячейку B14/ инструмент **Вставка функций**/ в поле **Категории** щелкните **Математические**/ в поле **Функция** найдите и выберите **ОКРУГЛ** и **ОК**/ в окне вставки функции установите курсор в поле **Число разрядов** и введите 0 (округление до целого числа), установите курсор в поле **Число** и введите 5* (множитель)/ в инструменте выбора функции (левый верхний угол рабочей книги) щелкните кнопку списка и выберите позицию **Другие функции…**/ в окне **Мастер функций** выберите категорию **LOG10** из категории **Математические** и **ОК**/ в окне вставки функции в поле **Число 1** введите ссылку с числом выборок **H14** и **ОК**;

– вставьте формулу для расчета ширины интервала при помощи ввода с клавиатуры: выделите ячейку A15 и введите «ширина интервала», выделите ячейку B15/ введите знак =(равно) и знак **(**(скобка)/ щелкните ячейку с максимальным значением H10 и нажмите клавишу **F4** для перехода к абсолютной ссылке/ введите знак **–** /щелкните ячейку с минимальным значением H9 и нажмите клавишу F4/ введите знак **)** (скобка) и знак **/** (наклонная черта) и щелкните ячейку B14 с числом интервалов/ **Enter**;

– аналогично в ячейки A20–A27 вставьте формулы для вычисления правых границ интервалов: щелкните ячейку A20, введите знак =(равно)/ щелкните ячейку с минимальным значением H9 и нажмите клавишу **F4** для перехода к абсолютной ссылке/ введите знак + (плюс) и щелкните ячейку с значением ширины интервала B15/ **Enter;** в ячейку A21 введите формулу =A20+ B15; в ячейки A22 и ниже растяните формулу из ячейки A20 при помощи автозаполнения: после ввода формулы в A21 укажите на нижний правый угол ячейки A21 до появления маркера автозаполнения в форме +, нажмите левую кнопку мыши и, удерживая ее, протяните выделение ячейки до A27 и отпустите кнопку мыши;

– вставьте формулу для расчета частот с применением функции массивов: выделите диапазон ячеек B20–B27/ инструмент **Вставка функций**/ найдите и выберите функцию **ЧАСТОТА** из категории **Статистические** и **ОК**/ в окне вставки функции справа от поля **Массив данных** щелкните кнопку сворачивания/ выделите мышью диапазон ячеек A2:E11/ щелкните кнопку разворачивания/ справа от поля

Массив интервалов щелкните кнопку сворачивания/ выделите мышью диапазон ячеек A20:A27/ щелкните кнопку разворачивания/ одновременно нажмите клавиши **Ctrl, Shift, Enter** для фиксации функции массива.

3. Постройте гистограмму для исследуемой величины с применением мастера диаграмм: выделите диапазон ячеек с таблицей частот A20:B27/ инструмент **Мастер диаграмм**/ на вкладке **Нестандартные** в поле **Тип** выберите **График/Гистограмма 2** и кнопка **Далее**/в окне **Исходные данные** на вкладке **Диапазон данных** включите переключатель **в столбцах**/ на вкладке **Ряд** щелкните кнопку сворачивания справа от поля **Подписи по оси X**/ выделите диапазон ячеек A20:A27 и щелкните кнопку разворачивания/ в поле **Подписи второй оси X** внесите диапазон ячеек B20:B27 и кнопка **Далее**/ в окне размещение диаграммы включите переключатель **на имеющемся листе** и **ОК**.

Пример 2. Выполните процедуру генерации случайных чисел и проанализируйте их с помощью средств **Анализ данных** и **Мастер диаграмм** MS Excel.

1. Перейдите на свободный рабочий лист книги и переименуйте его в **Генерация данных**.

2. Подключите надстройку **Пакет анализа** MS Excel: **Сервис/Надстройки**/ в окне Надстройки установите флажок **Пакет анализа** и **ОК**.

3. Выполните генерацию 30 случайных чисел, распределенных в соответствии с нормальным законом с нулевым средним и дисперсией 1: щелкните ячейку A1 и **Сервис/Анализ данных**/ в поле со списком **Инструмент анализа** щелкните позицию **Генерация случайных чисел** и **ОК**/ в поле **Число переменных** введите 1, в поле **Число случайных чисел** введите 30, раскройте список поля **Распределение** и выберите позицию **Нормальное**, введите в полях **Среднее** – 0, **Стандартное отклонение** – 1, в разделе **Параметры вывода** включите переключатель **выходной интервал**, щелкните кнопку сворачивания/ щелкните ячейку A1 и кнопку разворачивания, **ОК**.

4. Измените разрядность данных, уменьшите число знаков после запятой до двух: выделите диапазон ячеек A1:A30/ щелкните инструмент **Уменьшить разрядность** четыре раза.

5. Выполните процедуру описательной статистики по сгенерированным данным: **Сервис/Анализ данных/Описательная статистика** и **ОК**/ в окне Описательная статистика в поле Входной интервал введите ссылку на диапазон ячеек A1:A30/ в разделе **Группирование** включите переключатель **по столбцам** и уберите флажок **Метки в первой строке**/ в разделе **Параметры вывода** включите переключатель **Выходной интервал** и щелкните ячейку C1/ установите флажок **Итоговая статистика** и **ОК.**

6. Постройте гистограмму по данным столбца А: **Сервис/Анализ данных/Гистограмма** и **ОК**/ в окне Гистограмма в разделе **Входные данные** в поле **Входной интервал** введите ссылку на диапазон ячеек А1:А30 и установите флажок Метки/ в разделе параметры вывода включите переключатель **Выходной интервал** и укажите любую свободную ячейку рабочего листа/ установите флажок **Интегральный процент** и **Вывод графика** и **ОК**. Добавьте на построенную гистограмму 2 линии тренда: полиноминального со степенью 4 и скользящего среднего на 2 точки: щелкните правой клавишей мыши по рядам значений, на появившемся контекстном меню выбрать позицию **Добавить линию тренда/** на вкладке тип выбрать **Полиноминальная**, в поле **степень** ввести 4 и т.д.

Лабораторная работа № 12
КОРРЕЛЯЦИОННЫЙ АНАЛИЗ, АНАЛИЗ ФАКТОРОВ В MS EXCEL

Цель работы: освоить применение стандартных функций MS Excel для решения задач анализа связей, применение пакета анализа MS Excel для решения задач анализа связей.

Контрольные вопросы

1. Построение диаграммы рассеяния.
2. Расчет корреляции с помощью стандартных функций MS Excel.
3. Однофакторный дисперсионный анализ средствами пакета анализа MS Excel.

Пример 1. Проведите визуальный анализ данных и расчет коэффициента корреляции в рамках задачи проверки наличия связи между двумя переменными с использованием средств **Вставка функций** и **Мастер диаграмм** MS Excel.

1. Запустите MS Excel и переименуйте ярлычок рабочего листа в **Диаграмма рассеяния.**
2. Сформируйте массив исходных данных результатов измерений длины первого молярного *x* и второго молярного *y* зубов у ископаемого млекопитающего по образцу.
3. Используя средство **Мастер диаграмм**, постройте диаграмму рассеяния: выделите диапазон ячеек **A2:B21**/ инструмент **Мастер диаграмм/** тип диаграммы **Точечная** вида 1 и **Далее**/ на вкладке **Диапазон данных** установите флажок **Ряды в столбцах** и **Далее**/ введите заголовки диаграммы **Диаграмма рассеяния**, оси X – **первый молярный**, оси У – **второй молярный**/ снимите отображение легенды/установите отображение основных и промежуточных линий

сетки по обеим осям/ расположите диаграмму на имеющемся листе/ добавьте линейный тренд с включением отображения уравнения (вкладка **Параметры** окна **Линия тренда**).

4. Проанализируйте полученные данные и сделайте предварительный вывод о наличии линейной связи между рассматриваемыми признаками.

x	y
10,7	11,2
10,8	10,9
10,6	10,5
10,7	9,6
10,1	11,2
11,2	11,3
11,4	12,2
12,1	12,1
12,3	11,7
12,0	11,0
12,3	13,2
12,7	13,0
12,9	12,2
12,8	13,4
13,1	12,6
13,3	12,2
13,3	12,0
13,4	11,2
12,7	11,4
12,5	11,4

5. Рассчитайте коэффициент корреляции для исследуемых выборок. Выделите ячейку **A22** и введите текст **Коэффициент корреляции**/ выделите ячейку **B22**/щелкните инструмент **Вставка функций**/в окне **Мастер функций** в поле **Категории** щелкните **Статистические,** в поле **Функция** при помощи полосы прокрутки пролистайте список названия функций, найдите и щелкните **КОРЕЛЛ** и **ОК**/в полях **Массив1** и **Массив2** введите последовательно ссылки на диапазоны **A2:A21** и **B2:B21** соответственно и **ОК**.

6. Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента. Выделите ячейку **A23** и введите текст **t-статистика**/выделите ячейку **B23** и введите формулу по образцу =**B22*КОРЕНЬ(СЧЕТ(A2:A21)-2)/КОРЕНЬ(1-B22*B22)**. Выделите ячейку **A24** и введите текст **Критическое значение**. Выделите ячейку **B24** и введите формулу по образцу **=СТЬЮДРАСПОБР (0,05;СЧЕТ(A2:A21)-2).** Если значение **t** больше табличного (критического), то принимается наличие значимой линейной связи (отвергается предположение об отсутствии связи).

Пример 2. Проведите визуальный и корреляционный анализ данных в рамках задачи проверки наличия связи между двумя переменными с использованием средств **Анализ данных** и **Мастер диаграмм** MS Excel.

1. Перейдите на **Лист 2** книги и переименуйте его ярлычок в **Корелл анализ**. Скопируйте предыдущий пример на новый лист: выделите диапазон ячеек **A2:B21** листа **Диаграмма рассеяния**/щелкните правой кнопкой мыши для вызова контекстного меню/**Копировать**/перейдите на новый рабочий лист/выделите ячейку **A1**/щелкните правой кнопкой мыши и **Вставить**.

2. Рассчитайте корреляцию для исследуемых данных, используя группу **Корреляция** средства **Анализ данных**: **Сервис/Анализ данных**/в поле со списком **Инструмент анализа** щелкните позицию

Корреляция и **ОК**/в поле **Входной интервал** введите ссылку на диапазон **A2:B21**/установите флажки в **Группирование по столбцам** и **Метки в первой строке**/ в поле выходной интервал укажите ссылку на ячейку D1 и **ОК**.

Пример 3. Проведите однофакторный дисперсионный анализ влияния одного фактора на характеристики нескольких экспериментальных групп с использованием средств **Анализ данных** MS Excel.

1. Перейдите на **Лист 3** книги и переименуйте его ярлычок в **Однофакт анализ**.

2. Сформируйте массив исходных данных лабораторных испытаний влияния фактора погоды на изменение продолжительности систолической остановки сердца при введении хлорида бария для четырех групп экспериментальных животных по образцу:

	A	B	C	D	E
1		Показатели экспериментальных групп			
2	Погода	1	2	3	4
3	Тихая погода	13,8	11	13,7	12,1
4	Ветер и вьюга	16	12,2	15,8	14,3

3. Запустите процедуру однофакторного дисперсионного анализа: **Сервис**/**Анализ данных**/**Однофакторный дисперсионный анализ**/укажите диапазон входных значений **B3:E4**, группирование **по столбцам**, флажок **Метки** снимите, укажите ячейку выходного диапазона ячеек **B6**.

4. Проанализируйте полученные результаты: сравните дисперсию внутри групп (характеризует влияние случайной составляющей) и между группами (характеризует влияние изучаемого фактора – погоды). Если они значимо отличаются (уровень значимости P=0,05), то фактор считается оказывающим статистически значимое влияние на исследуемую переменную. Сравните расчетное F и критическое значения статистики Фишера. Отличие считается значимым, если расчетное значение больше критического.

5. Сформируйте дополнительную расчетную формулу для принятия решения **Влияет ли фактор?** по образцу. Выделите ячейку E23 и введите текст **Влияет ли фактор?**/щелкните ячейку F23 и введите формулу =ЕСЛИ(F18>H18;«да»;«нет»).

Лабораторная работа № 13

ПРОВЕРКА ГИПОТЕЗ В MS EXCEL. ПАРАМЕТРИЧЕСКИЕ И НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

Цель работы: освоить применение стандартных функций MS Excel для решения задач проверки гипотез, применение пакета анализа для решения задач проверки гипотез.

Контрольные вопросы

1. Нулевая гипотеза.
2. Эмпирический тест на нормальность.
3. Проверка гипотезы о равенстве среднего заданному значению.
4. Проверка гипотезы о распределении по критерию хи-квадрат.

Пример 1. Проведите анализ данных в рамках задачи проверки гипотезы о распределении при помощи эмпирического теста на нормальность с использованием средства **Вставка функций** MS Excel.

1. Запустите MS Excel и сохраните созданную при запуске книгу под именем **Примеры гипотезы.**
2. Переименуйте ярлычок рабочего листа в **Тест норм**.
3. Сформируйте массив исходных данных результатов 100 замеров отклонений от номинального размера объекта по образцу 1 в диапазоне ячеек **A2:J11**:

Образец 1

x (норм)									
48	39	43	44	34	34	32	43	40	46
25	31	34	49	39	37	45	48	41	49
43	46	34	35	42	32	41	34	42	42
38	40	46	47	34	42	38	40	38	36
30	43	41	40	40	35	35	41	38	45
37	42	38	36	44	39	332	48	43	39
43	30	44	36	42	34	49	49	49	51
37	30	50	48	44	35	45	34	33	41
43	45	44	34	33	39	41	39	46	31
40	52	45	39	35	45	33	42	42	36

4. Рассчитайте среднее и среднеквадратичное отклонение, разместив расчетные формулы в ячейках **L12, L13** соответственно: =СРЗНАЧ(A2:J11) и =СТАНДОТКЛОН (B2:J11).

5. Рассчитайте массив отклонений выборочных значений от среднего: активизируйте ячейку **A15**/инструмент **Вставка функций**/в поле **Категории** выберите **Математические**/ в поле **Функция** найди-

те и выберите **ABS** и **ОК**/в окне вставки функции установите курсор в поле **Число,** введите ссылку **A2**/введите знак разности «-» /введите ссылку **L12**, перейдите к абсолютной ссылке клавишей **F4** и **ОК**/используя маркер автозаполнения ячейки **A15**, растяните формулу в ячейки **B15:J15/**не снимая выделение с диапазона **A15:J15**, используя маркер автозаполнения выделенного диапазона, растяните формулу в ячейки **A16:J24**.

6. Сформируйте таблицу проверки условий эмпирического теста на нормальность и вставьте расчетные формулы согласно Образцу 2 в ячейки **L15:L22:**

Образец 2

	L	M
15	Значение 3S	=3*L13
16	Значение 0,625 S	=L13*0,625
17	Число выборок	100
18		
19	Условие	Выполняется ли условие
20	<3S	=ЕСЛИ(СЧЕТЕСЛИ(A15:J24;”<16,52”)>0,997*100; “да”; ”нет”)
21	<S	=ЕСЛИ(СЧЕТЕСЛИ(A15:J24;”<5,508”)>0,683*100; “да”;“нет”)
22	<0,625S	=ЕСЛИ(СЧЕТЕСЛИ(A15:J24;”<3,44”)>0,5*100; “да”;“нет”)

7. Проинтерпретируйте полученные результаты: в случае невыполнения хотя бы одного из условий эмпирического теста необходима дополнительная проверка исходной гипотезы о нормальности при помощи, например, критерия хи-квадрат. При выполнении всех трех условий гипотеза о нормальном законе распределения принимается.

Пример 2. Проведите анализ данных в рамках задачи проверки гипотезы о распределении при помощи критерия согласия хи-квадрат с использованием средств **Вставка функций** и **Анализ данных** MS Excel.

1. Перейдите на рабочий лист 2 книги **Примеры гипотезы** и переименуйте его в **Тест хи-квадрат**. В ячейке **A1** введите заголовок столбца данных **x, норм**.

2. Скопируйте исходные данные из диапазона ячеек листа 1 **A2:J11** в позицию, начиная с ячейки **A2**.

3. Реорганизуйте скопированный массив данных на листе при помощи приема перемещения диапазонов ячеек так, чтобы данные располагались в одном столбце **A**: результирующий массив должен занимать диапазон ячеек **A2:A101**.

4. При помощи средства **Сервис/Анализ данных** рассчитайте по исходным данным описательную статистику и постройте таблицу частот и гистограмму (см. лабораторную работу №. 11) в диапазоне ячеек **C1:G12**.

5. Используя построенную таблицу частот и рассчитанные среднее и среднеквадратичное, а также стандартные и встроенные функции, сформируйте таблицу для расчета статистики хи-квадрат по образцу 3. Обратите внимание на ввод максимального значения вместо текста **Еще** в исходной таблице частот (ячейка **C29**) и расчетной формулы для дополнительного значения интервала группирования в ячейке **C30**. Эти действия необходимы для корректного вычисления теоретических частот. В столбце скорректированных частот выполнено объединение тех карманов, где значение частот менее 3. Это условие правильного применения критерия. Расчетная формула для числа степеней свободы распределения хи-квадрат определяется разностью числа карманов (с учетом их объединения) и числа 3 (количеству налагаемых связей плюс 1).

Образец 3

	C	D	E	F	G
18	карман	часто-та	скор-ректир. часто-та	теоретич. частота	хи-квадрат
19	25	1			
20	27,7	0			
21	30,4	3	=D19+D20+D21	=(НОРМРАСП(C22;F3;F7;ИСТИНА)-НОРМРАСП(C19;F3;F7;ИСТИНА))	=(E21-F21)^2/F21
22	33,1	8	8	=(НОРМРАСП(C23;F3;F7;ИСТИНА)-НОРМРАСП(C22;F3;F7;ИСТИНА))	=(E22-F22)^2/F22
23	35,8	14	14	=(НОРМРАСП(C24;F3;F7;ИСТИНА)-НОРМРАСП(C23;F3;F7;ИСТИНА))	=(E23-F23)^2/F23
24	38,5	12	12	=(НОРМРАСП(C25;F3;F7;ИСТИНА)-НОРМРАСП(C24;F3;F7;ИСТИНА))	=(E24-F24)^2/F24
25	41,2	19	19	=(НОРМРАСП(C26;F3;F7;ИСТИНА)-НОРМРАСП(C25;F3;F7;ИСТИНА))	=(E25-F25)^2/F25
26	43,9	15	15	=(НОРМРАСП(C27;F3;F7;ИСТИНА)-НОРМРАСП(C26;F3;F7;ИСТИНА))	=(E26-F26)^2/F26
27	46,6	15	15	=(НОРМРАСП(C28;F3;F7;ИСТИНА)-НОРМРАСП(C27;F3;F7;ИСТИНА))	=(E27-F27)^2/F27

	C	D	E	F	G
28	49,3	10	10	=(НОРМРАСП(C29;F3;F7;ИСТИНА)-НОРМ-РАСП(C28;F3;F7;ИСТИНА))	=(E28-F28)^2/F28
29	52	3	3	=(НОРМРАСП(C30;F3;F7;ИСТИНА)-НОРМ-РАСП(C29;F3;F7;ИСТИНА))	=(E29-F29)^2/F29
30	=C29+(C29-C28)			Статистика хи-квадрат	=СУММ(G21:G29)
31				Ошибка	0,05
32				Число степеней свободы	=СЧЕТ(E21:E29)-3
33				Табл. значение	=ХИ2ОБР(G31:G32)
34				Проверка условия	=ТСКВ(G33>G30; “да”; “нет”)

Пример 3. Проведите анализ данных в рамках задачи проверки гипотезы о равенстве среднего некоторому определенному значению (для данных, взятых из нормально распределенной генеральной совокупности) с использованием средств **Вставка функций** MS Excel.

1. Перейдите на рабочий **Лист 3** книги и переименуйте его в **Тест среднее**.

2. Сформируйте массив данных, используя функцию генерации случайных чисел средства **Анализ данных** для нормального распределения со средним 0,15 и стандартным отклонением 0,1 при числе выборок 20.

3. Рассчитайте среднее и дисперсию выборки, используя стандартные функции MS Excel.

4. Сформируйте таблицу для проверки критерия, исходя из того, что критериальное значение вычисляется по формуле $t = (m - A)\sqrt{n} / S^2$, где A – предполагаемая величина среднего (например, 0,09), а критическое значение вычисляется для распределения Стьюдента со значимостью α и числом степеней свободы n-1. Гипотеза о равенстве среднего отвергается, если по абсолютной величине критериальное значение больше верхней α/2%-й точки распределения Стьюдента: $|t_{критериальное}| > T\{ \alpha/2, n-1\}$. Для расчета распределения Стьюдента используйте функцию **СТЬЮДРАСПОБР**().

Применение пакета анализа MS Excel для решения задач проверки гипотез

Пример 4. Проведите анализ данных в рамках задачи проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности (следовательно, их равенстве) по критерию Фишера с использованием средств **Анализа данных** MS Excel.

1. Перейдите на рабочий **Лист 4** книги и переименуйте его в **Тест. рав. дисп**.

2. Сформируйте массив исходных данных двух независимых выборок значений измерения веса опытных животных по образцу 5.

Образец 5

№ группы	Результаты измерений									
1	55	73	50	71	63	59	66	74	58	69
2	43	39	50	47	47	38	51	48	37	42

3. Запустите процедуру проверки гипотезы: **Сервис/Анализ данных/Двухвыборочный тест для дисперсии**/в одноименном окне укажите диапазоны ячеек для 1 и 2 выборок в полях **Интервал переменной** .., введите уровень значимости **0,05** в поле **Альфа**, укажите верхнюю левую ячейку размещения результатов в поле **Выходной интервал** и **ОК**.

4. Проанализируйте полученные результаты. По условиям критерия нулевая гипотеза отвергается, если значение F статистики Фишера больше верхнего критического или меньше нижнего.

5. Постройте для обеих выборок гистограммы с полиномиальным трендом.

Пример 5. Проведите анализ данных в рамках задачи проверки гипотезы о равенстве средних при неравных дисперсиях и объемах выборок по критерию Стьюдента с использованием средств **Анализа данных** MS Excel.

1. Перейдите на рабочий **Лист 5** книги и переименуйте его в **Тест. рав. сред1**.

2. Сформируйте массив данных, используя функцию генерации случайных чисел средства **Анализ данных** для нормального распределения со средним 1 и стандартным отклонением 1 и 2 для двух выборок объемом 15 и 20 чисел соответственно, разместив их в столбцах **B** и **C**.

3. Запустите процедуру проверки гипотезы **Сервис/Анализ данных/Двухвыборочный t-тест с различными дисперсиями**/ в одноименном окне укажите диапазоны ячеек для 1 и 2 выборок в полях **Интервал переменной** .., введите уровень значимости **0,05** в поле **Альфа**, укажите верхнюю левую ячейку размещения результатов в поле **Выходной интервал** и **ОК**.

4. Проанализируйте полученные результаты. По условиям критерия нулевая гипотеза отвергается, если значение t-статистики Стьюдента по абсолютной величине больше верхней точки распределения или критического значения.

5. Постройте для обеих выборок гистограммы с трендом скользящего среднего.

***Пример 6*.** Проведите анализ данных в рамках задачи проверки гипотезы о равенстве средних при равных дисперсиях по критерию Стьюдента с использованием средств **Анализа данных** MS Excel.

1. Перейдите на рабочий **Лист 6** книги и переименуйте его в **Тест. рав. сред2**.

2. Сформируйте массив данных, используя функцию генерации случайных чисел средства **Анализ данных** для нормального распределения со средним 2 и стандартным отклонением 2 для двух выборок объемом 40 чисел каждая, разместив их в столбцах **B** и **C**.

3. Запустите процедуру проверки гипотезы **Сервис/Анализ данных/Двухвыборочный t-тест с одинаковыми дисперсиями**/ в одноименном окне укажите диапазоны ячеек для 1 и 2 выборок в полях **Интервал переменной** .., введите уровень значимости **0,05** в поле **Альфа**, укажите верхнюю левую ячейку размещения результатов в поле **Выходной интервал** и **ОК**. При вводе значения в поле **Гипотетическая средняя разность** проверяется гипотеза о разности значений средних двух выборок.

4. Проанализируйте полученные результаты и примите решение.

5. Постройте для обеих выборок гистограммы с полиномиальным трендом.

Лабораторная работа № 14
ИСПОЛЬЗОВАНИЕ ПАКЕТА STATISTICA ДЛЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

1. Первичная обработка опытных данных помощи модуля Basic Statistics/Tables.

Расчет описательных статистик производится при помощи модуля Basic Statistic/Tables. В этом модуле объединены наиболее часто использующиеся на начальном этапе обработки данных процедуры.
В стартовой панели модуля приводится перечень статистических процедур этого модуля (рис. 1).

Descriptive statistics – Описательные статистики;
Correlation matrices – Корреляционные матрицы;
t-test for independent samples – t-тест для независимых выборок;
t-test for dependent samples – t-тест для зависимых выборок;
Breakdown = one-way ANOVA – Классификация и однофакторный дисперсионный анализ и др.

