

Министерство образования и науки Российской Федерации
Ярославский государственный университет им. П. Г. Демидова

С. И. Сиделев

**Математические методы
в биологии и экологии:
введение в элементарную биометрию**

Учебное пособие

*Рекомендовано
Научно-методическим советом университета для студентов,
обучающихся по направлениям
Биология и Экология и природопользование*

Ярославль 2012

УДК 51-7:57
ББК Е 0с21я73
С 34

*Рекомендовано
Редакционно-издательским советом университета
в качестве учебного издания. План 2011 года*

Рецензенты:

Крылов В. В., кандидат биологических наук;
лаборатория экологии водных беспозвоночных
Института биологии внутренних вод РАН

Сиделев, С. И. Математические методы в биологии и экологии: введение в элементарную биометрию: учебное пособие / С. И. Сиделев; Ярослав. гос. ун-т им. П. Г. Демидова. – Ярославль : ЯрГУ, 2012. – 140 с.

ISBN 978-5-8397-0859-4

Пособие представляет собой подробное изложение базовых понятий биометрии и некоторых приемов первичной количественной обработки биологических и экологических данных. Приводится разбор часто используемых в биологии и экологии статистических показателей, условий их применимости и алгоритмов расчета, раскрыты основы выборочного метода исследований, статистического оценивания и проверки гипотез. Рассмотрены основы дисперсионного анализа. Большое внимание уделено применению специального программного обеспечения в процессе анализа данных.

Предназначено для студентов, обучающихся по направлениям 020400.62 Биология и 020200.62 Экология и природопользование (дисциплины «Математические методы в биологии» и «Математические методы в экологии», блок Б 2), очной формы обучения.

УДК 51-7:57
ББК Е 0с21я73

ISBN 978-5-8397-0859-4

© Ярославский государственный
университет им. П. Г. Демидова, 2012

*Посвящается светлой памяти
Сергея Владимировича Тихонова*

Введение

Основоположниками преподавания математических методов на факультете биологии и экологии Ярославского государственного университета им. П. Г. Демидова являются профессор Л. А. Жаков и старший преподаватель кафедры экологии и зоологии С. В. Тихонов. Методической основой дисциплины «Математические методы в биологии и экологии», преподаваемой на факультете в течение многих лет С. В. Тихоновым, явилось применение в курсе «Основы рыбного хозяйства» (1980–1990 гг.) учебных имитационных моделей.

Задачи освоения дисциплины – обучение студентов применению современных методов обработки и анализа биологических и экологических данных, основанных на использовании математической статистики и современной вычислительной техники. Основное назначение предлагаемого учебного пособия значительно скромнее – ознакомление студентов биологических и экологических специальностей с базовыми понятиями и элементарными методами количественного анализа данных наблюдений (экспериментов). Освоение самых простых приемов биометрической обработки данных позволит студентам осознанно подойти к использованию более сложных математических методов при решении разнообразных исследовательских задач.

Необходимость преподавания курса на факультете биологии и экологии ЯрГУ определяется тремя обстоятельствами. Во-первых, курс помогает в освоении и более глубоком понимании учебного материала по дисциплинам специализации, поскольку практические занятия и лекции