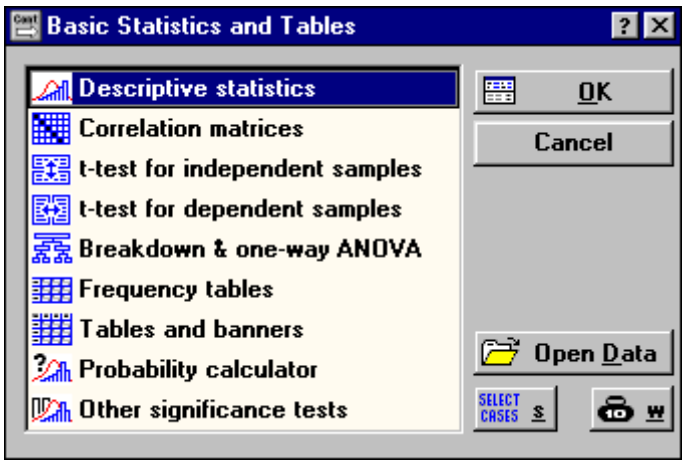


Рис. 1. **Стартовое окно модуля с перечнем статистических процедур.**

1.1. Процедура Descriptive statistics (Описательные статистики)

Рассмотрим возможности этой процедуры на примере.

Имеется выборка объемом 50 измерений, представляющая собой результаты обмера 1-летних сеянцев сосны обыкновенной. Файл данных (рис. 2) содержит 4 переменных:

Data: PRIMER1.STA 4v * 50c

NUMVAL	Биометрические показатели 1-летних сеянцев сосны с			
	1 VAR1	2 VAR2	3 VAR3	4 VAR4
1	3,6	1,33	20,5	3,100
2	2,3	1,00	19,5	2,500
3	2,9	1,54	14,2	2,600
4	2,6	1,10	14,9	2,500
5	3,1	1,45	14,5	2,600
6	3,2	1,33	15,8	2,500
7	3,5	1,25	11,4	2,800
8	4,2	1,65	17,6	3,300
9	5,0	1,42	19,4	2,600
10	3,0	1,36	21,5	2,700
11	3,1	1,20	17,6	2,800
12	2,9	1,17	18,3	2,700
13	4,0	1,75	17,5	2,300
14	2,2	1,62	21,9	3,500
15	3,1	1,24	16,0	2,400

Рис. 2. **Окно файла данных.**

VAR1 – длина надземной части сеянцев, см;
VAR2 – диаметр у корневой шейки, мм;
VAR3 – длина корней, см;

VAR4 – длина хвои, см.

После выбора процедуры Descriptive statistics на экране появится одноименное диалоговое окно (рис. 3).

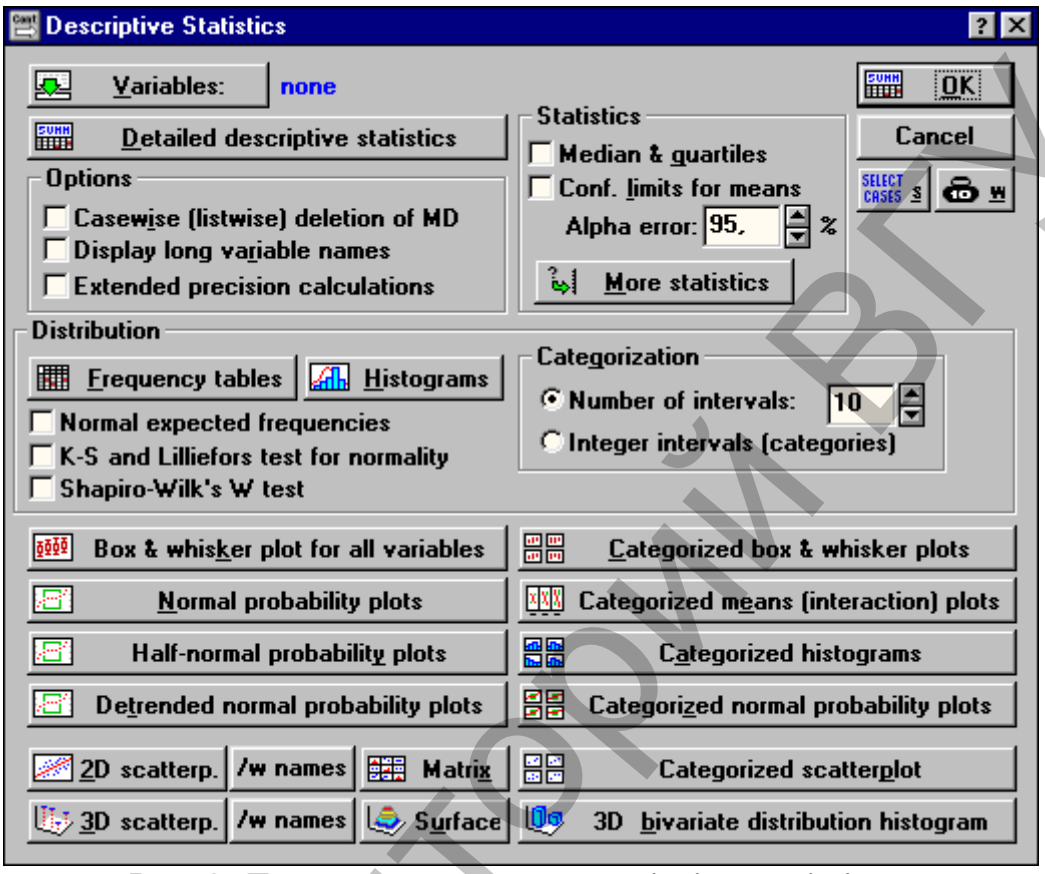


Рис. 3. **Диалоговое окно «Descriptive statistics».**

В этом окне при помощи кнопки **Variables** следует выбрать переменные для анализа (рис. 4);

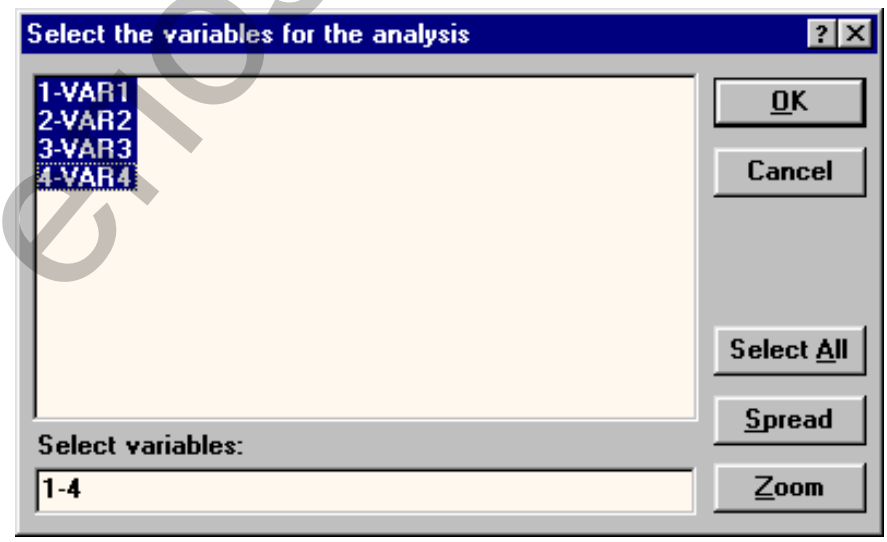


Рис. 4. **Окно выбора переменных.**

На первом этапе обработки данных часто возникает необходимость в их группировке. Группировка позволяет представить первичные данные в компактном виде, выявить закономерности варьирования изучаемого признака. Количество классов можно приблизительно наметить, пользуясь следующими придержками (Лакин, 1990): при количестве наблюдений 25–40 – 5–6 классов, при количестве наблюдений 40–60 – 6–8 классов, 60–100 – 7–10, 100–200 наблюдений – 8–12, более 200 наблюдений – 10–15 классов.

Для построения гистограмм и таблиц частот используется группа кнопок **Distribution** окна Descriptive statistics. Число классов (интервалов) группировки данных устанавливается при помощи счетчика переключателя **Number of intervals** окна Descriptive statistics. Справа от кнопок Distribution находятся две опции **Categorization** (Группировка), позволяющие задать число интервалов группировки или установить величину интервала равную целому числу. Если заактивировать переключатель **Integer intervals (categories)**, то классы (интервалы) группировки будут представлять из себя целые числа.

Результаты группировки длины сеянцев (переменная Var1) представлены в табл. 1.

Таблица 1

Результаты группировки замеров высот

		Cumul.	Percent	Cumul %	% of all
Интервал	Count	Count	of Valid	of Valid	Cases
длин, м	(к-во)	(к-во с накоплением)	(%)	(% с накоплением)	(% от общего к-ва)
1,0 < x <= 2,0	0	0	0	0	0
2,0 < x <= 3,0	15	15	30	30	30
3,0< x <= 4,0	23	38	46	76	46
4,0 < x <= 5,0	5	43	10	86	10
5,0< x <= 6,0	6	49	12	98	12
6,0 < x <= 7,0	1	50	2	100	2

Представим распределение переменных на гистограммах. Для этого предназначена кнопка **Histograms** окна Descriptive statistics.

На гистограмму при необходимости можно наложить плотность нормального распределения, проверить близость распределения к нормальному виду при помощи критериев Колмогорова–Смирнова, Лилиефорса; вычислить статистику Шапиро–Уилкса. Для этого в группе опций Distribution необходимо установить флажок напротив соответствующих статистик. Значения статистик показываются прямо на гистограммах.

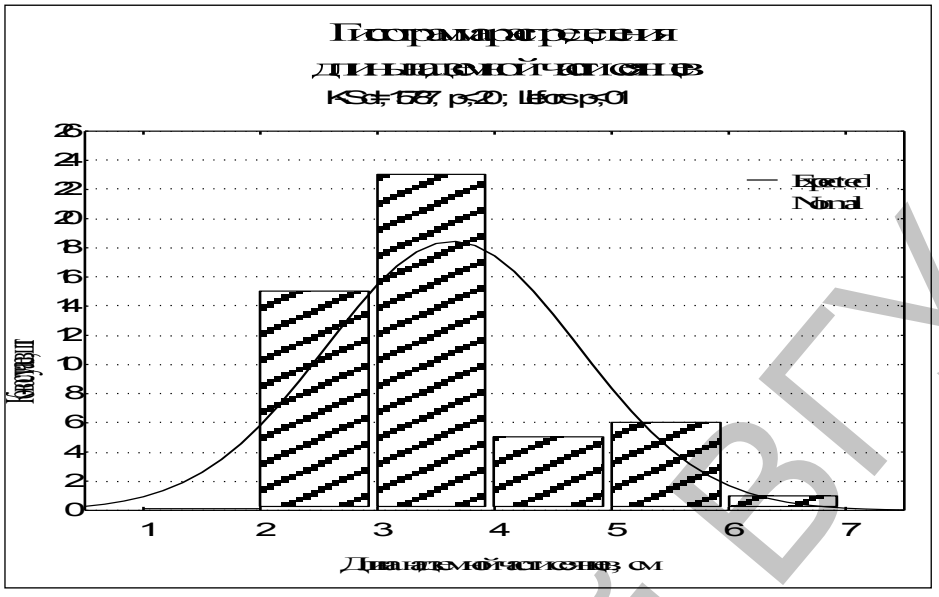


Рис. 5. **Гистограмма распределения длины надземной части сеянцев.**

На рис. 5 в качестве примера приводится гистограмма распределения длины надземной части сеянцев (переменной Var1).

На гистограмме показана кривая плотности нормального распределения, а также критерий Колмогорова–Смирнова (d). Статистика Колмогорова–Смирнова оказалась равной 0,157. Чем меньше величина этой статистики, тем ближе распределение случайной величины к нормальному. Вероятность нулевой гипотезы (p) менее 0,20.

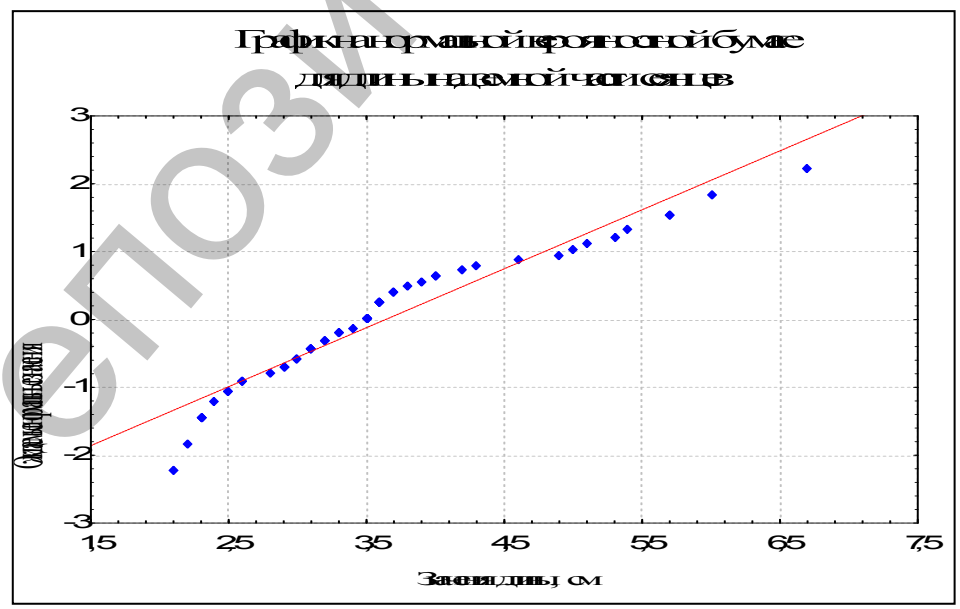


Рис. 6. **График на нормальной вероятностной бумаге для выборки длин надземной части сеянцев.**

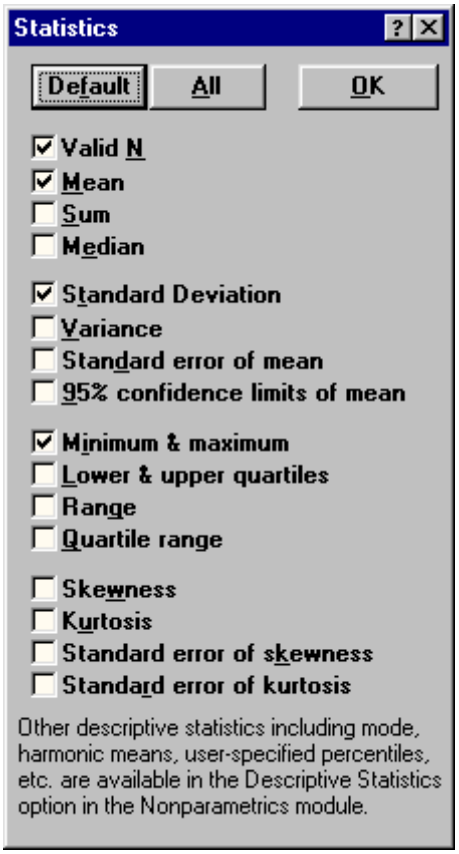


Рис. 7. **Окно выбора статистик.**

О нормальности распределения можно судить по графику на нормальной вероятностной бумаге. Его легко построить при помощи опции **Normal probability plots** окна «Descriptive statistics» (рис. 3). Чем ближе распределение к нормальному виду, тем лучше значения ложатся на прямую линию (рис. 6). Этот метод оценки является фактически глазомерным. В сомнительных случаях проверку на нормальность можно продолжить с использованием специальных статистических критериев (Колмогорова-Смирнова, Омега-квадрат (w^2)). Однако детальная проверка гипотезы о нормальности выборки требует довольно значительных объемов выборки (по мнению некоторых авторов не менее 100 наблюдений).

Чтобы выбрать статистики, подлежащие вычислению, удобнее всего воспользоваться кнопкой **More statistics** (рис. 7).

Valid N – объем выборки;

Mean – средняя арифметическая;

Среднее значение случайной величины представляет собой наиболее типичное, наиболее вероятное ее значение, своеобразный центр, вокруг которого разбросаны все значения признака.

Sum – сумма;

Median – медиана;

Медианой является такое значение случайной величины, которое разделяет все случаи выборки на две равные по численности части.

Standard Deviation – стандартное отклонение;

Стандартное отклонение (или среднее квадратическое отклонение) является мерой изменчивости (вариации) признака. Оно показывает на какую величину в среднем отклоняются случаи от среднего значения признака. Особенно большое значение имеет при исследовании нормальных распределений. В нормальном распределении 68% всех случаев лежит в интервале $\pm$ одного отклонения от среднего, 95% – $\pm$ двух стандартных отклонений от среднего и 99,7% всех случаев – в интервале $\pm$ трех стандартных отклонений от среднего.

Variance – дисперсия.

Дисперсия является мерой изменчивости, вариации признака и представляет собой средний квадрат отклонений случаев от среднего значения признака. В отличие от других показателей вариации дисперсия может быть разложена на составные части, что позволяет тем самым оценить влияние различных факторов на вариацию признака. Дисперсия – один из существеннейших показателей, характеризующих явление или процесс, один из основных критериев возможности создания достаточно точных моделей.

Standard error of mean – стандартная ошибка среднего;

Стандартная ошибка среднего – это величина, на которую отличается среднее значение выборки от среднего значения генеральной совокупности при условии, что распределение близко к нормальному. С вероятностью 0,68 можно утверждать, что среднее значение генеральной совокупности лежит в интервале $\pm$ одной стандартной ошибки от среднего, с вероятностью 0,95 – в интервале $\pm$ двух стандартных ошибок от среднего и с вероятностью 0,99 – среднее значение генеральной совокупности лежит в интервале $\pm$ трех стандартных ошибок от среднего.

95% confidence limits of mean – 95%-й доверительный интервал для среднего;

Интервал, в который с вероятностью 0,95 попадает среднее значение признака генеральной совокупности.

Minimum, maximum – минимальное и максимальное значения;

Lower, upper quartiles – нижний и верхний квартили;

Верхний квартиль – это такое значение случайной величины, больше которого по величине 25% случаев выборки. Верхний квартиль – это такое значение случайной величины, меньше которого по величине 25% случаев выборки.

Range – размах;

Расстояние между наибольшим (maximum) и наименьшим (minimum) значениями признака.

Quartile range – интерквартильная широта;

Расстояние между нижним и верхним квартилями.

Skewness – асимметрия;

Асимметрия характеризует степень смещения вариационного ряда относительно среднего значения по величине и направлению. В симметричной кривой коэффициент асимметрии равен нулю. Если правая ветвь кривой, начиная от вершины, больше левой (правосторонняя асимметрия), то коэффициент асимметрии больше нуля. Если левая ветвь кривой больше правой (левосторонняя асимметрия), то коэффициент асимметрии меньше нуля. Асимметрия менее 0,5 считается малой.

Standard *error of skewness* – стандартная ошибка асимметрии;

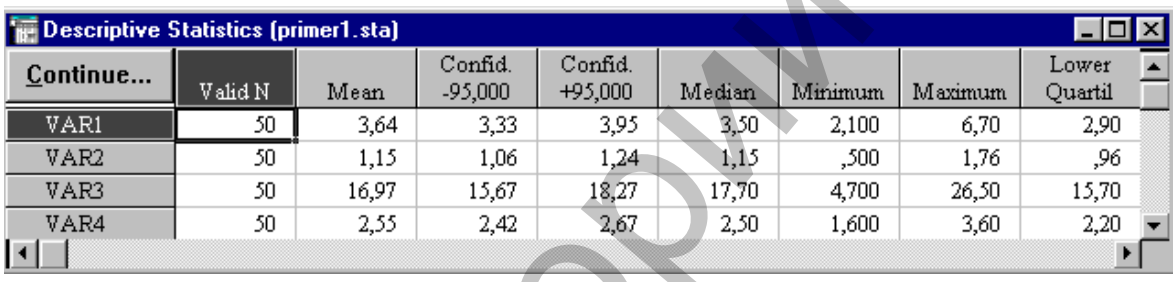
Kurtosis – эксцесс.

Эксцесс характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутости кривой. В кривой нормального распределения эксцесс равен нулю. Если эксцесс больше нуля, то кривая распределения характеризуется островершинностью, т.е. является более крутой по сравнению с нормальной, а случаи более густо группируются вокруг среднего. При отрицательном эксцессе кривая является более плосковершинной, т.е. более пологой по сравнению с нормальным распределением. Отрицательным пределом величины эксцесса является число -2, положительного предела – нет.

Standard error of kurtosis **–** стандартная ошибка эксцесса.

Напротив статистик, подлежащих вычислению (рис. 7), следует поставить флажок.

После нажатия на кнопку OK окна Descriptive statistics на экране появится таблица с результатами расчетов описательных статистик (рис. 8).



Descriptive Statistics (primer1.sta)

Continue...	Valid N	Mean	Confid. -95,000	Confid. +95,000	Median	Minimum	Maximum	Lower Quartil
VAR1	50	3,64	3,33	3,95	3,50	2,100	6,70	2,90
VAR2	50	1,15	1,06	1,24	1,15	,500	1,76	,96
VAR3	50	16,97	15,67	18,27	17,70	4,700	26,50	15,70
VAR4	50	2,55	2,42	2,67	2,50	1,600	3,60	2,20

Рис. 8. **Окно с результатами расчета описательных статистик.**

В таблице 2 эти данные представлены после копирования в текстовый редактор Word.

К сожалению, пакет Statistica не рассчитывает такие часто применяемые статистики, как коэффициент вариации и относительная ошибка среднего значения (точность опыта). Но их определение не представляет большого труда. Коэффициент вариации (%) есть отношение стандартного отклонения к среднему значению, умноженное на 100%:

$$\text{Коэффциентвариации} = \frac{\text{StandardDeviation}}{\text{Mean}} \cdot 100\%.$$

Коэффициент вариации, как дисперсия и стандартное отклонение, является показателем изменчивости признака. Коэффициент вариации не зависит от единиц измерения, поэтому удобен для сравнительной оценки различных статистических совокупностей. При величине коэффициента вариации до 10% изменчивость оценивается как слабая, 11–25% – средняя, более 25% – сильная (Лакин, 1990).

Относительная ошибка среднего значения (%) – отношение стандартной ошибки среднего к среднему значению, умноженное на 100% (для вероятности 0,68).

Этот процент расхождения между генеральной и выборочной средней показывает на сколько процентов можно ошибиться, если утверждать, что генеральная средняя равна выборочной средней. Если относительная ошибка не превышает 5%, то точность исследований (точность опыта) оценивается как хорошая, до 10% – удовлетворительная.

Точность 3–5% при вероятности 0,95, а в некоторых случаях и при вероятности 0,68, является вполне достаточной для большинства задач лесного хозяйства.

При необходимости обработки сгруппированных данных нужно воспользоваться кнопкой **Weight** окна Descriptive statistics (рис. 3). В появляющемся диалоговом окне (рис. 9) следует указать переменную, являющуюся весами для других переменных (Weight variables), а переключатель Status установить в положение ON. Необходимо иметь в виду, что весы действуют сразу для всех переменных. Поэтому обрабатывать сгруппированные и не сгруппированные данные нужно отдельно.

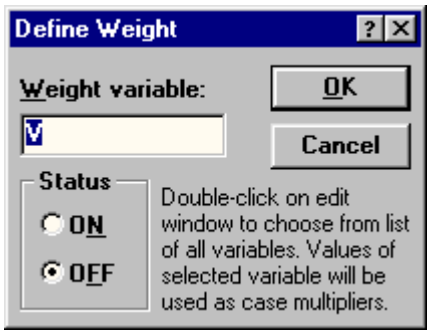


Рис. 9. **Окно задания переменной весов.**

При помощи опции **Alpha error** (рис. 3) выбирается уровень доверительной вероятности статистического анализа. В биологических исследованиях наиболее часто используется вероятность 0,95 (95%). Вероятности 0,95 соответствует уровень значимости 0,05 (5%).

Кнопка **Select cases** позволяет установить условия включения (include if) или исключения (exclude if) случаев (строк файла данных) из статистической обработки (рис. 10). Операторы, которые могут использоваться при написании выражений, а также примеры самих выражений имеются непосредственно на самом диалоговом окне Case Selection Conditions (рис. 10) в нижней его части.

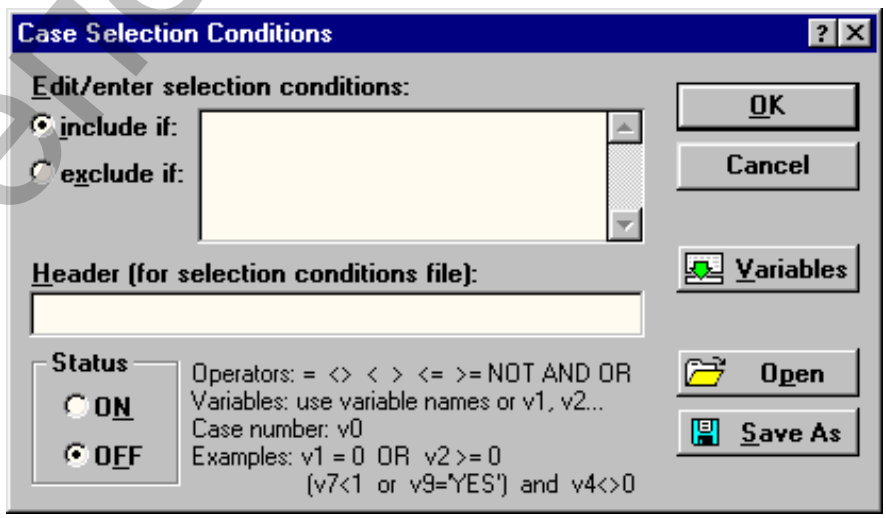


Рис. 10. **Окно задания условий выбора случаев.**

Таблица 2

Основные описательные статистики выборки 1-летних сеянцев сосны обыкновенной

Переменная			Confid.	Confid.				Lower	Upper
	Valid N	Mean	-95%	+95%	Median	Minimum	Maximum	Quartile	Quartile
VAR1	50	3,64	3,33	3,95	3,50	2,1	6,70	2,90	4,00
VAR2	50	1,15	1,06	1,24	1,15	0,5	1,76	0,96	1,37
VAR3	50	16,97	15,67	18,27	17,70	4,7	26,50	15,70	19,70
VAR4	50	2,55	2,42	2,67	2,50	1,6	3,60	2,20	2,80

Переменная		Quartile			Standard		Std.Err.		Std.Err.
	Range	Range	Variance	Std.Dev.	Error	Skewness	Skewness	Kurtosis	Kurtosis
VAR1	4,60	1,10	1,169	1,081	0,153	0,921	0,337	0,403	0,662
VAR2	1,26	0,41	0,098	0,313	0,044	-0,080	0,337	-0,451	0,662
VAR3	21,80	4,00	20,865	4,568	0,646	-0,834	0,337	0,772	0,662
VAR4	2,00	0,60	0,200	0,447	0,063	0,386	0,337	0,036	0,662

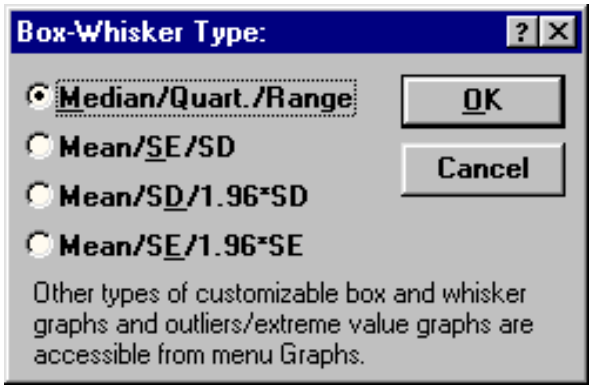


Рис. 11. **Окно выбора статистик для графика коробок.**

Для визуализации описательных статистик можно построить статистические графики типа «коробок» (или «ящиков с усами»). Это легко можно сделать при помощи кнопки **Box & Whisker plot for all variable** окна Descriptive statistics. На графике можно отобразить 3 статистики, установив переключатель в одно из 4-х положений (рис. 11):

1. ***Median/Quart./Range*** – Медиана / Квартили / Размах;
2. ***Mean/SE/SD*** – Среднее / Ошибка среднего / Стандартное отклонение;
3. ***Mean/SD/1.96SD*** – Среднее / Стандартное отклонение / Интервал 1,96* стандартного отклонения;
4. ***Mean/SE/1.96*SE*** – Среднее / Ошибка среднего / Интервал 1,96* ошибки среднего.

Визуализация описательных статистик переменных VAR1, VAR3 и VAR4 рассматриваемого примера при помощи графика коробок представлена на рис. 12.

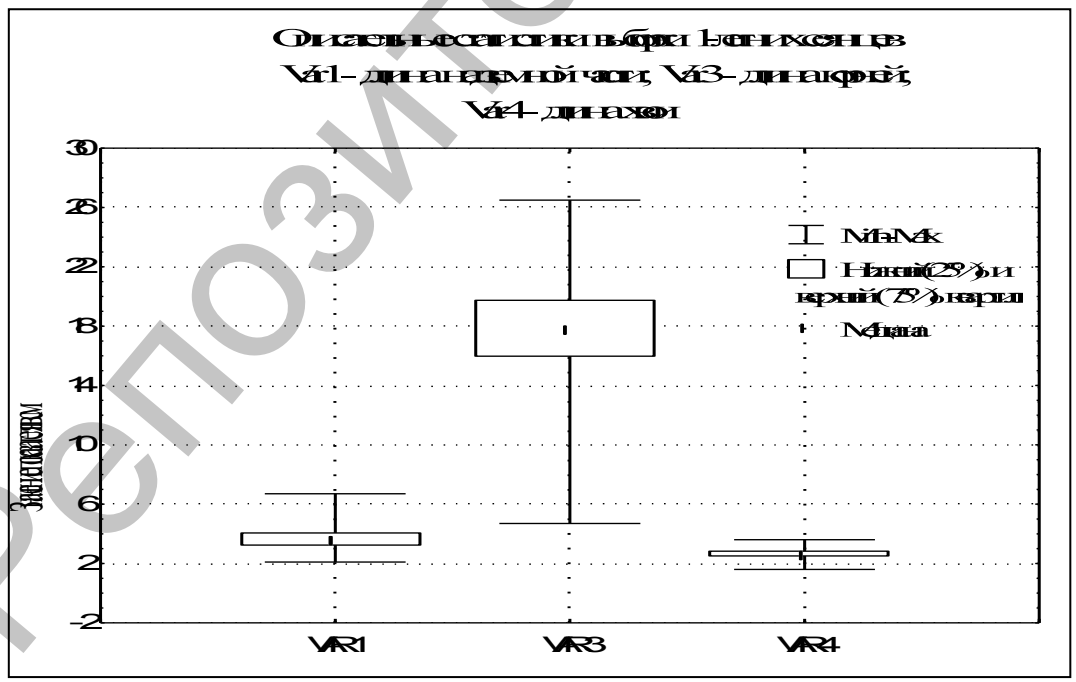


Рис. 12. **Описательные статистики в графическом виде.**

2.2. Процедура Correlation matrices (Корреляционные матрицы)

Эта процедура предназначена для проведения корреляционного анализа, установления тесноты линейной связи между переменными.

Установим тесноту взаимосвязей между таксационными показателями дубовых древостоев. Фрагмент окна файла данных представлен на рис. 13. Данные представляют собой таксационные показатели древостоев 93 пробных площадей, заложенных в низкоствольных дубравах 4 класса бонитета. По названию переменных понятно, какие таксационные показатели они содержат.

Data: DUB4.STA 16v * 93c

NUM VAL	Таксационные показатели древостоев порослевого дуба 4 кла						
	2 %ДУБА	3 N_ДУБА	4 G_ДУБА	5 D_ДУБА	6 H_ДУБА	7 A	8 M_ДУБА
1	50	1710	4,47	5,7	4,8	14	12,68
2	98	813	18,68	17,1	13,7	56	133,00
3	88	797	21,46	18,3	15,4	68	149,30
4	97	1258	26,51	16,4	14,1	56	190,13
5	36	311	7,25	17,2	14,8	57	52,80
6	85	1062	18,90	15,1	14,0	47	127,30
7	95	1150	21,26	15,3	11,2	38	131,40
8	92	873	20,56	17,3	13,8	55	139,40
9	61	1620	3,48	5,2	5,7	16	15,00
10	56	1790	5,27	6,1	5,2	14	16,10
11	54	1210	7,80	9,0	7,5	19	26,60
12	69	300	9,06	19,6	14,4	67	65,65
13	100	1145	21,72	15,6	12,8	54	154,50
14	48	278	10,85	22,3	15,4	61	90,00
15	77	1105	12,56	12,0	10,0	38	52,70

Рис. 13. **Окно файла данных.**

В стартовом окне этой процедуры **«Pearson Product-Moment Correlation»** (Корреляция Пирсона) (рис. 14) для расчета квадратной матрицы используется кнопка **One variable list (square matrix).**

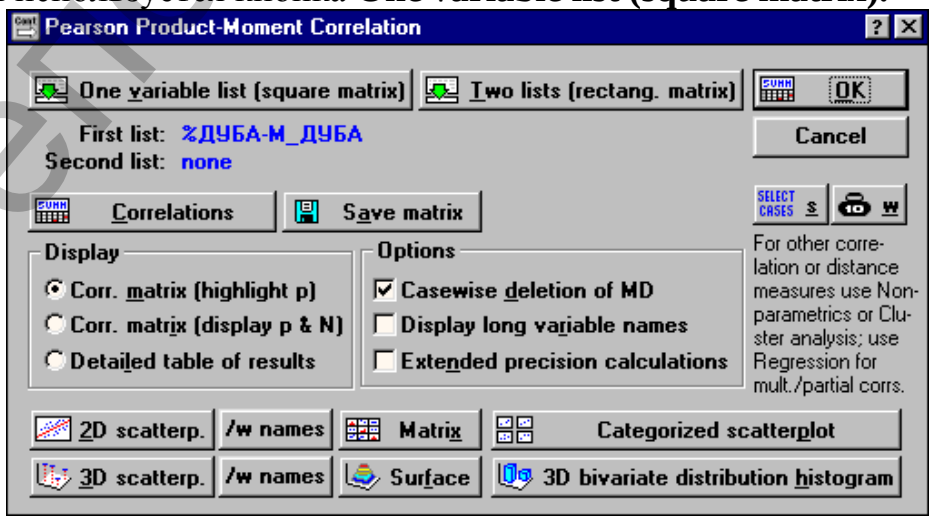
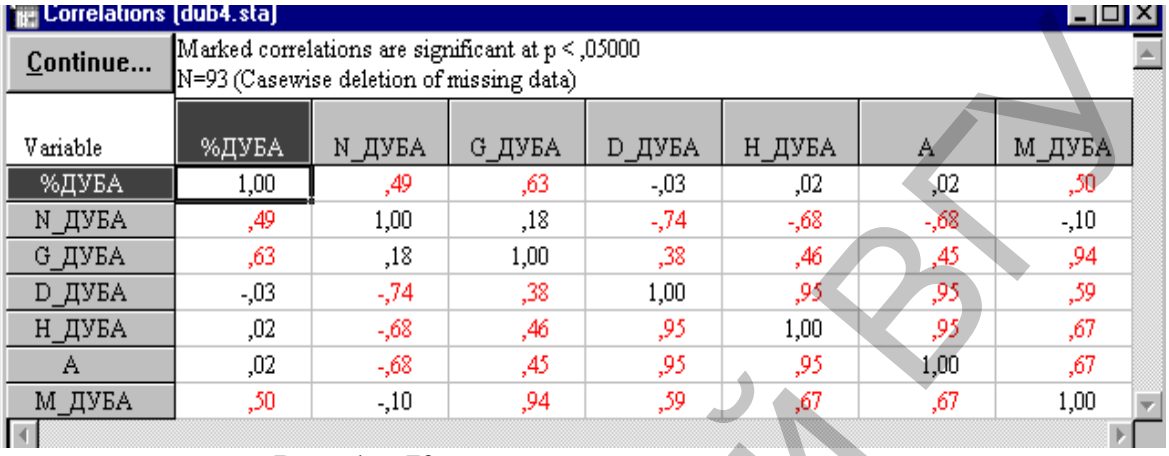


Рис. 14. **Окно Pearson Product-Moment Correlation.**

В списке переменных выбирают переменные, между которыми будут рассчитаны парные коэффициенты корреляции Пирсона. После нажатия на кнопку OK или Correlations на экране появится корреляционная матрица (рис. 15).


Correlations (dub4.sta)

Continue...

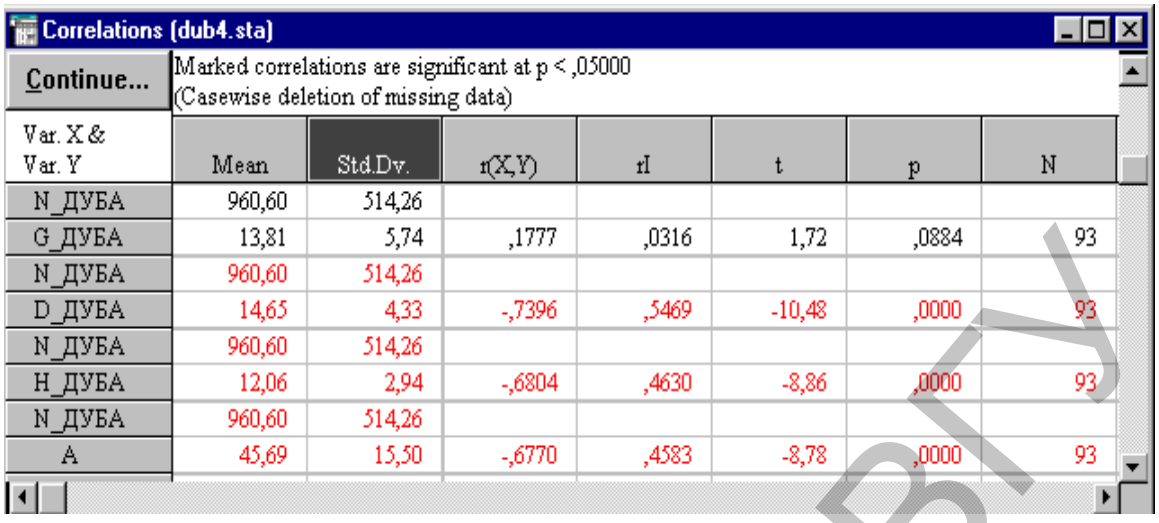
Marked correlations are significant at p < ,05000
N=93 (Casewise deletion of missing data)

Variable	%ДУБА	N_ДУБА	G_ДУБА	D_ДУБА	H_ДУБА	A	M_ДУБА
%ДУБА	1,00	,49	,63	-,03	,02	,02	,50
N_ДУБА	,49	1,00	,18	-,74	-,68	-,68	-,10
G_ДУБА	,63	,18	1,00	,38	,46	,45	,94
D_ДУБА	-,03	-,74	,38	1,00	,95	,95	,59
H_ДУБА	,02	-,68	,46	,95	1,00	,95	,67
A	,02	-,68	,45	,95	,95	1,00	,67
M_ДУБА	,50	-,10	,94	,59	,67	,67	1,00

Рис. 15. **Корреляционная матрица.**

Коэффициент корреляции – это показатель, оценивающий тесноту линейной связи между признаками. Он может принимать значения от -1 до +1. Знак «–» означает, что связь обратная, «+» – прямая. Чем ближе коэффициент к +1, тем теснее линейная связь. При величине коэффициента корреляции (по Дворецкому) менее 0,3 связь оценивается как слабая, от 0,31 до 0,5 – умеренная, от 0,51 до 0,7 – значительная, от 0,71 до 0,9 – тесная, 0,91 и выше – очень тесная. Для практических целей Дворецкий рекомендует использовать значительные, тесные и очень тесные связи.

Процедура Correlation matrices сразу же дает возможность проверить достоверность рассчитанных коэффициентов корреляции. Значение коэффициента корреляции может быть высоким, но не достоверным, случайным. Чтобы увидеть вероятность нулевой гипотезы (p), гласящей о том, что коэффициент корреляции равен 0, нужно в опции **Display** окна Pearson Product-Moment Correlation (рис. 14) установить переключатель на вторую строку **Corr. matrix (display p & N)**. Но даже если этого не делать и оставить переключатель в первом положении **Corr. matrix (highlight p),** статистически значимые на 5%-м уровне коэффициенты корреляции будут выделены в корреляционной матрице на экране монитора цветом, а при распечатке помечены звездочкой. Третье положение переключателя опции Display – **Detailed table of results** позволяет просмотреть результаты корреляционного анализа в деталях (рис. 16). Флажок опции **Casewise deletion of MD** устанавливается для исключения из обработки всей строки файла данных, в которой есть хотя бы одно пропущенное значение.



Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	rI	t	p	N
N_ДУБА	960,60	514,26					
G_ДУБА	13,81	5,74	,1777	,0316	1,72	,0884	93
N_ДУБА	960,60	514,26					
D_ДУБА	14,65	4,33	-,7396	,5469	-10,48	,0000	93
N_ДУБА	960,60	514,26					
H_ДУБА	12,06	2,94	-,6804	,4630	-8,86	,0000	93
N_ДУБА	960,60	514,26					
A	45,69	15,50	-,6770	,4583	-8,78	,0000	93

Рис. 16. **Вариант детального просмотра результатов корреляционного анализа.**

2.3. Процедура t-test for independent samples (t-критерий для независимых выборок)

Эта процедура используется для установления достоверной статистической разницы между средними значениями выборок на основе t-критерия Стьюдента.

Имеются результаты определения водопроницаемости почвы на площадках с различным характером напочвенного покрова (табл. 3). Создадим файл с данными с четырьмя переменными:

VAR1 –	Водопроницаемость на площадке 1 (мертвый покров, лесная подстилка 2.5см)
VAR2 –	Водопроницаемость на площадке 2 (травяной покров, проективное покрытие 40–50%, задернение 10%)
VAR3 –	Водопроницаемость на площадке 3 (травяной покров, проективное покрытие 100%, задернение 70%)
VAR4 –	Водопроницаемость на площадке 4 (травяной покров, проективное покрытие 30–40%, задернения нет)

Таблица 3

Значения переменных VAR1, VAR2, VAR3, VAR4
(Водопроницаемость почвы (мм/мин) в зависимости от характера напочвенного покрова)

Переменная			
VAR1	VAR2	VAR3	VAR4
1	2	3	4
303	78,7	53,5	67,9

238	82	68	105,3
303	58,1	38,8	149,3
238	97,1	49,5	138,9
303	73	70,4	45,5
200	142,9	40,5	98
400	55,6	25,1	61,3
238	108,7	12,2	75,8
263	69,9	33,6	71,4
303	120,5	28,3	35,7

Окно с файлом данных этого примера приводится на рис. 17.

Рис. 17. **Окно с файлом данных.**

Влияет ли характер напочвенного покрова на водопроницаемость почвы с ее поверхности? Воспользуемся процедурой t-test for independent samples для расчета средних величин водопроницаемости по вариантам опыта и одновременно проверим достоверность различий между средними значениями.

В окне «T-Test for independent Samples (Groups)» (рис. 18) в опции **Input file** следует указать тип файла с данными:

– **One record percase (use a grouping variable)** – одна запись на случай (используя группирующую переменную);

– **Each variable contains the data for one group** – каждая переменная содержит данные одной группы.

Использующийся нами файл данных (рис.19) относится ко второму типу (Each variable contains the data for one group).

При помощи кнопки **Variables** выбираются переменные для попарного сравнения. При этом должны быть выбраны переменные в

обоих списках. Чтобы сравнить попарно сразу все варианты опыта друг с другом, следует выбрать переменные так, как показано на рис. 19.

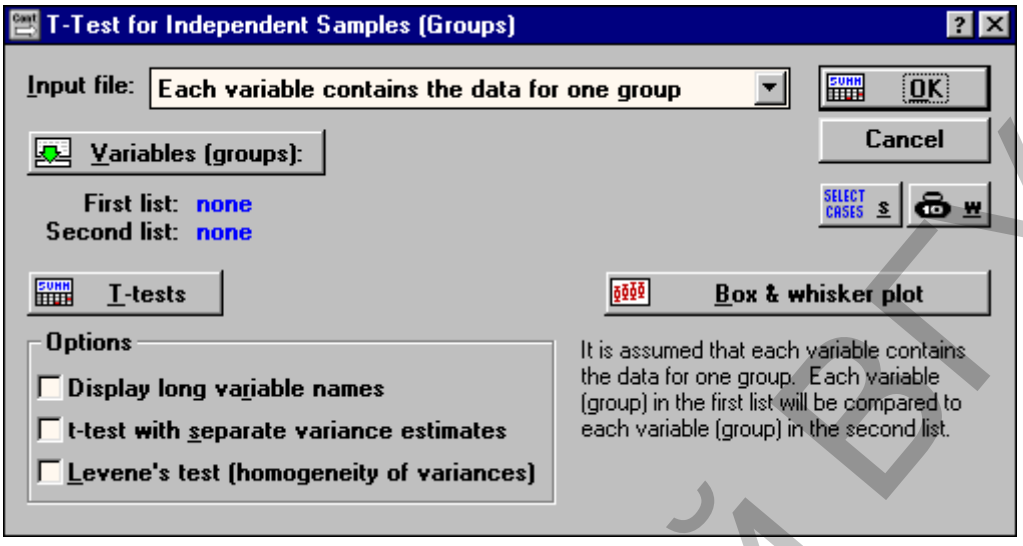


Рис. 18. **Окно «T-Test for independent samples (Groups)».**

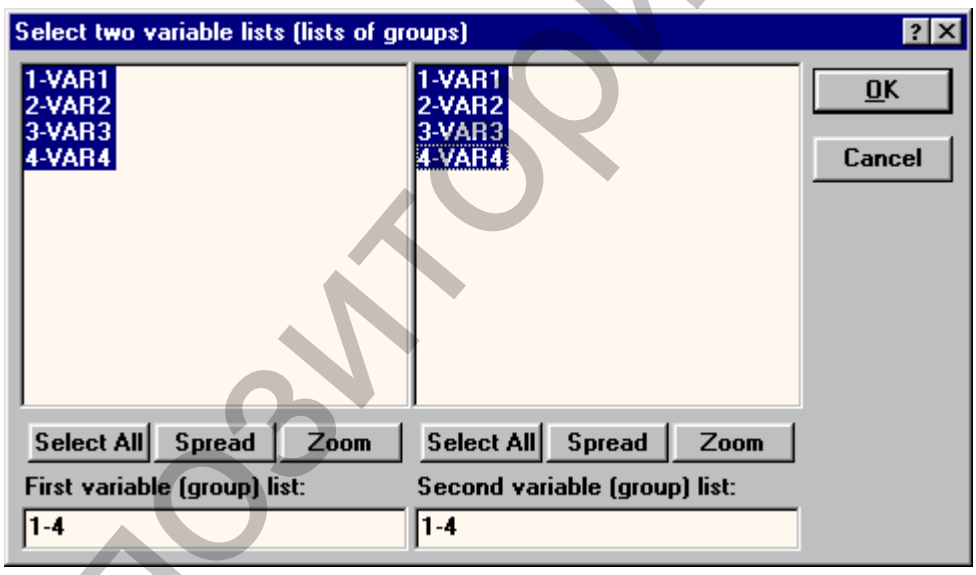
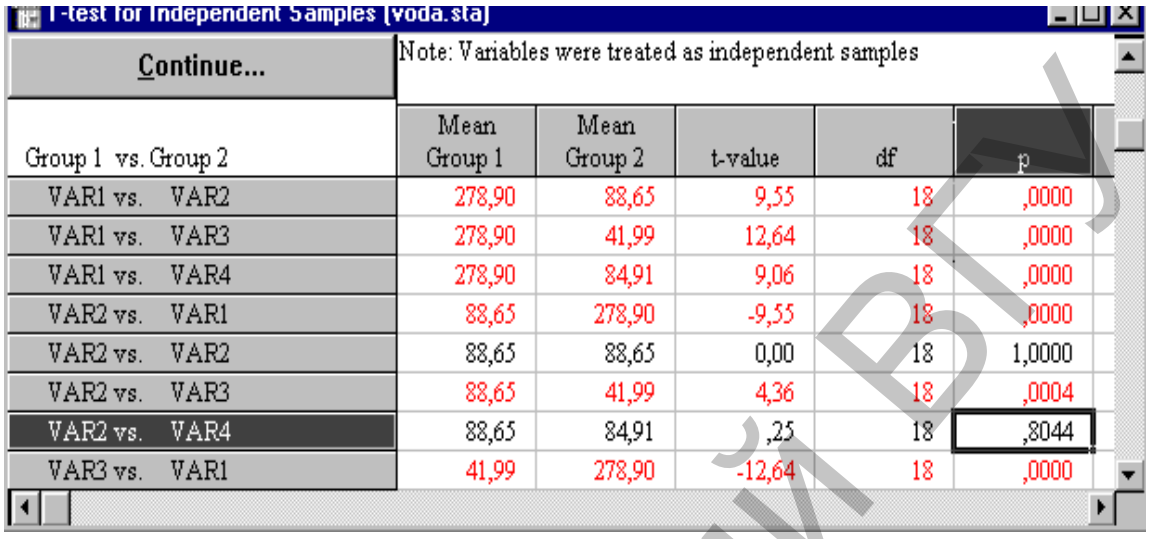


Рис. 19. **Выбор переменных для попарного сравнения.**

После нажатия на кнопку **OK** или **T-test** на экране появляется таблица с результатами сравнения по t-критерию.

Фрагмент окна с результатами проведения процедуры приводится на рис. 20. Согласно нулевой гипотезы между средними значениями водопроницаемости достоверного различия нет, т.е. две выборки однородны и представляют одну генеральную совокупность. Если вероятность нулевой гипотезы (p) меньше 5% (т.е. p < 0, 05), то с вероятность 0,95 нулевую гипотезу можно отбросить. По парное сравнение средних величин водопроницаемости показало достоверное

различие между всеми вариантами опыта, кроме вариантов 2 и 4. Нулевую гипотезу в последнем случае отбросить нельзя, так как ее вероятность чересчур высока (p=0,804).


T-test for Independent Samples (voda.sta)

Continue...

Note: Variables were treated as independent samples

Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p
VAR1 vs. VAR2	278,90	88,65	9,55	18	,0000
VAR1 vs. VAR3	278,90	41,99	12,64	18	,0000
VAR1 vs. VAR4	278,90	84,91	9,06	18	,0000
VAR2 vs. VAR1	88,65	278,90	-9,55	18	,0000
VAR2 vs. VAR2	88,65	88,65	0,00	18	1,0000
VAR2 vs. VAR3	88,65	41,99	4,36	18	,0004
VAR2 vs. VAR4	88,65	84,91	,25	18	,8044
VAR3 vs. VAR1	41,99	278,90	-12,64	18	,0000

Рис. 20. **Результаты проведения процедуры t-test for independent samples.**

1.4. Процедура Breakdown / one way ANOVA (Классификация и однофакторный дисперсионный анализ)

Эта процедура используется для проведения простейшего варианта однофакторного дисперсионного анализа данных по схеме полной рендомизации (неорганизованных повторений). Не позволяя вычленить дисперсию блоков (повторений), рядов, столбцов, процедура не предназначена для обработки данных, полученных по активным опытным схемам (рендомизированных блоков, смехе латинского квадрата, расщепленных делянок и блоков).

Воспользуемся исходными данными примера из раздела 2.3. и, проведя дисперсионный анализ, выясним, влияет ли характер напочвенного покрова на водопроницаемость почв с ее поверхности. Для проведения процедуры Breakdown / one way ANOVA следует создать файл с данными из двух переменных (табл. 3):

VAR1 –	Водопроницаемость почвы с поверхности (мм/мин) по всем вариантам опыта
VAR2 –	Номер варианта опыта (1, 2, 3 или 4)

Таблица 4

Значения переменных VAR1 и VAR2

VAR1	VAR2	VAR1	VAR2	VAR1	VAR2	VAR1	VAR2
303	1	78,7	2	53,5	3	67,9	4
238	1	82	2	68	3	105,3	4
303	1	58,1	2	38,8	3	149,3	4
238	1	97,1	2	49,5	3	138,9	4
303	1	73	2	70,4	3	45,5	4
200	1	142,9	2	40,5	3	98	4
400	1	55,6	2	25,1	3	61,3	4
238	1	108,7	2	12,2	3	75,8	4
263	1	69,9	2	33,6	3	71,4	4
303	1	120,5	2	28,3	3	35,7	4

На рис. 21 представлен вид окна с файлом данных.

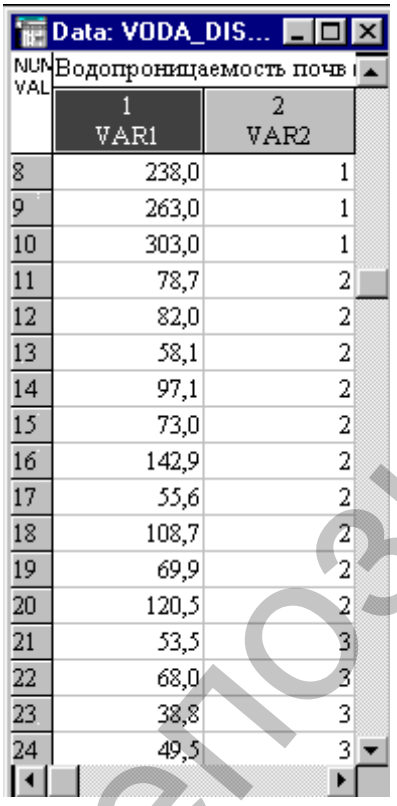


Рис. 21. **Вид окна файла данных.**

В окне Descriptive Statistics and Correlations by Groups (Breakdown) (рис. 22) в опции **Analysis** следует выбрать: **Detailed analysis of individual tables**. Вторая строка в списке Analysis – Bach process (and print) list of table предназначена для создания таблицы частот сгруппированных данных и разбитых на интервалы зависимых переменных. Флажок опции **Casewise (listwise) deletion of MD** устанавливается для исключения из обработки всей строки файла данных, в которой есть хотя бы одно пропущенное значение.

Через кнопку **Variables** выбирается зависимая переменная (Dependent variables) и группирующая переменная (Grouping variables), с помощью которой случаи будут разбиты на группы. Группирующей (Grouping) переменной в нашем примере является переменная VAR2, с ее помощью данные по водопроницаемости из зависимой (Dependent) переменной VAR1 группируются по четырем вариантам опыта.

После возвращения в диалоговое окно Descriptive Statistics and Correlations by Groups (Breakdown) и нажатия на кнопку **ОК** на экране появится окно Results (рис. 23) с результатами дисперсионного анализа. При помощи кнопок и опций этого окна в удобном виде можно просмотреть результаты обработки сгруппированных данных.

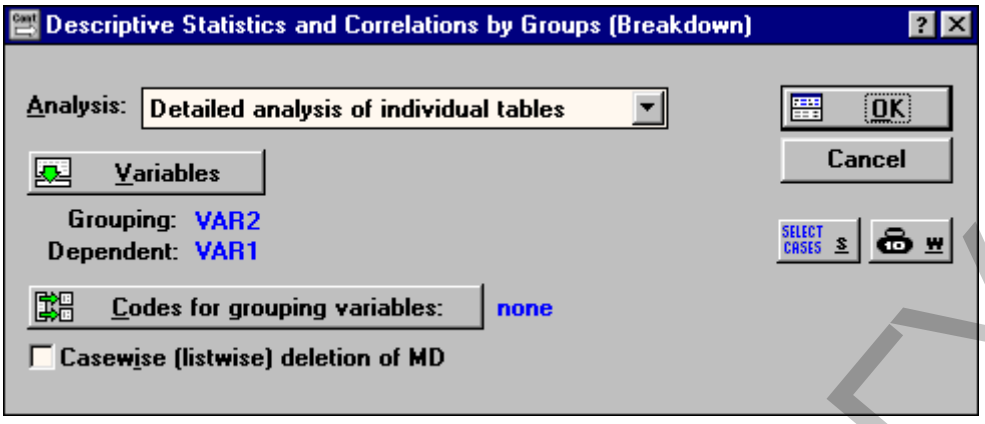


Рис. 22. **Окно «Descriptive Statistics and Correlations by Groups (Breakdown)».**

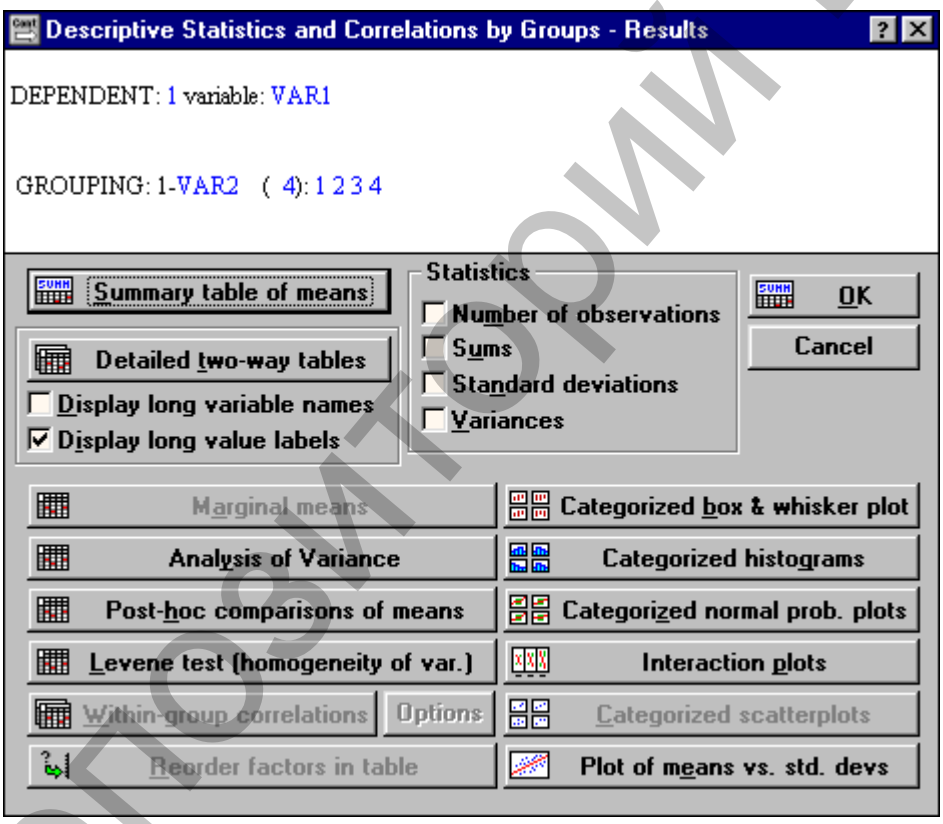


Рис. 23. **Окно Results процедуры Breakdown.**

Дисперсионный анализ заключается в разложении общей изменчивости признака на составные части: с одной стороны, на вариацию, определяемую действием изучаемого конкретного фактора, а с другой – вариацию, вызываемую случайными, неконтролируемыми в данном опыте факторами. Основные результаты дисперсионного анализа и проверку нулевой гипотезы однофакторного дисперсионного анализа (утверждающей, что фактор не влияет на вариацию зависимой

переменной, т.е. вся вариация сводится к случайной) можно просмотреть при помощи кнопки **Analysis of Variance** (Анализ дисперсий) окна Result (табл. 5).

Таблица 5

Результаты дисперсионного анализа

	SS	df	MS	SS	df	MS		p
	Effect	Effect	Effect	Error	Error	Error	F	
	(сумма квадратов фактора)	(число степеней свободы фактора)	(средний квадрат фактора)	(сумма квадратов ошибки)	(число степеней свободы ошибки)	(средний квадрат ошибки)		(вероятность нулевой гипотезы)
VAR1	334967	3	111655,67	51531,70	36	1431,436	78,00	0,0000000

Проверка нулевой гипотезы осуществляется при помощи F-критерия (Критерия Фишера). F-критерий используется как общий критерий, подтверждающий или опровергающий значимое влияние фактора на общую вариацию признака. В нашем примере низкая вероятность нулевой гипотезы (p=0,000000) позволяет ее отвергнуть и говорить о достоверном влиянии характера напочвенного покрова на водопроницаемость почвы.

Просмотрим средние значения водопроницаемости по вариантам опыта при помощи кнопки **Summary table of means** окна Results (табл. 6).

Таблица 6

Средние значения водопроницаемости по вариантам опыта

Вариант опыта	Водопроницаемость, мм/мин
1	278,90
2	88,65
3	41,99
4	84,91

Несмотря на то, что значимое влияние фактора доказано, это автоматически не означает, что каждый вариант опыта существенно отличается от всех других. Поэтому следующим важным этапом дисперсионного анализа является установление существенности частных различий, т.е. сравнение средних значений водопроницаемости по вариантам опыта. Для этого используется процедура **Post-hoc comparisons of means** (Post-hoc сравнения средних) (рис. 24). Сравнение групповых средних может производиться при помощи различных критериев:

LSD test or planned comparison – LSD – тест плановых сравнений. Этот критерий сравнения в отечественной литературе по статистике известен как наименьшая существенная разница (НСР).

Scheffii test – тест Шеффе.

Tukey (HSD) test – тест Тьюки. Тесты Шеффе и Тьюки считаются устаревшими (Литтл, Хиллз, 1981).

Duncan's multiple range test & critical ranges – Многоранговый критерий Дункана.

Выбор критериев осуществляется в диалоговом окне Post-hoc Comparisons of Means (рис. 24)

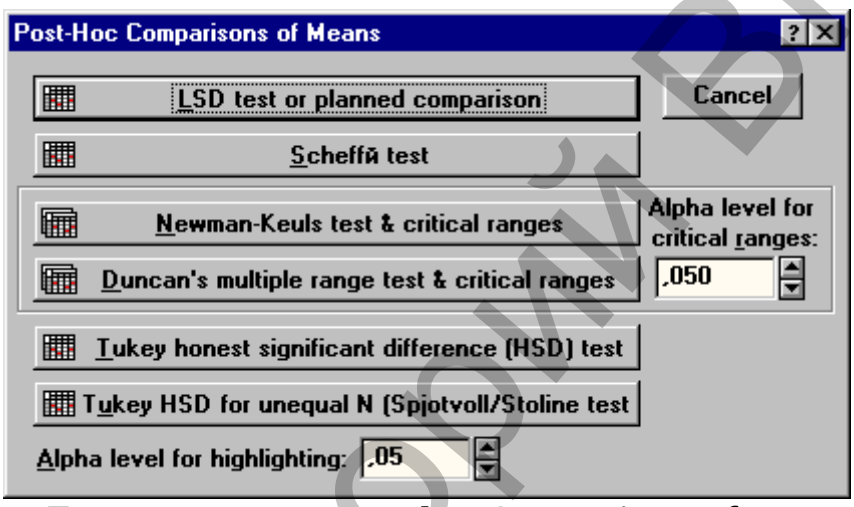


Рис. 24. **Диалоговое окно Post-hoc Comparisons of Means.**

Проведем сравнение средних значений водопроницаемости по вариантам опыта при помощи такого широко применяемого точечного критерия как НСР (LSD test or planned comparison) (рис. 25).

Анализируя результаты теста, представляющие собой вероятность нулевой гипотезы попарного сравнения средних величин водопроницаемости, мы видим достоверное различие на 5%-м уровне между всеми вариантами опыта, кроме вариантов 2 и 4. Нулевую гипотезу в последнем случае отбросить нельзя, так как ее вероятность высока (p=0,826).

LSD Test; Variable: VAR1 (voda_dis.sta)

Continue...

Marked differences are significant at p < ,05000

VAR2	{1} M=278,90	{2} M=88,650	{3} M=41,990	{4} M=84,910
G_1:1 {1}		,000000	,000000	,000000
G_2:2 {2}	,000000		,009087	,826310
G_3:3 {3}	,000000	,009087		,015671
G_4:4 {4}	,000000	,826310	,015671	

Рис. 25. **Результаты сравнения групповых средних по НСР.**

Сама величина HCP на экран не выводится, но если она потребуется, то она может быть легко рассчитана:

$$HCP_{0,05} = t_{0,05}\sqrt{\frac{2 \cdot MSError}{n}}$$

где: $t_{0,05}$ – величина t-критерия для 5%-ного уровня значимости (определяемся для числа степеней свободы, равному df Error); MS Error – средний квадрат ошибки; n – повторность опыта.

В нашем примере: $HCP_{0,05} = 2{,}04\sqrt{\frac{2 \cdot 1431{,}36}{10}} = 34{,}5$.

Лабораторная работа № 15
ПРОВЕДЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА ПРИ ПОМОЩИ МОДУЛЯ MULTIPLE REGRESSIONS

В стартовом диалоговом окне этого модуля (рис. 26.) при помощи кнопки **Variables** указываются зависимая (dependent) и независимые (ая) (independent) переменные. В поле **Input file** указывается тип файла с данными:

Raw Date – данные в виде строчной таблицы;

Correlation Matrix – данные в виде корреляционной матрицы.

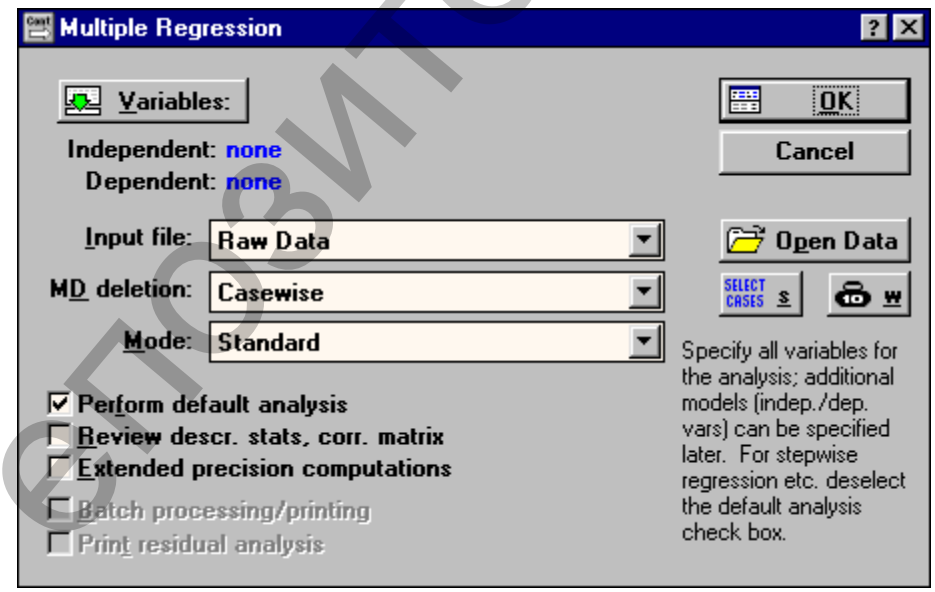


Рис. 26. **Стартовое диалоговое окно модуля Multiple Regression.**

В поле **MD deletion** указывается способ исключения из обработки недостающих данных:

casewise – игнорируется вся строка, в которой есть хотя бы одно пропущенное значение;

mean Substitution – взамен пропущенных данных подставляются средние значения переменных;

pairwise – попарное исключение данных с пропусками из тех переменных, корреляция которых вычисляется.

В поле **Mode** указывается тип регрессионной модели:

Standard – стандартная линейная модель вида:

$Y = a_1 + a_2X_1 + a_3X_2 + a_3X_3 + \ldots\ldots + a_nX_n$

Fixed non linear – фиксированная нелинейная, т.е. нелинейная модель, но которая может быть приведена к линейному виду путем преобразования переменных.

Рассмотрим проведение регрессионного анализа на примере. Имеются данные обмера и таксации 380 модельных деревьев различных древесных пород. В файле данных (рис. 29) 10 переменных:

1	PORODA	Древесня порода (d- дуб, lp- липа, k- клен, o – осина)
2	A	Возраст дерева, лет
3	D	Таксационный диаметр ствола дерева в коре, см
4	H	Высота дерева, м
5	VK	Объем ствола в коре, куб.м
6	V	Объем ствола без коры, куб.м
7	Q2	Второй коэффициент формы
8	L	Длина кроны дерева, м
9	DKR	Диаметр кроны дерева, м
10	F	Старое видовое число

Data: MODL.STA 10v * 380c

TEXT VALUE: Таксационные показатели модельных деревьев

	1 PORODA	2 A	3 D	4 H	5 VK	6 V	7 Q2	8 L	9 DKR	10 F
196	d	21	6,8	9,8	,0170	,0141	,69	6,00	1,10	,478
197	d	37	8,5	10,0	,0272	,0203	,68	7,00	3,10	,479
198	d	35	10,2	12,8	,0556	,0506	,73	10,50	1,60	,532
199	d	36	13,3	14,0	,1018	,0710	,71	7,20	2,10	,523
200	d	42	15,3	15,0	,1375	,1122	,72	7,00	4,60	,499
201	d	46	18,0	15,0	,1748	,1402	,69	9,50	3,70	,458
202	d	44	18,9	17,0	,2148	,1807	,66	13,00	5,60	,450
203	d	41	19,7	16,5	,2375	,2007	,69	10,00	6,30	,472
204	d	45	23,5	16,5	,3216	,2636	,66	7,50	4,00	,449
205	k	30	8,5	10,4	,0287	,0243	,71	8,40	2,65	,486
206	k	53	10,6	13,7	,0696	,0680	,82	4,50	3,65	,576
207	k	9	3,7	5,9	,0033	,0029	,67	4,40	2,00	,520
208	k	38	12,6	13,0	,0746	,0630	,69	5,40	3,50	,460

Рис. 29. **Вид окна с файлом данных.**

Найдем параметры регрессионного уравнения линейной связи объема ствола дуба в коре (переменная VK) от диаметра (D) и высоты (H) ствола. Вид уравнения: $VK = a_1 + a_2D + a_3H$.

Выставим опции стартового окна регрессионного анализа (рис. 29):

Variables: зависимая (dependent) переменная – VK; независимые (independent) – D, H (рис. 30); **Input file – Raw Date** (данные файла в виде строчной таблицы); **MD deletion – pairwise; Mode – Standard.**

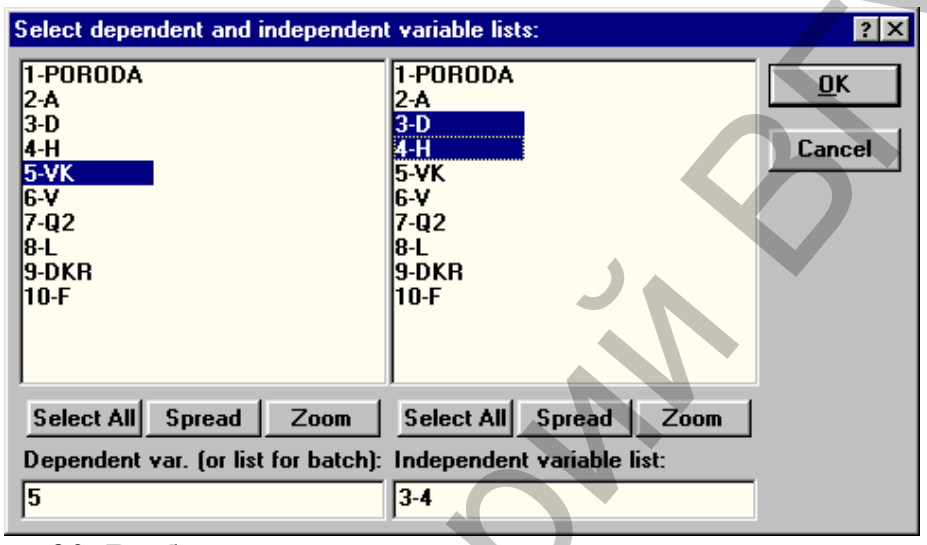


Рис. 30. **Выбор зависимой и независимых переменных.**

Так как в файле данных содержится информация о модельных деревьях разных пород, а уравнение регрессии мы хотим получить для дуба, нужно воспользоваться кнопкой Select cases диалогового окна Multiple Regressions, чтобы установить условие включения случаев (строк файла данных) в статистическую обработку. В обработку должны включаться только те строки файла данных, для которых значение первой переменной V1 = 'd' (т.е. дуб) (рис. 31).

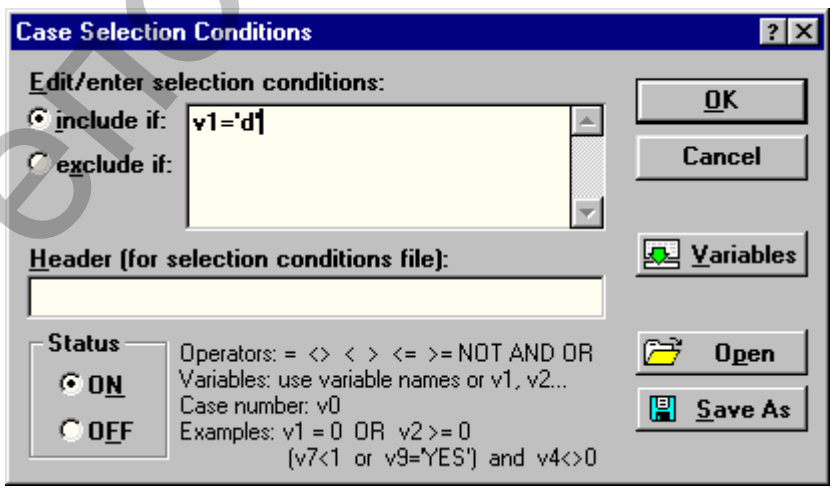


Рис. 31. **Задание условия включения в обработку случаев со значением переменной V1 – дуб.**

После того, как все опции стартового диалогового окна регрессионного анализа выставлены, нажатие на кнопку ОК приведет к появлению окна Multiple Regression Results (результаты регрессионного анализа) (рис. 32), с помощью которого можно просмотреть результаты анализа в деталях.

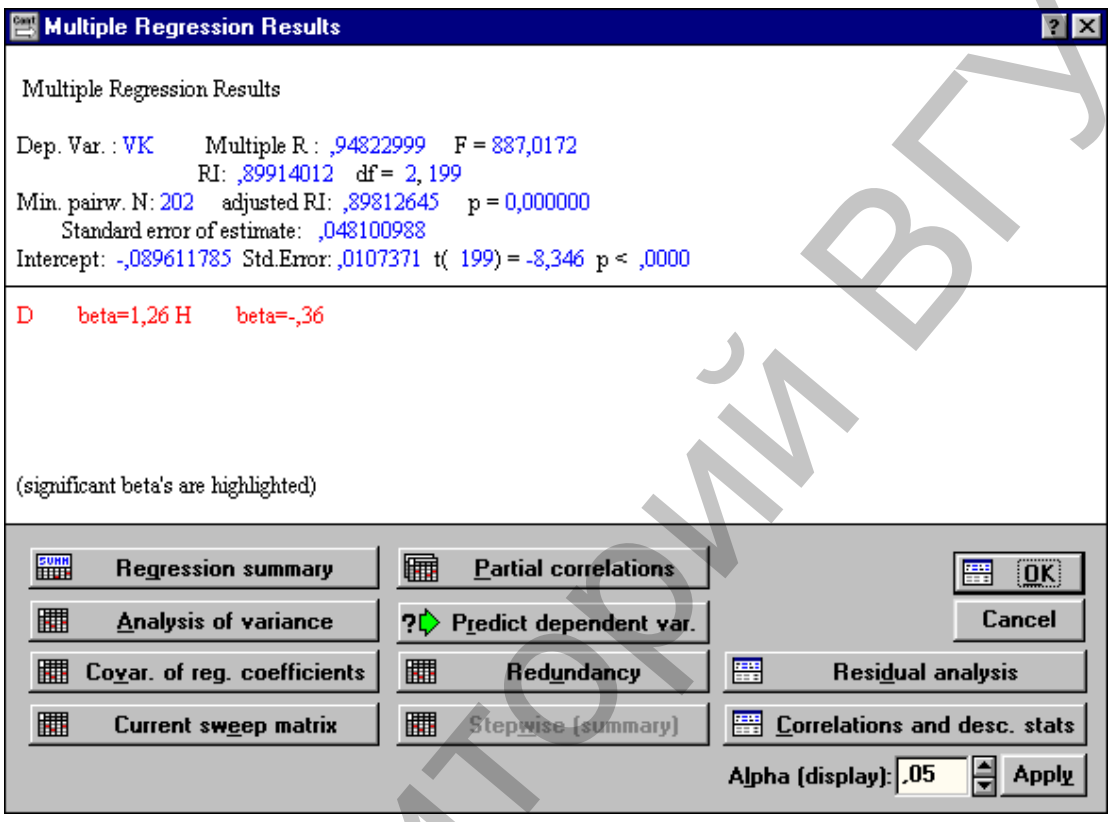


Рис. 32. **Окно просмотра результатов регрессионного анализа.**

В верхней части окна приводятся наиболее важные параметры полученной регрессионной модели:

Multiple R – коэффициент множественной корреляции.

Характеризует тесноту линейной связи между зависимой и всеми независимыми переменными. Может принимать значения от 0 до 1.

R^2 или **RI** – коэффициент детерминации.

Численно выражает долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения. Чем больше R^2, тем большую долю вариации объясняют переменные, включенные в модель;

adjusted R – скорректированный коэффициент множественной корреляции.

Этот коэффициент лишен недостатков коэффициента множественной корреляции. Включение новой переменной в регрессионное уравнение увеличивает RI не всегда, а только в том случае, когда ча-

стный F-критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение RI и adjusted R^2;

adjusted R^2 или **adjusted RI** – скорректированный коэффициент детерминации.

Скорректированный R^2 можно с большим успехом (по сравнению с R^2) применять для выбора наилучшего подмножества независимых переменных в регрессионном уравнении.

F – F-критерий;

df – число степеней свободы для F-критерия;

p – вероятность нулевой гипотезы для F-критерия;

Standard error of estimate – стандартная ошибка оценки (уравнения);

Intercept – свободный член уравнения;

Std.Error – стандартная ошибка свободного члена уравнения;

t – t-критерий для свободного члена уравнения;

p – вероятность нулевой гипотезы для свободного члена уравнения.

Beta – β-коэффициенты уравнения.

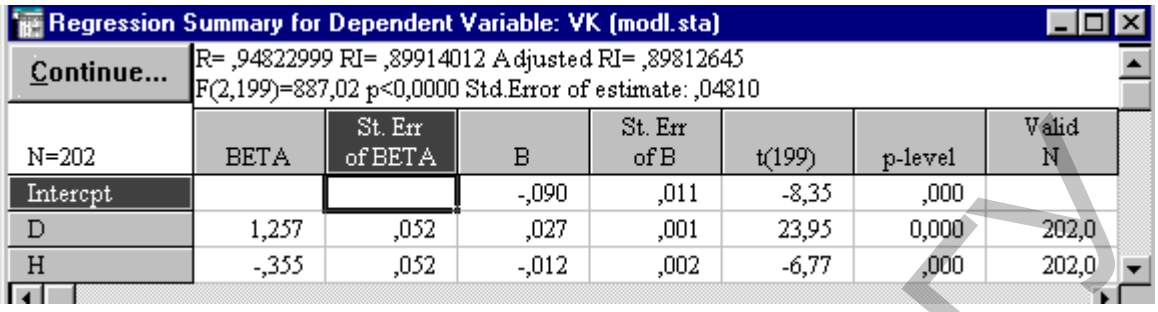
Это стандартизированные регрессионные коэффициенты, рассчитанные по стандартизированным значениям переменных. По их величине можно сравнить и оценить значимость зависимых переменных, так как β-коэффициент показывает, на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной при условии постоянства остальных независимых переменных. Свободный член в таком уравнении равен 0.

При помощи кнопок диалогового окна Multiple Regression Results (рис. 32) результаты регрессионного анализа можно просмотреть более детально.

Кнопка **Regression summary** – позволяет просмотреть основные результаты регрессионного анализа (рис. 33): **BETA** – β-коэффициенты уравнения; **St. Err. of BETA** – стандартные ошибки β-коэффициентов; **B** – коэффициенты уравнения регрессии; **St. Err. of B** – стандартные ошибки коэффициентов уравнения регрессии; **t (95)** – t-критерии для коэффициентов уравнения регрессии; **p-level** – вероятность нулевой гипотезы для коэффициентов уравнения регрессии.

Таким образом, в результате проведенного регрессионного анализа получено следующее уравнение взаимосвязи между объемом ствола дуба в коре (VK), диаметром (D) и высотой (H) ствола: VK = -0,090 + 0,027D – 0,012H. Все коэффициенты уравнения значимы на 5%-м уровне (p-level < 0,05). Это уравнение объясняет 89,9%

(R^2 = 0,899) вариации зависимой переменной. Ограничения модели: 2<=D>=31; 1,6<=H>=19,5.



Regression Summary for Dependent Variable: VK (modl.sta)

Continue...

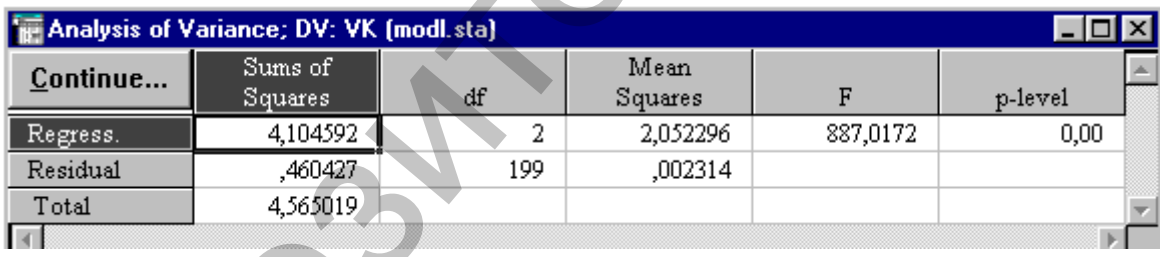
R= ,94822999 RI= ,89914012 Adjusted RI= ,89812645
F(2,199)=887,02 p<0,0000 Std.Error of estimate: ,04810

N=202	BETA	St. Err of BETA	B	St. Err of B	t(199)	p-level	Valid N
Intercpt			-,090	,011	-8,35	,000	
D	1,257	,052	,027	,001	23,95	0,000	202,0
H	-,355	,052	-,012	,002	-6,77	,000	202,0

Рис. 33. **Основные результаты регрессионного анализа.**

Кнопка **Analysis of variance** – позволяет ознакомиться с результатами дисперсионного анализа уравнения регрессии (рис. 34). В строках таблицы дисперсионного анализа уравнения регрессии – источники вариации: Regress. – обусловленная регрессией, Residual – остаточная, Total – общая. В столбцах таблицы: Sums of Squares – сумма квадратов, df – число степеней свободы, Mean Squares – средний квадрат, F – значение F-критерия, p-level – вероятность нулевой гипотезы для F-критерия.

F-критерий полученного уравнения регрессии значим на 5%-м уровне. Вероятность нулевой гипотезы (p-level) значительно меньше 0,05, что говорит об общей значимости уравнения регрессии.

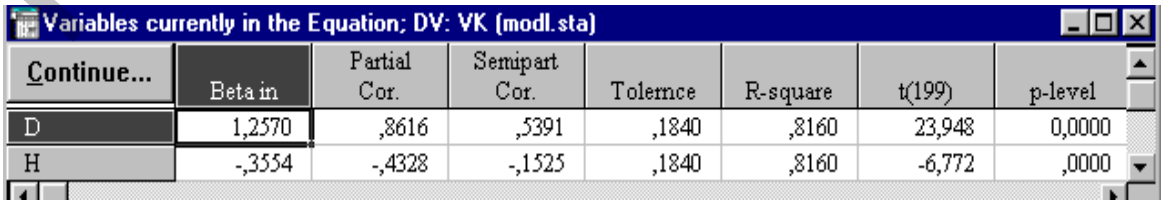


Analysis of Variance; DV: VK (modl.sta)

Continue...	Sums of Squares	df	Mean Squares	F	p-level
Regress.	4,104592	2	2,052296	887,0172	0,00
Residual	,460427	199	,002314		
Total	4,565019				

Рис. 34. **Результаты дисперсионного анализа уравнения регрессии.**

Кнопка **Partial correlations –** позволяет просмотреть частные коэффициенты корреляции (Partial Cor.) между переменными (рис. 35). Частная корреляция – это корреляция между двумя переменными, когда одна или больше из оставшихся переменных удерживаются на постоянном уровне (т.е. имеют постоянное значение). Частные коэффициенты корреляции, как и парные, могут принимать значения от -1 до +1.



Variables currently in the Equation; DV: VK (modl.sta)

Continue...	Beta in	Partial Cor.	Semipart Cor.	Tolernce	R-square	t(199)	p-level
D	1,2570	,8616	,5391	,1840	,8160	23,948	0,0000
H	-,3554	-,4328	-,1525	,1840	,8160	-6,772	,0000

Рис. 35. **Результаты расчета частных коэффициентов корреляции.**

Сильная взаимная коррелированность независимых переменных в нашем уравнении затрудняет анализ влияния отдельных факторов на зависимую переменную. Отрицательный знак коэффициента уравнения перед высотой (H), отрицательный знак частного коэффициента корреляции VK с H противоречат реальному положению дел. Положительный знак парного коэффициента корреляции между высотой и объемом ствола говорит о прямой взаимосвязи между ними.

В идеальной регрессионной модели независимые переменные вообще не коррелируют друг с другом. Однако в моделях, разрабатываемых для природных объектов, сильная коррелированность переменных является довольно частым явлением. Это приводит к увеличению ошибок уравнения, уменьшению точности оценивания, снижается эффективность использования регрессионной модели. Поэтому выбор независимых переменных, включаемых в регрессионную модель, должен быть очень тщательным.

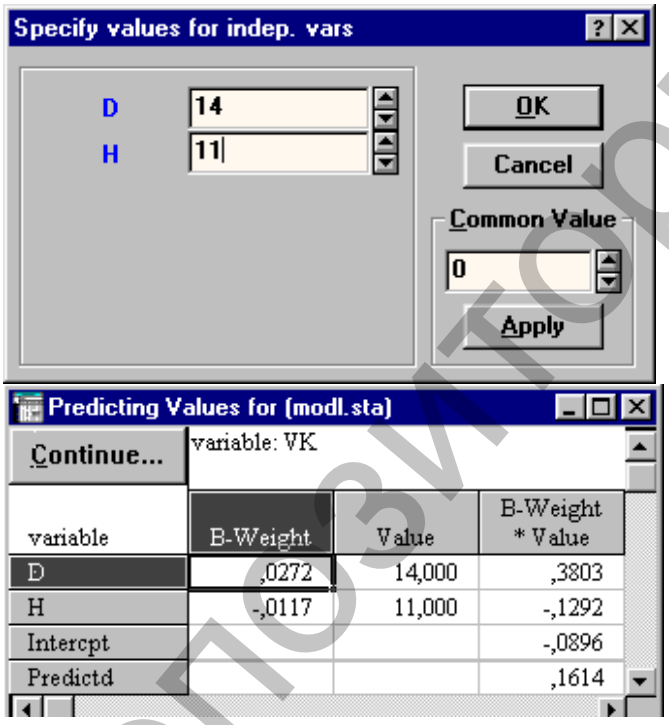


Рис. 36. **Окно задания значений независимых переменных и результаты расчета по регрессионному уравнению зависимой переменной.**

Кнопка **Predict dependent var.** позволяет рассчитать по полученному регрессионному уравнению значение зависимой переменной по значениям независимых переменных. На рис. 36 приводится пример расчета объема ствола дуба в коре при величине диаметра ствола – 14 см и высоты – 11 м. Предсказанный (Predictd) объем составил 0,1614 куб.м.

Кнопка **Correlations and desc. stats** позволяет просмотреть описательные статистики и корреляционную матрицу с парными коэффициентами корреляции переменных, участвующих в регрессионной модели (рис. 37).

Кнопка **Residual analysis** запускает процедуру всестороннего анализа остатков регрессионного уравнения (рис. 38). Остатки – это разности между опытными и предсказанными значениями зависимой переменной в построенной регрессионной модели.

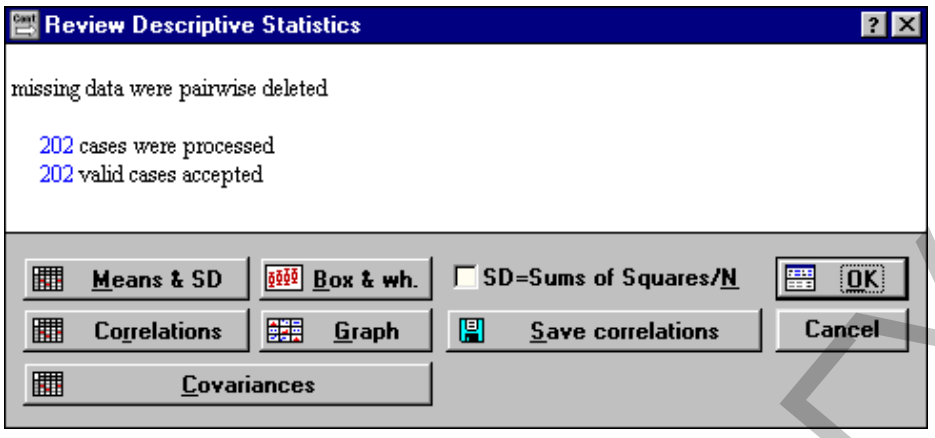


Рис. 37. **Диалоговое окно Review Descriptive Statistics.**

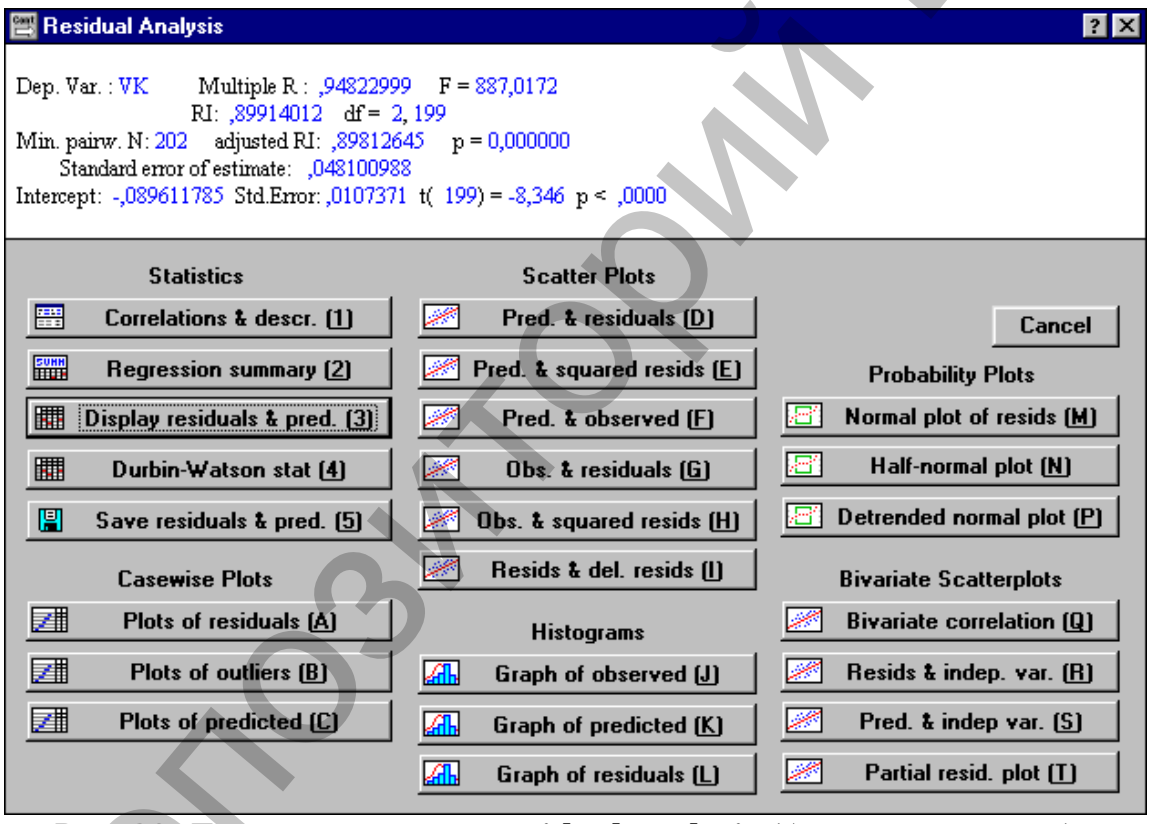


Рис. 38. **Диалоговое окно Residual analysis (Анализ остатков).**

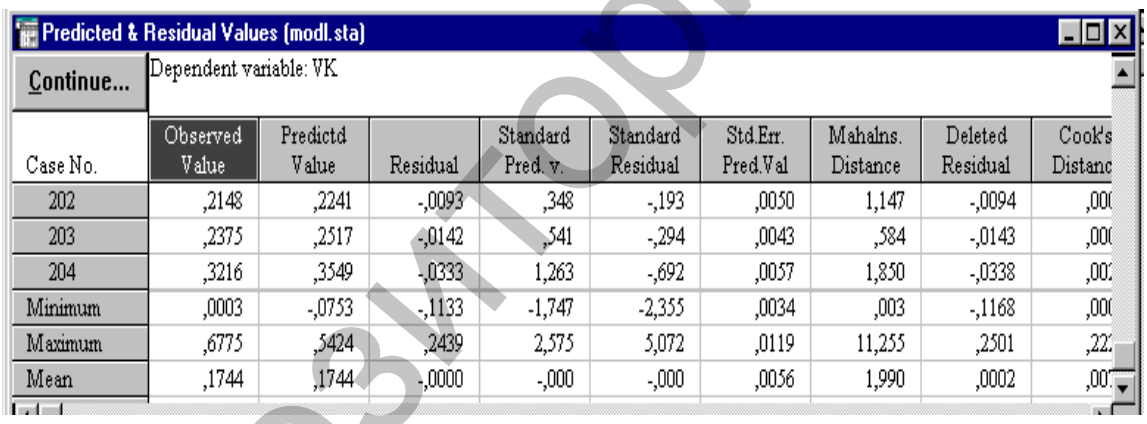
Кнопка **Redundancy** предназначена для поиска выбросов. Выбросы – это остатки, которые значительно превосходят по абсолютной величине остальные. Выбросы показывают опытные данные, которые являются не типичными по отношению к остальным данным, и требует выяснения причин их возникновения. Выбросы должны исключаться из обработки, если они вызваны ошибками регистрации, измерения. Для выделения имеющихся в регрессионных остатках выбросов предложен ряд показателей:

Показатель Кука (Cook's Distance) – принимает только положительное значение и показывает расстояние между коэффициентами уравнения регрессии после исключения из обработки i-й точки данных. Большое значение показателя Кука указывает на сильно влияющий случай.

Расстояние Махаланобиса (Mahalns. Distance) – показывает насколько каждый случай или точка в p-мерном пространстве независимых переменных отклоняется от центра статистической совокупности.

Внимательный анализ остатков позволяет оценить адекватность модели. Остатки должны быть нормально распределены, со средним значением равным нулю и постоянной, независимо от величин зависимой и независимой переменных, дисперсией. Модель должна быть адекватна на всех отрезках интервала изменения зависимой переменной.

Просмотр величин остатков и специальных критериев, их оценивающих, осуществляется при помощи кнопки **Display residuals & pred.** окна Residual analysis. Для нашего примера фрагмент окна с этими данными представлен на рис. 39.


Predicted & Residual Values (modl.sta)

Continue... Dependent variable: VK

Case No.	Observed Value	Predictd Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalns. Distance	Deleted Residual	Cook's Distanc
202	,2148	,2241	-,0093	,348	-,193	,0050	1,147	-,0094	,00
203	,2375	,2517	-,0142	,541	-,294	,0043	,584	-,0143	,00
204	,3216	,3549	-,0333	1,263	-,692	,0057	1,850	-,0338	,00
Minimum	,0003	-,0753	-,1133	-1,747	-2,355	,0034	,003	-,1168	,00
Maximum	,6775	,5424	,2439	2,575	5,072	,0119	11,255	,2501	,22
Mean	,1744	,1744	-,0000	-,000	-,000	,0056	1,990	,0002	,00

Рис. 39. **Окно со значениями остатков (Residuals), показателями Кука (Cook's Distance), расстояния Махаланобиса (Mahalns. Distance), опытными (Observed Value) и предстказанными по уравнению (Predictd Value) значениями зависимой переменной.**

Вполне достаточно бывает одного графического анализа остатков. О нормальности остатков можно судить по графику остатков на нормальной вероятностной бумаге. Чем ближе распределение к нормальному виду, тем лучше значения остатков ложатся на прямую линию. Он строится при помощи кнопки **Normal plot of resids** окна Residual analysis (рис. 40).

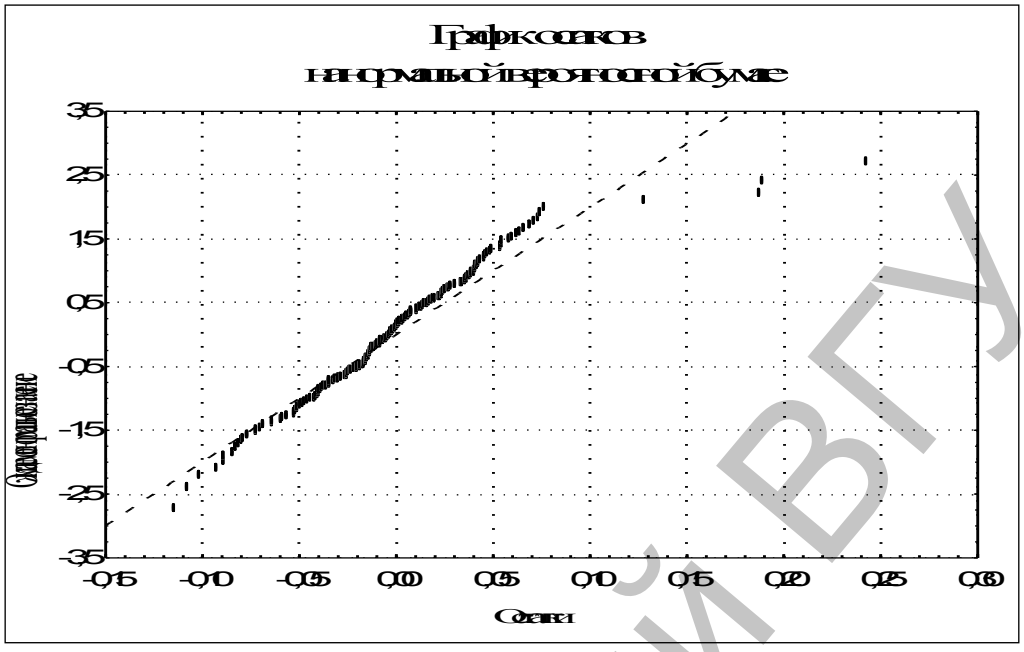


Рис. 40. **График остатков на нормальной вероятностной бумаге.**

Важно просмотреть графики зависимости остатков от каждой из независимых переменных. Их легко просмотреть при помощи кнопки **Resids & indep. var.** окна Residual analysis. Остатки должны быть нормально распределены, т.е. на графике они должны представлять приблизительно горизонтальную полосу одинаковой ширины на всем ее протяжении. Коэффициент корреляции (r) между регрессионными остатками и переменными должен равняться нулю.

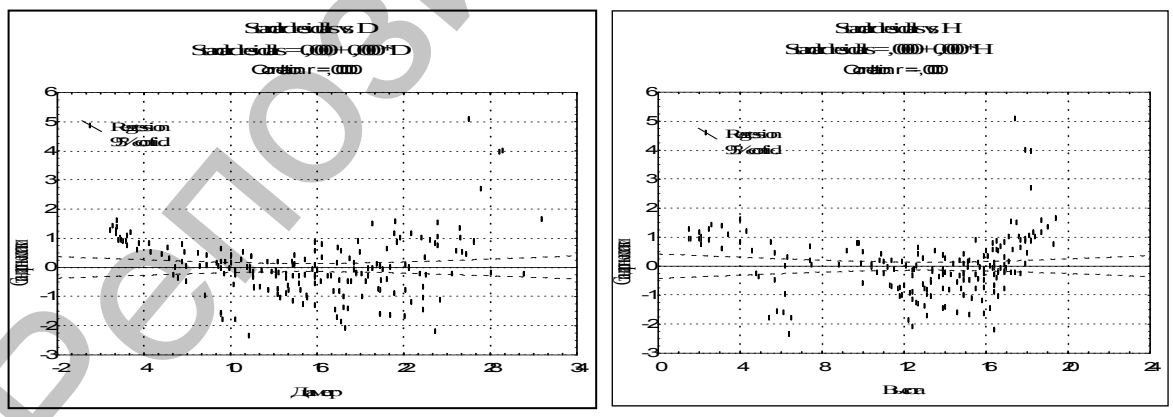


Рис. 41. **Зависимость остатков от независимых переменных: диаметра и высоты.**

В нашем случае на графиках остатков (рис. 42) хорошо просматривается нелинейный тренд, что вызывает сомнение в адекватности модели. Присутствие нелинейного тренда в регрессионных остатках гово-

рит о необходимости пересмотра модели (преобразования или ввода новых переменных, перехода от линейной модели к нелинейной).

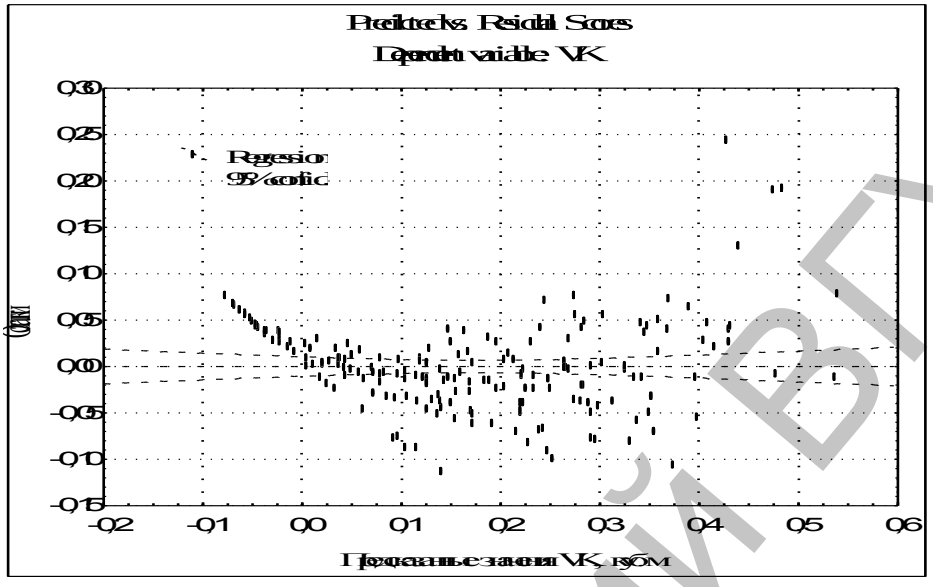


Рис. 42. **Зависимость регрессионных остатков от предсказанных значений зависимой переменной.**

Для выявления нестабильности дисперсии ошибки уравнения при помощи кнопки **Pred. & residuals** окна Residual analysis можно создать график зависимости регрессионных остатков от предсказанного значения зависимой переменной. Рис. 42 позволяет заключить о непостоянстве дисперсии ошибки уравнения (с увеличением значений зависимой переменной дисперсия увеличивается). Это еще одно подтверждение неадекватности анализируемой модели.

Очень удобным визуальным способом оценки адекватности регрессионной модели является анализ графического изображения опытных и полученных по регрессионному уравнению значений зависимой переменной. Оно строится при помощи кнопки **Pred. & observed** окна Residual analysis.

Из рис. 43 хорошо видно, что линейный вид нашей модели плохо описывает взаимосвязь объема ствола дуба в коре от его диаметра и высоты (модель при малых и больших значениях отклика занижает величину зависимой переменной). Эта связь носит нелинейный характер.

Рассмотрим порядок нахождения коэффициентов уравнений регрессии нелинейного вида, но которые через преобразования переменных могут быть приведены к линейной модели. Найдем параметры регрессионного уравнения связи объема ствола дуба в коре (переменная VK) от диаметра (D) ствола. Вид уравнения: $VK = a_1 + a_2D + a_3D^2$.

Опцию **Mode** стартового окна регрессионного анализа (рис. 26) выставим в положение **Fixed non linear**.

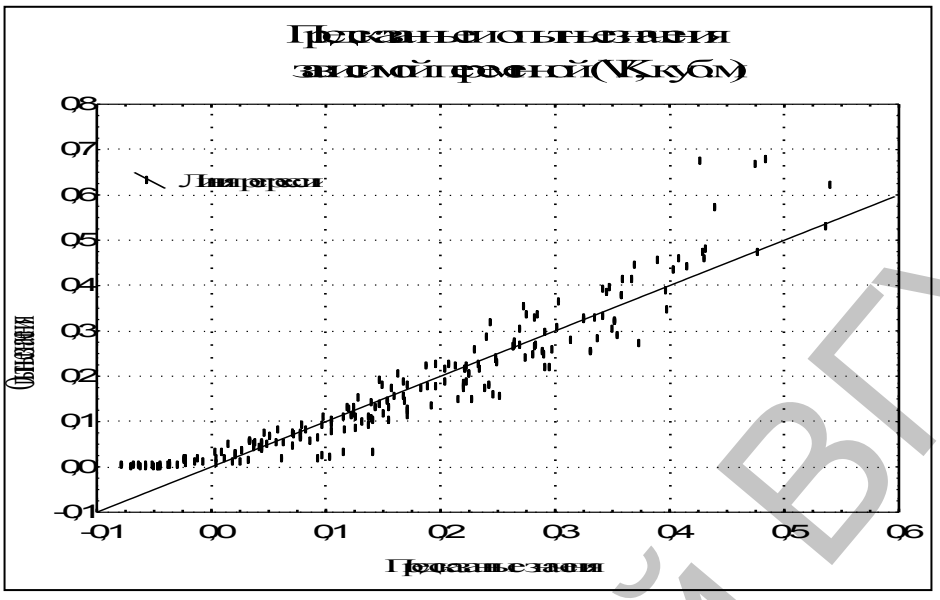


Рис. 43. **Линия регрессии, опытные и полученные по регрессионному уравнению значений зависимой переменной.**

Если выбран фиксированный нелинейный тип регрессионной модели, то после нажатия на кнопку OK в диалоговом окне Multiple Regressions (рис. 44), появляется окно Non-linear Components Regression (рис. 44), в котором можно выбрать следующие типы преобразования переменных: X^2, X^3, X^4, X^5, lnX (X >0), $lg_{10}X$ (X >0), e^X (40<X<-40), 10^X (-18 to +18). Если потребуются какие-либо иные преобразования переменных, то тогда в файле данных следует создать мнимые вичисляемые переменные и включить их в качестве зависимых переменных в регрессионную модель.

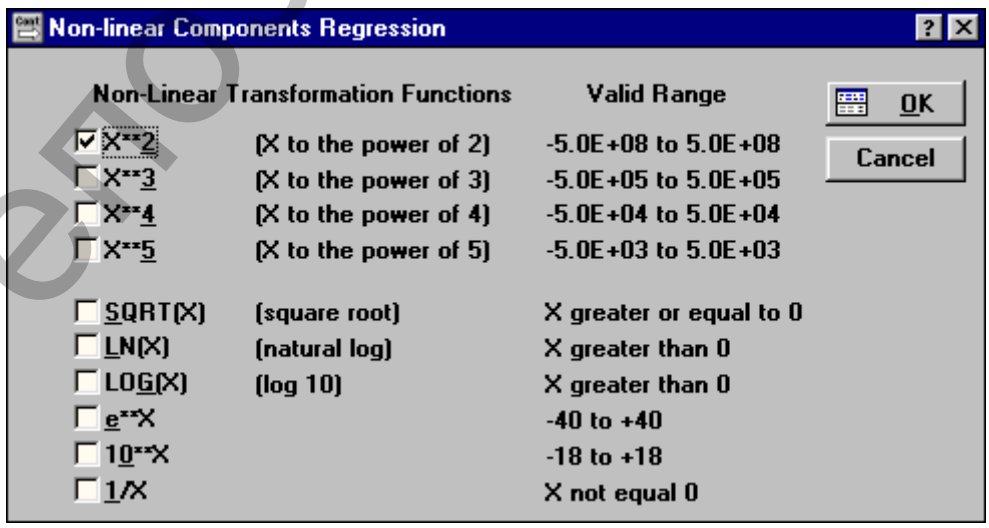


Рис. 44. **Окно выбора типов преобразования переменных.**

После того, как тип преобразования переменных определен (в нашем примере это возведение в квадрат), необходимо уточнение зависимой и независимых переменных фиксированной нелинейной регрессионной модели. Оно производится на следующем шаге при помощи кнопки Variables диалогового окна Model Definition (Уточнение модели) (рис. 45).

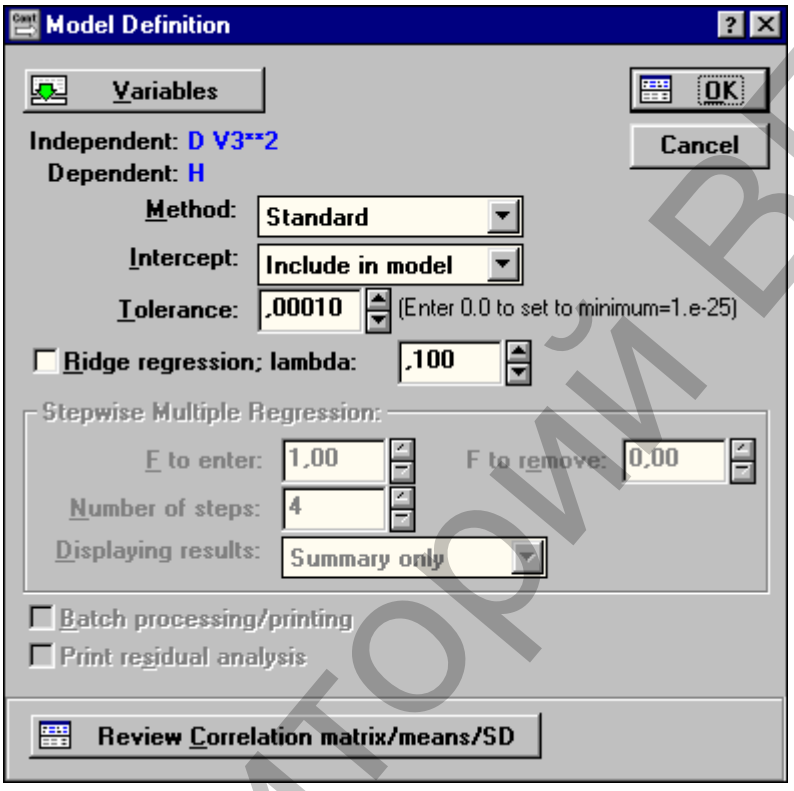


Рис. 45. **Диалоговое окно Model Definition (Уточнение модели).**

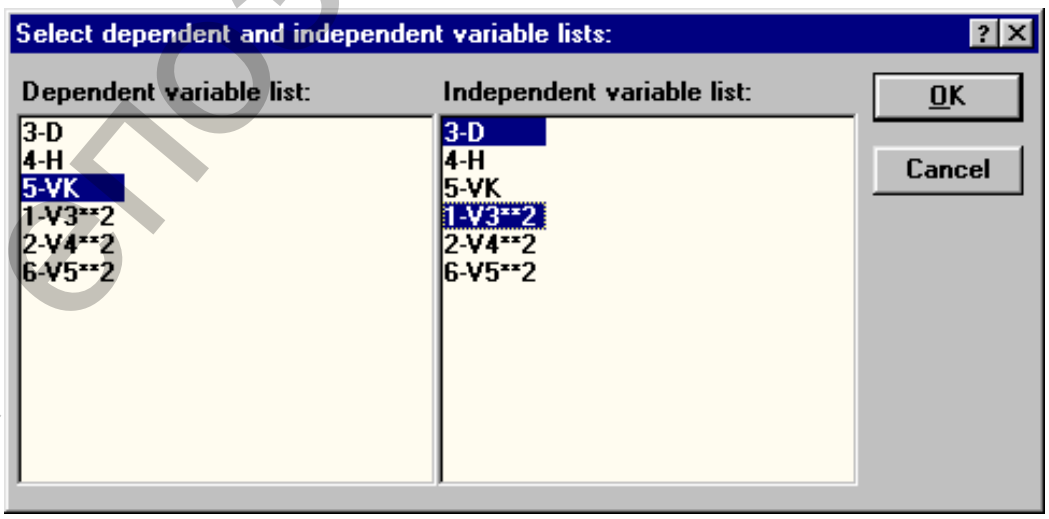


Рис. 46. **Выбор переменных для расчета уравнения**
$VK = a_1 + a_2D + a_3D^2$

Зависимой (dependent) переменной в нашем случае будет – VK; независимыми (independent) – D и D^2 (рис. 46). Переменная D^2 значится в списке переменных как V3**2, так как переменная D является третьей в списке переменных.

Уравнение взаимосвязи между объемом ствола дуба в коре (VK) от его диаметром (D) оказалось следующим: VK = 0,00023 – 0,0034D + + 0,0008D^2. Все коэффициенты уравнения (за исключением свободного члена) значимы на 5%-м уровне (p-level < 0,05). Это уравнение объясняет 95,8% (R^2 = 0,958) вариации зависимой переменной (рис. 47).

Regression Summary for Dependent Variable: VK (modl.sta)

Continue...

R= ,97900982 RI= ,95846023 Adjusted RI= ,95804274
F(2,199)=2295,8 p<0,0000 Std.Error of estimate: ,03087

N=202	BETA	St. Err. of BETA	B	St. Err. of B	t(199)	p-level
Intercpt			,0023	,0087	,268	,7891
D	-,1561	,0568	-,0034	,0012	-2,750	,0065
V3**2	1,1292	,0568	,0008	,0000	19,888	0,0000

Рис. 47. **Результаты регрессионного анализа модели $VK = a_1 + a_2D + a_3D^2$**

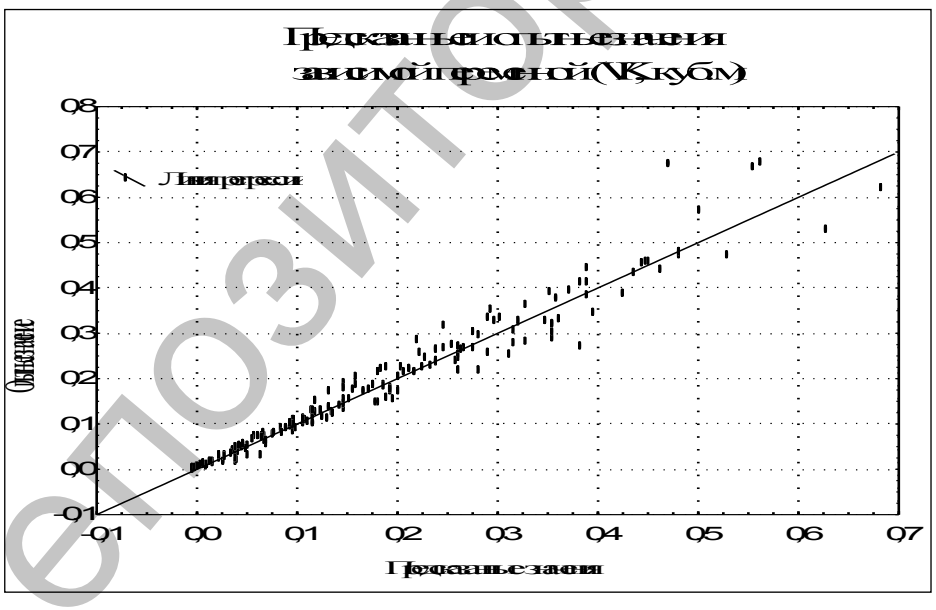


Рис. 48. **Линия регрессии, опытные и полученные по регрессионному уравнению значений зависимой переменной.**

По всем стандартным параметрам второе уравнение регрессии значительно лучше первого. Это наглядно подтверждает и график на рис. 49.

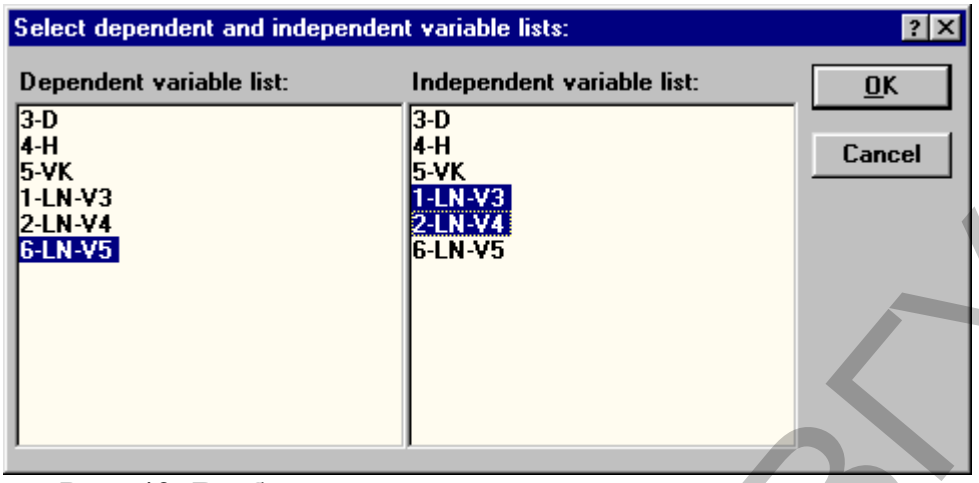


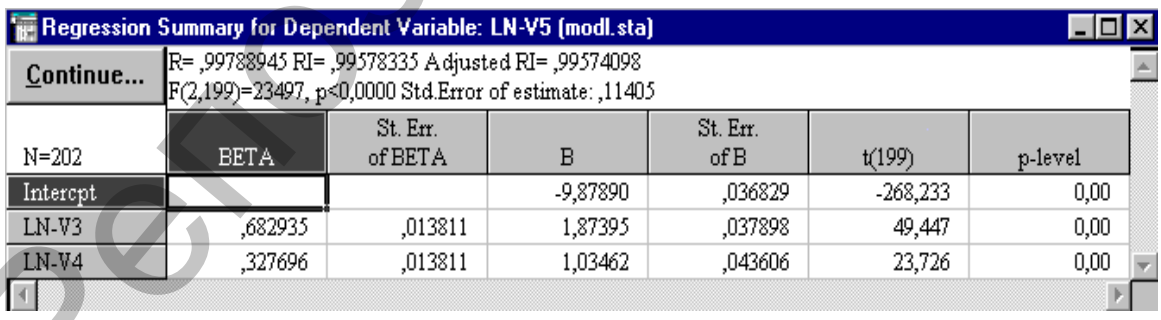
Рис. 49. **Выбор переменных для расчета уравнения**
$\ln VK = \ln a_1 + a_2 \ln D + a_3 \ln H$.

Найдем параметры еще одного регрессионного уравнения. Вид уравнения: $VK = a_1 D^{a2} H^{a3}$. Это степенное уравнение может быть приведено к линейному виду через логарифмирование:

$$\ln VK = \ln a_1 + a_2 \ln D + a_3 \ln H.$$

При помощи кнопки Variables укажем зависимую – VK и независимые переменные – D, H. Опцию Mode стартового окна регрессионного анализа (рис. 26) выставим в положение Fixed non linear. В качестве типа преобразования переменных выберем натуральный логарифм (ln (X)). В диалоговом окне Model Definition при помощи кнопки Variables уточним модель, переопределив зависимую и независимые переменные так, как это показано на рис. 49.

Основные результаты регрессионного анализа представлены на рис. 50.



N=202	BETA	St. Err. of BETA	B	St. Err. of B	t(199)	p-level
Intercpt			-9,87890	,036829	-268,233	0,00
LN-V3	,682935	,013811	1,87395	,037898	49,447	0,00
LN-V4	,327696	,013811	1,03462	,043606	23,726	0,00

Рис. 50. **Результаты регрессионного анализа модели**
$\ln VK = \ln a_1 + a_2 \ln D + a_3 \ln H$.

Уравнение выглядит следующим образом: lnVK = -9,8789 + + 1,8739lnD + 1,0346lnH или в степенном виде: $VK = 0{,}00005\ D^{1{,}8739}\ H^{1{,}0346}$. Все коэффициенты уравнения значимы на 5%-м уровне (p-level < 0,05). Это уравнение объясняет 99,6% ($R^2 = 0{,}996$) вариации зависи-

мой переменной. Ошибка уравнения 0,11405. Чтобы выразить ее в процентах, сравним абсолютную величину ошибки со средним значением зависимой переменной (lnVK): 0,11405/2,46166*100% = 4,6%.

Проверим адекватность полученной модели через анализ остатков. В целом он даст положительное заключение. В качестве иллюстрации приведем лишь несколько графиков (рис. 51, 52), подтверждающих такой вывод.

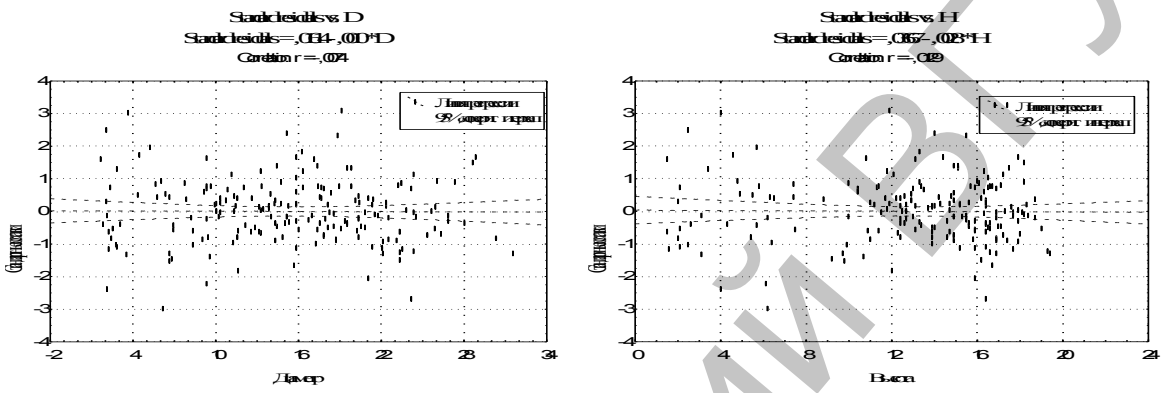


Рис. 51. **Зависимость остатков степенного уравнения от независимых переменных: диаметра и высоты.**

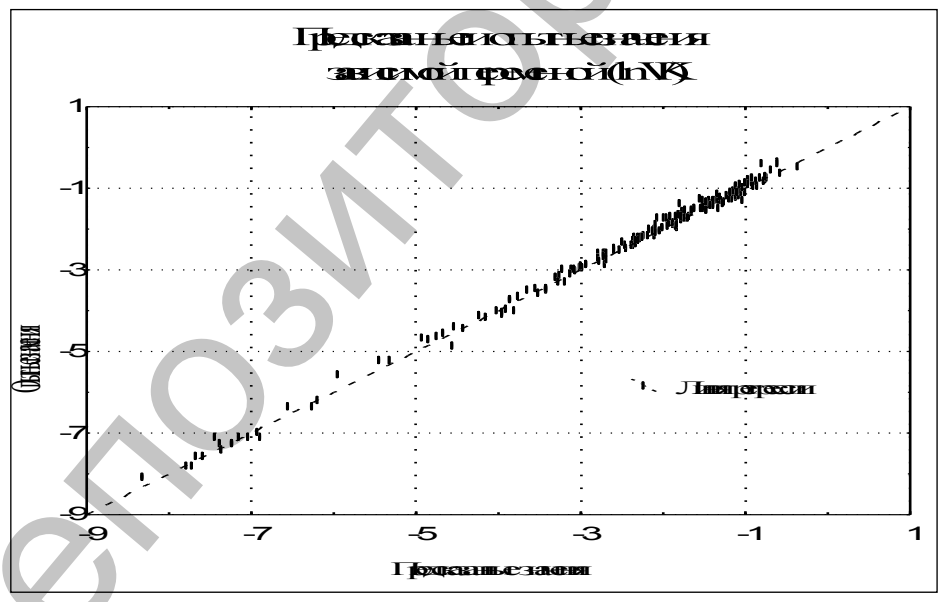


Рис. 52. **Линия регрессии, опытные и полученные по степенному регрессионному уравнению значений зависимой переменной.**

Поиск наилучшей регрессионной модели представляет собой довольно громоздкий процесс. При помощи опции **Method** (рис. 26) пользователь может отказаться от стандартного проведения регрессионного анализа (Standard) и воспользоваться методами пошагового включения переменных в регрессионную модель (**Forward stepwise**)

или пошагового исключения переменных (**Backward stepwise**) из регрессионной модели. Опция **Displaying results** позволяет просматривать или же только итоговые результаты регрессионного анализа (Summary only) или после каждого шага включения или исключения переменных (At each step). Если необходимо получить регрессионную модель без свободного члена уравнения, тогда в списке поля **Intercept** нужно выбрать – Set to zero.

Воспользуемся методом пошагового включения переменных для нахождения наилучшего регрессионного уравнения, описывающего объем ствола дуба в коре (VK). В качестве независимых переменных, которые потенциально могут быть включены в модель, примем: диаметр ствола (D), квадрат диаметра (D^2), высота ствола (H), квадрат высоты ствола (H^2), произведение диаметра ствола на его высоту (DH), квадрат произведения диаметра ствола на его высоту ($(DH)^2$).

В начале создадим новую переменную – DH. В файле данных она будет одиннадцатой по счету. Для расчета значений этой переменной вызовем окно с экспликацией этой переменной (рис. 53) и в поле **Long name** введем формулу, в соответствии с которой значения переменной должны быть рассчитаны, т.е. "=V3*V4".

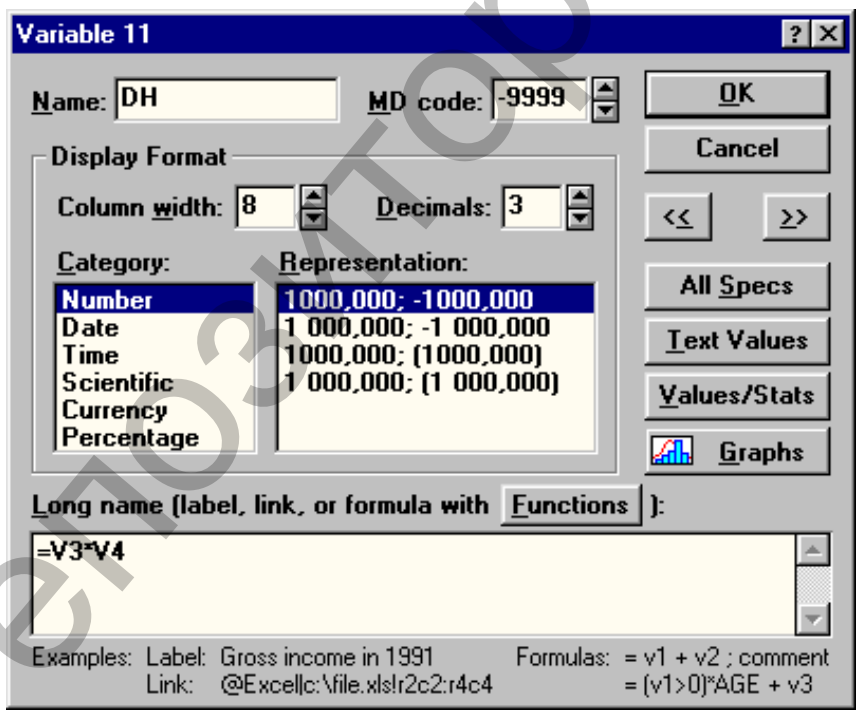


Рис. 53. **Окно экспликации 11-й переменной.**

Опцию **Mode** стартового окна регрессионного анализа (рис. 26) выставим в положение **Fixed non linear**.

Определим тип преобразования переменных – возведение в квадрат (рис. 44) и уточним зависимую и независимые переменные модели (рис. 54).

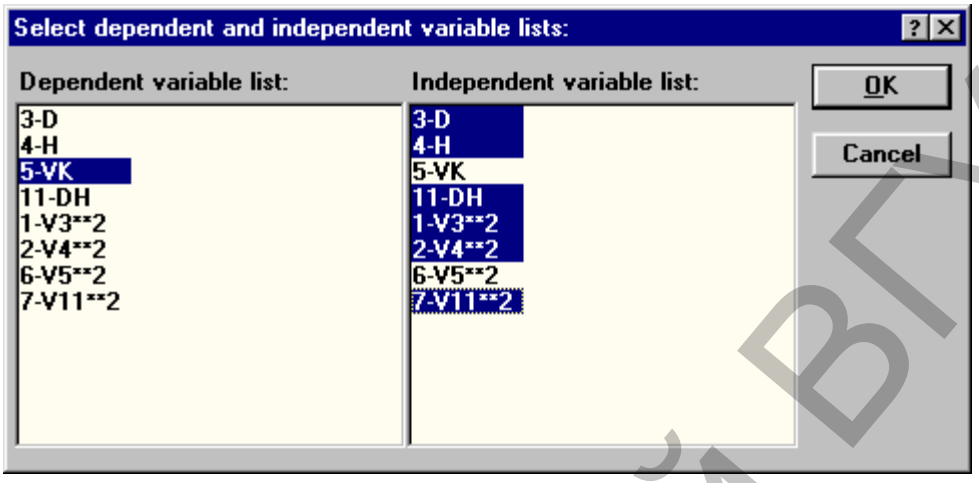


Рис. 54. **Уточнение зависимой и независимых переменных регрессионного анализа.**

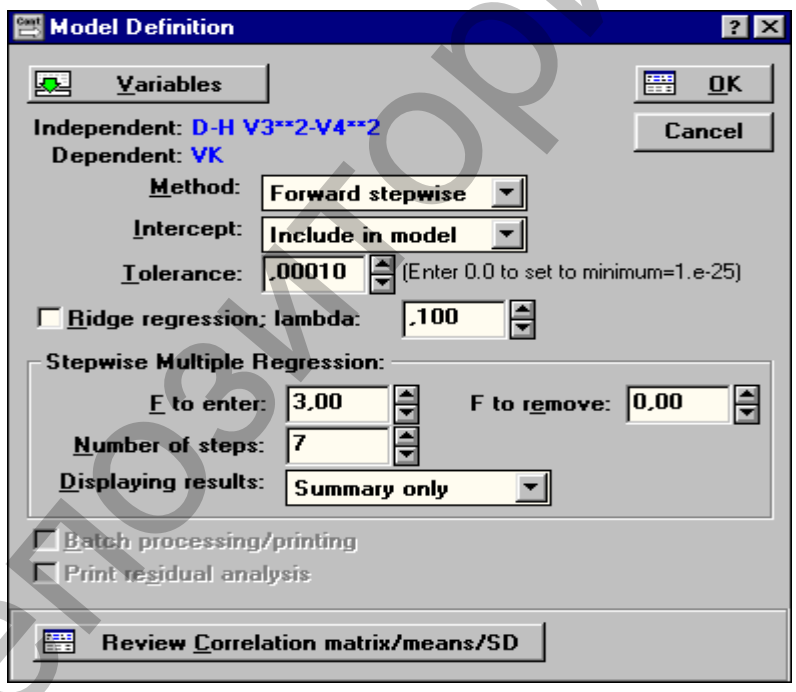
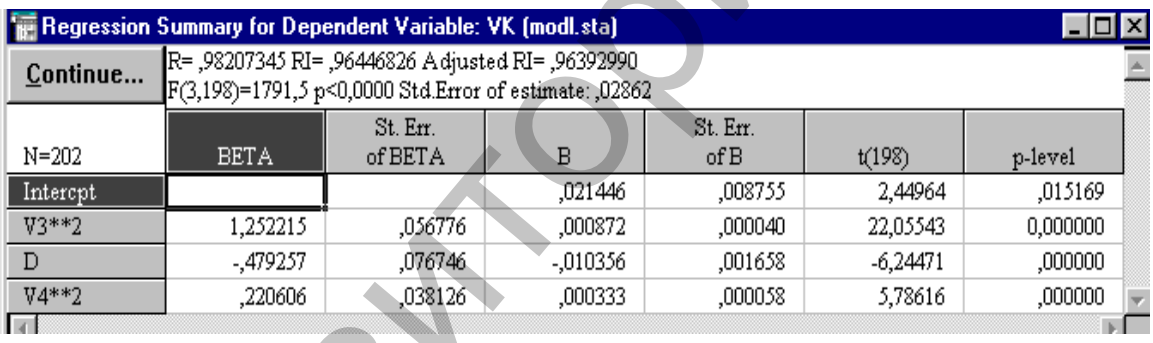


Рис. 55. **Диалоговое окно Model Definition при использовании метода пошагового включения переменных в модель.**

Для пошаговых методов регрессионного анализа важно установить величину **Tolerance** (толерантность) и величины частного F-критерия для включения в модель (**F to enter**) и исключения из нее (**F to remove**). Установив величину толерантности, мы создаем барьер для включения в модель переменных, толерантность которых меньше

установленной. Если величина толерантности переменной мала, то переменная несет малую дополнительную информацию и включение ее в модель не целесообразно. Какая-либо новая независимая переменная, включаемая в модель, может сильно влиять на зависимую переменную, но если она включается в модель после других переменных, то она может уже мало влиять на переменную отклика (например, из-за сильной коррелированности с переменными, уже включенными в модель). По умолчанию в пакете Statistica переменная включается в модель, если частный F-критерий больше или равен 1. Численное значение F-критерия для включения никогда не выбирается меньшим, чем численное значение F-критерия для исключения.

Выставим опции окна Model Definition так, как показано на рис. 55. В результате процедуры пошагового включения переменных в регрессионную модель получено следующее уравнение (рис. 54): $VK = 0{,}0214 + 0{,}0009D^2 - 0{,}0104D + 0{,}0003(DH)^2$. Все коэффициенты уравнения значимы на 5%-м уровне (p-level < 0,05). Это уравнение объясняет 96,4% ($R^2 = 0{,}964$) вариации зависимой переменной (рис. 56). Средняя ошибка уравнения составляет 0,02862 $м^3$.


Regression Summary for Dependent Variable: VK (modl.sta)

Continue...

R= ,98207345 RI= ,96446826 Adjusted RI= ,96392990
F(3,198)=1791,5 p<0,0000 Std.Error of estimate: ,02862

N=202	BETA	St. Err. of BETA	B	St. Err. of B	t(198)	p-level
Intercpt			,021446	,008755	2,44964	,015169
V3**2	1,252215	,056776	,000872	,000040	22,05543	0,000000
D	-,479257	,076746	-,010356	,001658	-6,24471	,000000
V4**2	,220606	,038126	,000333	,000058	5,78616	,000000

Рис. 56. **Характеристика уравнения, полученного методом Forward stepwise.**

При поиске лучшей регрессионной модели следует руководствоваться следующими наиболее общими требованиями (Дрейпер, Смит, 1981):

1. Регрессионная модель должна объяснять не менее 80% вариации зависимой переменной, т.е. $R^2 \geq 0.8$.
2. Стандартная ошибка оценки зависимой переменной по уравнению должна составлять не более 5% среднего значения зависимой переменной.
3. Коэффициенты уравнения регрессии и его свободный член должны быть значимы на 5%-м уровне.

4. Остатки от регрессии должны быть без заметной автокорреляции (r<0,30), нормально распределены и без систематической составляющей.

Чем меньше сумма квадратов остатков, чем меньше стандартная ошибка оценки и чем больше R^2 , тем лучше уравнение регрессии.

Одним из недостатков классического регрессионного анализа, в основе которого лежит метод наименьших квадратов, является недостаточная устойчивость к изменениям входной информации. Сейчас довольно широко стали применяться альтернативные регрессионные модели, одной из которых является **гребневая регрессия,** которая отличается устойчивостью для случаев сильной коррелированности зависимых переменных друг с другом. В отличии от метода наименьших квадратов, дающего несмещенные оценки коэффициентов уравнения, в методе гребневой регрессии оценки смещенные, но при этом они имеют меньшую дисперсию. Поэтому такие оценки могут давать более точные и приемлемые для практического использования модели (Забелин, 1983).

Для расчета гребневой регрессии следует установить флажок в опции **Ridge regression** диалогового окна Model Definition.

При практическом использовании метода гребневой регрессии одним из основных вопросов является выбор параметра **(λ lambda).** Существует несколько численных методов расчета параметра, но чаще используют простой эмпирический подход: выбирают такой параметр λ, при котором коэффициенты стабилизируются и при дальнейшем увеличении параметра изменяются мало. Значение принятого параметра λ является мерой смещения оценок от истинного значения, поэтому стараются не придавать λ слишком больших значений.

Обычно λ выбирают меньше 0,5, а шаг при подборе выбирают небольшим, например, 0,02 (Уланова, Забелин, 1990). При λ > 0,02, уравнение имеет коэффициенты классического метода наименьших квадратов.

ЗАДАНИЯ ДЛЯ КОНТРОЛЬНОЙ РАБОТЫ

Контрольная работа № 1
Вариант 1

1. Что такое корреляция?
2. Какая разница между корреляционной и функциональной зависимостями?
3. Какая разница между положительной и отрицательной корреляциями?
4. Что такое корреляционная решетка? Можно ли судить о характере корреляции по расположению данных в корреляционной решетке?
5. Коэффициент корреляции и его значения.
6. У окуня озера Баторино измерены длина головы *x* и длина грудного плавника *y*:

x 10,7 10,8 10,6 10,7 10,1 11,2 11,4 12,1 12,3 12,0
y 11,2 10,9 10,5 10,5 9,6 11,2 11,3 12,2 12,1 11,7

Определите коэффициент корреляции между признаками, оцените его достоверность и установите доверительные границы при P=0,05.

Вариант 2

1. Что такое корреляция?
2. Какая разница между корреляционной и функциональной зависимостями?
3. Какая разница между положительной и отрицательной корреляциями?
4. Что такое корреляционная решетка? Можно ли судить о характере корреляции по расположению данных в корреляционной решетке?
5. Коэффициент корреляции и его значения.
6. Получены следующие данные о продолжительности беременности у кроликов породы шиншилла (y) при различных размерах помета (x):

x 1 8 3 5 6 7 4 8 3 4
y 33 30 31 31 31 32 31 31 32 33

Определите коэффициент корреляции между признаками, оцените его достоверность и установите доверительные границы при P=0,05.

Вариант 3

1. Что такое корреляция?
2. Какая разница между корреляционной и функциональной зависимостями?
3. Какая разница между положительной и отрицательной корреляциями?

4. Что такое корреляционная решетка? Можно ли судить о характере корреляции по расположению данных в корреляционной решетке?
5. Коэффициент корреляции и его значения.
6. Учитывали плодовитость самок серебристо-черных лисиц (х) в совхозе «Белорусский» и плодовитость их дочерей (у):
х 6 7 5 6 5 5 4 5 5 4
у 4 5 4 4 6 2 3 3 2 6
Определите коэффициент корреляции между признаками, оцените его достоверность и установите доверительные границы при P=0,05.

Вариант 4

1. Что такое корреляция?
2. Какая разница между корреляционной и функциональной зависимостями?
3. Какая разница между положительной и отрицательной корреляциями?
4. Что такое корреляционная решетка? Можно ли судить о характере корреляции по расположению данных в корреляционной решетке?
5. Коэффициент корреляции и его значения.
6. Были получены следующие данные о весе х (в г) левой камеры сердца и длине ядер у (в μ) в мышцах сердца:
х 207 221 256 262 273 289 291 292 304 328
у 16,6 18,0 15,9 20,7 19,4 19,8 11,7 21,0 23,0 13,6
Определите коэффициент корреляции между признаками, оцените его достоверность и установите доверительные границы при P=0,05.

Вариант 5

1. Что такое корреляция?
2. Какая разница между корреляционной и функциональной зависимостями?
3. Какая разница между положительной и отрицательной корреляциями?
4. Что такое корреляционная решетка? Можно ли судить о характере корреляции по расположению данных в корреляционной решетке?
5. Коэффициент корреляции и его значения.
6. У 10 экземпляров днепровского ерша были изучены: длина тела х и вес у:
х 10,0 10,0 10,4 10,4 10,5 10,5 106, 10,7 10,7 10,7
у 19,0 20,0 28,0 35 27 26 28 28 30 27
Определите коэффициент корреляции между признаками, оцените его достоверность и установите доверительные границы при P=0,05.

Вариант 6

1. Что такое корреляция?
2. Какая разница между корреляционной и функциональной зависимостями?
3. Какая разница между положительной и отрицательной корреляциями?
4. Что такое корреляционная решетка? Можно ли судить о характере корреляции по расположению данных в корреляционной решетке?
5. Коэффициент корреляции и его значения.
6. У щук были измерены вес тела x и вес икры y:

x 456 375 484 56 788 7900 9581 3550 478 783

y 32 34 24 19 126 744 42 579 49 138

Определите коэффициент корреляции между признаками, оцените его достоверность и установите доверительные границы при P=0,05.

Контрольная работа № 2

Вариант 1

1. В таблице приведены данные результатов измерений предела текучести и предела прочности 10 марок стали. Постройте диаграмму рассеяния и сделайте предварительные выводы о наличии линейной связи между рассматриваемыми признаками. Рассчитайте коэффициент корреляции для исследуемых выборок, сделайте вывод о тесноте связи, используя таблицу 1.

Текучесть	58	145	94	113	85	121	119	112	85	41
Прочность	75	161	107	141	97	127	138	125	97	72

Рассчитайте ошибку коэффициента корреляции по формуле $m_r = \pm 1 - r^2 \big/ \sqrt{n}$.

Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$. Сделайте вывод значимости линейной связи.

Таблица 1

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Вариант 2

1. В таблице приведены данные результатов исследований зависимости оценки студентов по предмету от количества посещенных лекций.

Кол-во лекций	14	2	16	15	4	8	10	10	7	11
Оценка	4	3	5	4	2	3	4	3	3	4

Постройте диаграмму рассеяния и сделайте предварительные выводы о наличии линейной связи между рассматриваемыми признаками. Рассчитайте коэффициент корреляции для исследуемых выборок, сделайте вывод о тесноте связи. Рассчитайте ошибку коэффициента корреляции по формуле $m_r = \pm 1 - r^2 / \sqrt{n}$.

Таблица 1

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$. Сделайте вывод значимости линейной связи.

Вариант 3

1. В таблице приведены данные результатов изучения протяженности кроны от вершины и количества на дереве женских почек.

Протяженность кроны, м	0	1	2	3	4	5	6	7	8	9
Количество ж. почек	5	11	15	13	14	4	0	13	10	10

Постройте диаграмму рассеяния и сделайте предварительные выводы о наличии линейной связи между рассматриваемыми признаками. Рассчитайте коэффициент корреляции для исследуемых выборок, сделайте вывод о тесноте связи. Рассчитайте ошибку коэффициента корреляции по формуле $m_r = \pm 1 - r^2 / \sqrt{n}$.

Таблица 1

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$. Сделайте вывод значимости линейной связи.

Вариант 4

1. В таблице приведены данные результатов изучения протяженности кроны от вершины и количества на дереве мужских почек.

Протяженность кроны, м	0	1	2	3	4	5	6	7	8	9
Количество м. почек	0	12	0	27	41	66	135	190	31	64

Постройте диаграмму рассеяния и сделайте предварительные выводы о наличии линейной связи между рассматриваемыми признаками. Рассчитайте коэффициент корреляции для исследуемых выборок, сделайте вывод о тесноте связи. Рассчитайте ошибку коэффициента корреляции по формуле $m_r = \pm 1 - r^2 / \sqrt{n}$.

Таблица 1

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$. Сделайте вывод значимости линейной связи.

Вариант 5

1. В таблице приведены данные результатов измерений высот деревьев и диаметра стволов.

Высота, м	25	23	27	29	14	17	23	22	21	23
Диаметр, см	30	23	32	36	13	19	28	21	24	22

Постройте диаграмму рассеяния и сделайте предварительные выводы о наличии линейной связи между рассматриваемыми признаками. Рассчитайте коэффициент корреляции для исследуемых выборок, сделайте вывод о тесноте связи. Рассчитайте ошибку коэффициента корреляции по формуле $m_r = \pm 1 - r^2 / \sqrt{n}$.

Таблица 1

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$. Сделайте вывод значимости линейной связи.

Вариант 6

1. В таблице приведены данные результатов измерений урожайности растений (семян, шт.) в зависимости от количества внесенных удобрений (в мкг).

Кол-во семян	12	10	16	11	11	15	10	11	12	12
Кол-во удобрений	15	18	44	36	50	22	24	19	10	14

Постройте диаграмму рассеяния и сделайте предварительные выводы о наличии линейной связи между рассматриваемыми признаками. Рассчитайте коэффициент корреляции для исследуемых выборок, сделайте вывод о тесноте связи. Рассчитайте ошибку коэффициента корреляции по формуле $m_r = \pm 1 - r^2 / \sqrt{n}$. Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$. Сделайте вывод значимости линейной связи.

Таблица 1

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Вариант 7

1. В таблице приведены данные результатов измерений всхожести семян (%) при разной температуре.

Всхожесть	96	95	99	85	81	90	62	78	70	95
Температура	19	18	18	16	15	18	15	16	16	20

Постройте диаграмму рассеяния и сделайте предварительные выводы о наличии линейной связи между рассматриваемыми признаками. Рассчитайте коэффициент корреляции для исследуемых выборок, сделайте вывод о тесноте связи. Рассчитайте ошибку коэффициента корреляции по формуле $m_r = \pm 1 - r^2 / \sqrt{n}$.

Таблица 1

Коэффициент корреляции	Теснота связи
до 0,30	слабая
0,31-0,50	умеренная
0,51-0,70	значительная
0,71-0,90	высокая
0,91 и более	очень высокая

Проверьте значимость полученного значения коэффициента корреляции по критерию Стьюдента $t = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$. Сделайте вывод значимости линейной связи.

Контрольная работа № 3

Вариант 1

1. Дайте определение совокупности. Чем отличается выборочная совокупность от генеральной?
2. Среднеквадратичное отклонение. Общая формула.
3. Было сделано 5 определений содержания кальция в крови (в усл. единицах): 11,27; 11,36; 11,09; 11,16; 11,47. Вычислите среднее значение, его ошибку и дисперсию признака.
4. При изучении роста 10 лабораторных крыс коэффициент вариации веса крыс оказался равным 13%, а среднее значение веса – 200 г. Определите дисперсию признака и ошибку среднего значения.

Вариант 2

1. Что характеризует точность опыта? Основная формула для определения точности опыта.
2. Что такое мода и медиана вариационного ряда? Способы определения этих показателей. Могут ли совпадать значения средней, Мо и Ме?
3. Исследована длина тела (мм) плотвы оз. Швакшта. Получены следующие данные: 143, 143, 128, 130, 143, 127, 94, 157, 119, 127. Постройте вариационный ряд, вычислите величину асимметрии признака в выборке.
4. Применили три разных метода определения хлорофилла на выборках из 12 листьев растений, при этом получили следующие статистические показатели (в мг):

$$X_1 = 61{,}4\ \delta_1 = 5{,}22$$
$$X_2 = 337\ \delta_2 = 31{,}2$$
$$X_3 = 13{,}71\ \delta_3 = 1{,}2$$

Сравните коэффициенты вариации при разных методах и сделайте выводы.

Вариант 3

1. Что такое варианта? Классификация вариант.
2. Определение вариационного ряда, кривой распределения. Каковы возможные причины многовершинности вариационных кривых?
3. У студентов исследовали биение пульса (ударов в мин). Получены следующие данные: 43, 48, 52, 54, 55, 55, 60, 62, 65, 66, 70, 69, 69, 70, 72, 73, 78, 84, 88, 91. Ранжируйте варианты, разбейте на классы, определите среднее значение и его ошибку, дисперсию признака.
4. Были установлены следующие показатели высоты в холке (в см):

$$X_{ср} = 61{,}4\ \delta = 5{,}22$$

для телят 60,3

для молодых коров 100,5

Каждая выборка состояла из 15 животных. Оцените достоверность полученных средних. Отличаются ли эти показатели по степени изменчивости в изученных группах?

Вариант 4

1. Асимметрия и эксцесс. Формулы для вычисления.
2. Среднее значение. Типы средней, способы ее вычисления.
3. Обработайте следующие данные о длине третьего верхнего предкоренного зуба у ископаемого млекопитающего: 3,2; 3,1; 2,6; 2,8; 2,7; 3,0; 2,9; 3,4; 2,8; 3,0; 2,9; 3,0; 3,1; 3,0; 3,1; 3,3; 2,9; 2,9; 2,9; 2,8; 3,0. Вычислите основные статистические характеристики признака в совокупности.
4. Было установлено, что в группе свиней средняя скорость роста составляет 560 г в день. Определите дисперсию признака и ошибку средней, если известно, что точность опыта 2,4%, а изменчивость признака около 10%.

Контрольная работа № 4

Вариант 1

1. Были получены следующие данные о весе тушканчиков (*Dipus aegyptius*):

♂: 186, 190, 165, 182, 182, 182, 180, 173, 157, 179

♀: 162, 163, 190, 188, 147, 146, 145, 157, 162, 186

Выполните процедуру описательной статистики. Сделайте вывод об изменчивости веса самцов и самок. Оцените точность опыта, достоверность средних значений. Отличаются ли по весу самцы от самок?

2. У каждой из 10 самок было подсчитано число особей в помете.

♀: 162, 163, 190, 188, 147, 146, 145, 157, 162, 186

помет: 5, 4, 5, 3, 5, 5, 4, 4, 4, 4

Определите, есть ли зависимость между весом самки и величиной помета? С какой вероятностью?

Вариант 2

1. Температура тела тушканчиков (*Dipus aegyptius*) оказалась следующей:

♂: 37,5 37,9 37,4 37,8 36,8 37,8 37,5

♀: 37,8 38,1 37,0 37,5 37,7 37,8 37,6

Выполните процедуру описательной статистики. Сделайте вывод об изменчивости температуры тела самцов и самок. Оцените точность опыта, достоверность средних значений. Отличаются ли по температуре тела самцы от самок?

2. У самцов кроме температуры был измерен вес тела:
♂: 186, 190, 165, 182, 182, 182, 180
t, ºС: 37,5 37,9 37,4 37,8 36,8 37,8 37,5
Определите, есть ли зависимость между весом самки и величиной помета? С какой вероятностью?

Вариант 3

1. Для 7 коров известны следующие данные об их убойном весе (в кг) в теплом состоянии *x* и после охлаждения *y*:
x: 322,6 250,6 287,3 408,1 338,0 213,5 323,3
y: 318,9 247,0 279,7 403,0 334,7 209,3 319,2
Выполните процедуру описательной статистики. Сделайте вывод об изменчивости веса в теплом состоянии и после охлаждения. Оцените точность опыта, достоверность средних значений. Отличается ли теплый убойный вес от охлажденного?
2. Были получены следующие данные о весе (x) и длине туловища (y) у 7 серебристо-черных лисиц:
x: 4,7 4,6 5,2 5,1 5,3 5,3 4,6
y: 70 65 69 70 66 68 65
Определите, есть ли зависимость между весом и длиной туловища? С какой вероятностью?

Вариант 4

1. Для определения pH применили 2 типа электродов. При первом показания pH: 5,78; 5,74; 5, 84; 5,80; при втором – 5,82; 5,87; 5,96; 5,89. Выполните процедуру описательной статистики. Сделайте вывод об изменчивости показаний в первом и втором случаях. Следует ли отбросить нулевую гипотезу?
2. Имеются следующие данные об удоях 6 коров-матерей и их дочерей по полновозрастным лактациям:
Удой матерей: 3770 3817 2450 3463 3500 5544
Удой дочерей: 2991 4593 3529 4273 3130 3947
Определите, есть ли зависимость между показателями? С какой вероятностью?

ЛИТЕРАТУРА

1. Боровиков В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов – СПб.: Питер, 2001. – 656 с.

Книга представляет собой популярно написанный самоучитель по работе в статистическом пакете STATISTICA. Достаточно подробно рассмотрены технология работы в данном пакете, тогда как теория методов изложена гораздо скупее. К недостаткам книги можно отнести описания работы с рядом многомерных методов, реализованных в этом пакете.

2. Бейли Н. Математика в биологии и медицине. – М.: Мир, 1970. – 270 с.

Первая часть книги посвящена методологии применения различных разделов математики в экспериментальных исследованиях, рассматривается организация научных исследований и роль вычислительной техники в них. Во второй части проводится математический анализ конкретных биологических и медицинских проблем (теория эпидемий, экология и рост популяций, математические методы медицинской диагностики).

3. Броуди М. О статистическом рассуждении. – М.: Статистика, 1968. – 70 с.

В книге рассматриваются некоторые понятия математической статистики: «сбор», «классификация данных», «оценка разброса», «выборочные методы», «понятие о средних», анализ тенденций (анализ временных рядов, применение графиков, анализ корреляций).

4. Бююль А., Цёфель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. – СПб.: ДиаСофтЮП, 2002. – 608 с.

В книге дан минимально необходимый объем сведений по теории статистического анализа. Основное внимание сконцентрировано на особенностях использования отдельных методов, возможностях, которые этот метод предоставляет, а также интерпретации результатов применения данных методов. Книга предназначена для широкого круга читателей, специализирующихся в маркетинге, социологии, психологии, биологии и медицине.

5. Вайнберг Дж., Шумекер Дж. Статистика. – М.: Статистика, 1979. – 390 с.

Авторы последовательно излагают основные понятия статистики, дают характеристику важнейших статистических показателей и методов их получения (средние, вариация, группировка данных, распределения, проверка статистических гипотез, регрессия, корреляция, дисперсия, непараметрические критерии). Математический аппарат привлекается в самом необходимом объеме. Язык книги прост и доходчив. Основные понятия детально объясняются.

6. Гланц С. Медико-биологическая статистика. – М.: Практика, 1998. – 459 с.

В книге описан небольшой набор основных методов, которыми пользуются современная статистика. В частности, параметрические и непараметрические критерии, анализ выживаемости, анализ связей, планирование исследований и т.п. Учитывая, что автор книги является врач, самостоятельно освоивший описываемые статистические

методы, книга не свободна от некоторых недостатков. В частности, не рассмотрены имеющиеся ограничения на использование ряда популярных статистических критериев. Однако основным недостатком книги следует считать полное игнорирование автором современных методов многомерной статистики, которые в настоящее время приобретают все большую популярность.

7. Горя В.С. Алгоритмы математической обработки результатов исследований. – Кишинев: из-во Штиинца, 1978. – 118 с.

В пособии приведены распространенные методы статистической обработки данных, изложенные в виде алгоритмов и математических моделей (анализ вариационных рядов, ковариационный, дисперсионный анализ, анализ качественной изменчивости). Это позволило создать компактные схемы, удобные при массовых вычислениях. Изложение материала способствует лучшей ориентировке в разнообразных методах статистической обработки данных и выбору наиболее подходящего метода.

8. Лакин Г.Ф. Биометрия. – М.: Высшая школа, 1990. – 352 с.

В книге рассмотрены основные понятия биометрии, числовые характеристики описания совокупности эмпирических данных, законы распределения, построение статистических оценок, параметрические и непараметрические методы проверки статистических гипотез, дисперсионный, корреляционный, регрессионный анализ и некоторые вопросы планирования эксперимента.

9. Любищев А.А. Дисперсионный анализ в биологии. – М.: Изд-во Моск. ун-та, 1986. – 200 с.

Александр Александрович Любищев – человек удивительной судьбы! Его можно назвать современным Дон Кихотом, который всю свою жизнь боролся с косностью в науке и в нашем обществе. Единственный представитель СССР в Международном биометрическом обществе. 30 июля 1955 г. он закончил статью «Об аракчеевском режиме в биологии», посвятив ее изобличению уродливого монополизма Лысенко в биологии. Рекомендуемая нами книга посвящена специфике использования дисперсионного анализа в биологии. Насыщенность книги реальными примерами делает ее незаменимым учебником.

10. Налимов В.В., Голикова Т.И. Логические основания планирования эксперимента. 2-е изд., перераб. и доп. – М.: Металлургия, 1980. – 152 с.

Для тех, кто желает кратко и ясно узнать о том, что же такое есть «теория планирования эксперимента», можно смело порекомендовать эту книгу. Прочитав ее, вы поймете, что теория планирования эксперимента есть не что иное, как «теория здравого смысла» в экспериментальных исследованиях. Минимум математики и максимум здравомыслия – вот основное отличие этой книги от ей подобных. Прочитайте эту книгу, и вы поймете, почему необходимо применять планирование эксперимента.

11. Налимов В.В. Теория эксперимента. – М.: Наука, 1971. – 208 с.

В книге делается попытка показать, как под влиянием идей математической статистики формируется математическая теория эксперимента. Рассматриваются несколько наиболее интересных типов математических моделей. Излагаются основные концепции математической статистики – рандомизация условий проведения эксперимента, стратегия последовательного эксперимента и т.д. Излагаются методы статистического исследования, основанного на изучении рассеяния, и методы планирования эксперимента, основанные на оптимальном использовании пространства независимых переменных. Книга представляет собой свое-

образный путеводитель по идеям математической статистики, интересный для экспериментатора.

12. Никифоровский В.А. Вероятностный мир. – М.: Наука, 1992. – 174 с.

Теория вероятностей – одна из важнейших и интереснейших ветвей математики. Возникнув из задач, связанных с азартными играми, страхованием, обработкой результатов наблюдений, демографией, правосудием, она за сравнительно короткий срок выросла в ведущую науку; ее методы позволяют осознавать закономерности окружающего нас мира и широко применяются во многих теоретических и прикладных науках. В книге прослеживается возникновение и развитие теории вероятностей от ее основоположников – Паскаля, Ферма, Гюйгенса, Бернулли – до наших дней. Популярное издание может быть рекомендовано для знакомства с основами теории вероятностей читателям широкого круга – от школьников и студентов до исследователей в области медицины, биологии, химии, техники и т.д.

13. Плошко Б.Г., Елисеева И.И. История статистики. – М.: Финансы и статистика, 1990. – 295 с.

Показано зарождение и развитие статистической науки и практики за рубежом и в России. Особое внимание уделяется истории отечественной статистики, отмечаются ее плодотворность в 20-е годы XX в. и недооценка в последующий период. В книге много интересных фактов и исторических эпизодов, показывающих достаточно быстрое развитие статистической методологии в России конца прошлого и начала этого века. Авторы книги – ведущие специалисты Госкомстата (в недавнем прошлом – ЦСУ СССР).

14. Плохинский Н.А. Математические методы в биологии. – М.: МГУ, 1978. – 226 с.

Пособие может быть использовано для первоначального ознакомления с методами прикладной математики в биологии. Даны основные понятия о средних, корреляционном, дисперсионном анализе, рассмотрены математические модели биологических процессов (способ наименьших квадратов, способ Чебышева), информационные показатели (информация, энтропия). Вторая часть книги представляет собой справочник алгоритмов биометрии. Имеются необходимые математические таблицы.

15. Рейхман У.Дж. Применение статистики. – М.: Статистика, 1969. – 296 с.

Книга представляет собой популярное изложение вопросов применения статистики, рассчитанное на студентов и массового читателя. Книга охватывает довольно широкий круг вопросов: критерии достоверности статистической информации, приближенные и обобщенные показатели, применение теории вероятности, индексного метода, методы выборочных обследований и факторного анализа, линейное программирование и теория игр.

16. Румшиский Л.З. Математическая обработка результатов эксперимента. – М.: Наука, 1971. – 192 с.

В книге даются рекомендации по точечным и интервальным (доверительным) оценкам измеряемой величины. Даются простейшие методы проверки гипотез и основные сведения о корреляционных зависимостях, а также даются эффективные методы численного дифференцирования и интегрирования функций, заданных экспериментом. Все рекомендации сопровождаются примерами их практического применения. Книга предназначена для инженеров, а также студентов и аспирантов вузов.

17. Серафинович Л.П. Статистическая обработка опытных данных. – Томск, 1980. – 74 с.

В учебном пособии рассматриваются вопросы обработки статистических данных. Приведен систематизированный порядок обработки статистических данных, который иллюстрируется примерами. Рассматриваются вопросы построения рядов распределения, подбор и расчет наиболее распространенных теоретических законов распределения, определения числовых характеристик, а также правильность выбора теоретического закона распределения по критериям согласия. Пособие носит практический характер, компактно и удобно для выполнения курсовых и лабораторных работ.

18. Снедекор Дж.У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. – М.: Сельхозгиз, 1961. – 503 с.

В книге описаны методы математической статистики, используемые при обработке результатов экспериментов. Рассмотрены виды распределения, различные параметрические и непараметрические методы, регрессионный, корреляционный, дисперсионный анализ, ковариация, планирование и анализ выборочных наблюдений. Книга имеет характер практического руководства, в ней нет математического обоснования и вывода рекомендуемых формул. Материал изложен в доступной форме, приведено много подсобных таблиц, необходимых для оценки достоверности выводов, полученных при обработке экспериментальных данных.

19. Статистический словарь / гл. ред. М.А. Королев. – 2-е изд., перераб. и доп. – М.: Финансы и статистика, 1989. – 623 с.

Последнее издание словаря на данную тему. Нельзя сказать, что это издание достаточно полное и адекватно отражает современное состояние статистики, тем более что с момента выхода книги минуло много лет. Однако для тех, кто желает более систематизированно представить отдельные разделы статистики на уровне определений и основной терминологии, эта книга будет хорошим подспорьем. Однако надо понимать, что это далеко не учебник.

20. Факторный, дискриминантный и кластерный анализ: пер. с англ. / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др.; под ред. И.С. Енюкова. – М.: Финансы и статистика, 1989. – 215 с.

Лучшая из всех известных автору монографий, содержащих описание и примеры использования таких сложных многомерных методов, как факторный, дискриминантный и кластерный анализ. Книга содержит три отдельных очерка по этим методам, каждый из которых можно читать и изучать приведенные в них материалы отдельно. Используя минимум математики, авторам удалось рассказать просто о сложном. Книга может быть рекомендована как для первого чтения, так и для тех, кто уже знаком с этими методами.

21. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. – М.: Финансы и статистика, 1983. – 302 с.

Книга представляет собой пособие по широко применяемым методам статистической обработки данных – корреляционному и регрессионному анализу. Дает читателю, не знакомому со статистическим анализом, ключ к пониманию прикладных работ в интересующих его областях.

22. Хургин Я.И. Как объять необъятное. М.: Знание, 1979, – 192 с.

В этой книге, по своему характеру очень близкой популярному среди читателей типу изданий «занимательных наук», живо и ненавязчиво излагаются идеи

современной математической статистики, раскрывается ее основа – теория вероятностей, обсуждаются проблемы измерения, описания сложных систем и построения их математических моделей. Автор книги – доктор физико-математических наук, профессор. Предназначается для широкого круга читателей.

23. Афифи А., Эйзен С. Статистический анализ: подход с использованием ЭВМ / пер. с англ. – М.: Мир, 1982. – 488 с.

Монография американских ученых, рассчитанная на читателей, знакомых с основами математической статистики, но не имеющих опыта работы с ЭВМ и не знающих программирования. Изложение ориентировано на применение пакета прикладных программ, приведены примеры из биологии, медицины, гуманитарных наук.

24. Вараксин А.Н. Статистические модели регрессионного типа в экологии и медицине. – Екатеринбург: Изд-во «Гощинский», 2006. – 256 с.

В монографии рассмотрены вопросы построения и анализа моделей регрессионного типа в экологии и медицине. Дана краткая характеристика медико-экологической информации, приведены основные формулы, используемые при построении моделей, с комментариями применительно к экологии и медицине, обсуждаются вопросы применимости регрессионных моделей, дан обзор публикаций, посвящённых моделям регрессионного типа в экологии и медицине.

25. Вейр Б. Анализ генетических данных / пер. с англ. – М.: Мир, 1995. – 400 с.

В монографии известного американского ученого, написанной как учебное пособие, изложены современные методы статистического анализа дискретных популяционно-генетических данных. Эти методы получили новый импульс к развитию в связи с разработкой международной программы «Геном человека». В книге, наряду с теоретическим обоснованием определенных подходов, рассматриваются конкретные примеры расчетов, приводятся статистические таблицы и компьютерные программы для специалистов-генетиков и молекулярных биологов, студентов старших курсов и аспирантов.

26. Владимирский Б.М. Математические методы в биологии. – Ростов: изд-во Рост. ун-та, 1983. – 304 с.

В учебном пособии рассматриваются вопросы теории и практики использования математических методов и средств вычислительной техники для описания и моделирования процессов, протекающих в живых системах. На многочисленных примерах показано использование современных методов планирования и анализа экспериментальных данных, существенно повышающих эффективность исследований в самых разных разделах биологии и медицины. Описаны методы оценки взаимосвязей, описательные статистики, использование критериев согласия, элементы анализа динамических рядов, основные методы многомерного статистического анализа. В приложении даны некоторые статистические таблицы.

27. Гублер Е.В., Генкин А.А. Применение непараметрических критериев статистики в медико-биологических исследованиях. – Л.: Медицина, 1973. – 141 с.

В книге рассмотрены различные непараметрические критерии, приведены многочисленные примеры, отображающие опыт авторов в их применении, описана методика выбора наиболее адекватного критерия в каждом случае. Обширное приложение содержит необходимые таблицы.

28. Компьютерная биометрика / под ред. В.Н. Носова. – М.: МГУ, 1990. – 232 с.

В монографии рассмотрены современные методы статистической обработки данных и использование соответствующего программного обеспечения ЭВМ в биологических исследованиях. Изложены элементы теории вероятностей как математические основы статистических методов. Даны рекомендации по использованию пакетов прикладных программ (STATGRAPHICS, SYSTAT, SAS, ППП BMDP) и алгоритмических языков ЭВМ. Книга предназначена для специалистов биологов и медиков.

29. Лиепа И.Я. Математические методы в биологических исследованиях. Факторный и компонентный анализы. – Рига, 1980. – 104 с.

Учебное пособие предназначено для студентов 3–5 курсов биологических факультетов. Излагается общая теория факторного и компонентного анализов, методы определения факторных нагрузок, ортогональная и неортогональная ротация факторов, вычислительные матрицы факторных весов. Каждый приводимый метод иллюстрируется примером.

30. Лисенков А.Н. Математические методы планирования многофакторных медико-биологических экспериментов. – М.: Медицина, 1979. – 344с.

В книге рассматриваются принципы планирования многофакторных экспериментов. Описаны элементы дисперсионного анализа, планы 2^k, многоуровневые и несимметричные факторные планы, методы экстремального и отсеивающего эксперимента, процедуры множественных сравнений и непараметрические методы анализа многофакторных экспериментов. Излагаемый материал иллюстрируется примерами планирования и анализа, реальных медико-биологических экспериментов.

31. Максимов В.Н. Многофакторный эксперимент в биологии. – М.: МГУ, 1980. – 279 с.

На примерах, взятых из практики биологических исследований, изложены основные принципы планирования многофакторных экспериментов. Подробно рассматриваются вопросы статистической обработки результатов экспериментов и их интерпретация на основе получаемых уравнений регрессии. Описаны полный и дробный факторные планы. Книга для биологов, биохимиков, физиологов, микробиологов, гидробиологов и экологов.

32. Малиновский Л.Г. Классификация объектов средствами дискриминантного анализа. – М.: Наука, 1979. – 260 с.

Рассматриваются вопросы многомерного статистического анализа для построения классификационных алгоритмов. Уделяется внимание проблеме связи используемого математического аппарата, основанного на функциях нормальных распределений, с реальными объектами и измерениями. Развивается методология математического исследования статистических закономерностей. Предлагаются новые выборочные критерии классификации и критерии, построенные с ограничением вероятности ошибок. Работоспособность полученных алгоритмов иллюстрируется примерами классификации ЭКГ.

33. Методы современной биометрии: сб. статей / отв. ред. Н.А. Плохинский. – М.: МГУ, 1978. – 207 с.

Сборник посвящен памяти известного советского ученого П.В. Терентьева и открывается его статьей по истории возникновения и развития биометрии. Пуб-

ликуются статьи, посвященные проблеме практического использования биометрии в исследовательской работе.

34. Планирование эксперимента в биологии и сельском хозяйстве / под ред. В.Н. Максимова. – М.: МГУ, 1991. – 220 с.

В учебном пособии популярно изложены основные методы математической теории планирования экспериментов. Изложена теория оптимальности планов, принципы регрессионного анализа. Приведены примеры практического использования многофакторных планов. Описаны пакеты прикладных программ по планированию экспериментов, факторных и полиномиальных моделей.

35. Рокицкий П.Ф. Биологическая статистика. – Минск: Вышэйшая школа, 1973. – 320 с.

Учебное пособие для биологических факультетов университетов по курсу биологической статистики. В книге подробно и последовательно изложены необходимые для биологических исследований статистические методы: группировка материала, составление вариационных рядов, вычисление важнейших статистических показателей, характеризующих совокупности, корреляционный, регрессионный и дисперсионный анализы, применение критерия соответствия. Особое внимание уделено понятиям вероятности и достоверности и их значению для анализа биологических данных. Каждая глава содержит проверочные вопросы и задачи (на материале ботаники, зоологии, физиологии, генетики, медицины и др.).

36. Терентьев П.В., Ростова Н.С. Практикум по биометрии. – Л.: ЛГУ, 1977. – 152 с.

В пособии рассматриваются основные методы статистического анализа биологических явлений (описательные показатели, статистические заключения, корреляционный, регрессионный, дисперсионный анализ) и их непараметрические аналоги, указаны цели и границы применения методов, рекомендуемая литература. По каждому из методов приводятся данные для самостоятельных упражнений, подсобные таблицы. Пособие предназначено для студентов, преподавателей, научных работников.

37. Урбах В.Ю. Биометрические методы. Статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине. – М.: Наука, 1964. – 416 с.

Настоящее руководство предназначено для исследователей в области биологии, медицины, встречающихся с необходимостью статистической обработки опытных данных. В книге излагаются все основные методы биометрии. Чтение книги не требует от читателя специальной математической подготовки, кроме знаний в объеме средней школы. Теоретический материал иллюстрирован примерами из биологии и смежных дисциплин. Особое внимание обращено на подробный разбор техники расчетов, играющей во всех применениях биометрии первостепенную роль. Книга содержит необходимые математико-статистические таблицы.

П Р И Л О Ж Е Н И Я

Приложение 1

Стандартные значения критерия Стьюдента

Число степеней свободы $u=n_1+n_2-2$	Критерий Стьюдента t_{st} при вероятности безошибочного заключения *p*			
	0.1	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.952
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.684	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.732	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.723	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.714	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
∞	1.645	1.960	2.326	2.576

Приложение 2

Значения критерия Фишера F

U_2	Степень свободы для большей дисперсии U_1											
	3	4	5	6	8	10	12	16	24	30	50	∞
5%-ный уровень значимости $F_{0,05}$												
3	9,3	9,1	9,0	8,9	8,8	8,8	8,7	8,7	8,6	8,6	8,6	8,5
4	6,6	6,4	6,3	6,2	6,0	5,9	5,8	5,8	5,8	5,7	5,7	5,6
5	5,4	5,2	5,1	5,0	4,8	4,7	4,7	4,6	4,5	4,5	4,4	4,4
6	4,8	4,5	4,4	4,3	4,2	4,1	4,0	3,9	3,8	3,8	3,8	3,7
7	4,4	4,1	4,0	3,9	3,7	3,6	3,6	3,5	3,4	3,4	3,4	3,2
8	4,1	3,8	3,7	3,6	3,4	3,4	3,4	3,2	3,1	3,1	3,0	2,9
9	3,9	3,6	3,5	3,4	3,2	3,2	3,1	3,0	2,9	2,9	2,8	2,7
10	3,7	3,5	3,3	3,2	3,1	3,0	2,9	2,8	2,7	2,7	2,6	2,5
12	3,5	3,3	3,1	3,0	2,9	2,8	2,7	2,6	2,5	2,5	2,4	2,3
16	3,2	3,1	2,9	2,8	2,7	2,5	2,4	2,3	2,2	2,2	2,1	2,0
18	3,2	2,9	2,8	2,7	2,5	2,4	2,3	2,3	2,2	2,1	2,0	1,9
24	3,0	2,8	2,6	2,5	2,4	2,3	2,2	2,1	2,0	1,9	1,9	1,7
40	2,8	2,6	2,5	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,7	1,5
120	2,7	2,4	2,3	2,2	2,0	1,9	1,8	1,7	1,6	1,6	1,5	1,3
∞	2,6	2,4	2,2	2,1	1,9	1,8	1,8	1,6	1,5	1,5	1,4	1,0
1%-ный уровень значимости $F_{0,01}$												
3	29,5	28,7	28,2	27,9	27,5	27,0	26,8	26,6	26,5	26,5	26,4	26,1
4	16,7	16,0	15,5	15,2	14,8	14,5	14,4	14,2	13,9	13,8	13,7	13,5
5	12,1	11,4	11,0	10,7	10,3	10,1	9,9	9,7	9,5	9,4	9,2	9,0
6	9,8	9,2	8,8	8,5	8,1	7,9	7,7	7,5	7,3	7,2	7,1	6,9
7	8,4	7,8	7,5	7,2	6,8	6,6	6,5	6,3	6,1	6,0	5,8	5,6
8	7,6	7,0	6,6	6,4	6,0	5,8	5,7	5,5	5,3	5,2	5,1	4,9
9	7,0	6,4	6,1	5,8	5,5	5,3	5,1	4,9	4,7	4,6	4,5	4,3
10	6,6	6,0	5,6	5,4	5,1	4,9	4,7	4,5	4,3	4,3	4,1	3,9
12	6,0	5,4	5,1	4,8	4,5	4,3	4,2	4,0	3,8	3,7	3,6	3,4
16	5,3	4,8	4,4	4,2	3,9	3,7	3,6	3,4	3,2	3,1	3,0	2,8
20	4,9	4,4	4,1	3,9	3,6	3,4	3,2	3,0	2,9	2,8	2,6	2,4
30	4,5	4,0	3,7	3,5	3,2	3,0	2,8	2,7	2,5	2,4	2,2	2,0
60	4,1	3,6	3,3	3,1	2,8	2,6	2,5	2,3	2,1	2,0	1,9	1,6
120	3,9	3,5	3,2	3,0	2,7	2,4	2,3	2,2	1,9	1,9	1,7	1,4
∞	3,8	3,3	3,0	2,8	2,5	2,3	2,2	2,0	1,8	1,7	1,5	1,0

Приложение 3

Критические значения χ^2
для трех степеней доверительной вероятности

Число степеней свободы, U	Уровень вероятности			Число степеней свободы, U	Уровень вероятности		
	0.95	0.99	0.999		0.95	0.99	0.999
1	3.8	6.6	10.8	26	38.9	45.6	54.1
2	6.0	9.2	13.8	27	40.1	47.0	55.5
3	7.8	11.3	16.3	28	41.3	48.3	56.9
4	9.5	13.3	18.5	29	42.6	49.6	58.3
5	11.1	15.1	20.5	30	43.8	50.9	59.7
6	12.6	16.8	22.5	32	46.2	53.5	62.4
7	14.1	18.5	24.3	34	48.6	56.0	65.2
8	15.5	20.1	26.1	36	51.0	58.6	67.9
9	16.9	21.7	27.9	38	53.4	61.1	70.7
10	18.3	23.2	29.6	40	55.8	63.7	73.4
11	19.7	24.7	31.3	42	58.1	66.2	76.1
12	21.0	26.2	32.9	44	60.5	68.7	78.7
13	22.4	27.7	34.5	46	62.8	71.2	81.4
14	23.7	29.1	36.1	48	65.2	73.7	84.0
15	25.0	30.6	37.7	50	67.5	76.2	86.7
16	26.3	32.0	39.3	55	73.3	82.3	93.2
17	27.6	33.4	40.8	60	79.1	88.4	99.6
18	28.9	34.8	42.3	65	89.8	94.4	106.0
19	30.1	36.2	43.8	70	90.5	100.4	112.3
20	31.4	37.6	45.3	75	96.2	106.4	118.5
21	32.7	38.9	46.8	80	101.9	112.3	124.8
22	33.9	40.3	48.3	85	107.5	118.2	131.0
23	35.2	41.6	49.7	90	113.1	124.1	137.1
24	36.4	43.0	51.2	95	118.7	130.0	143.3
25	37.7	44.3	52.5	100	124.3	135.8	149.4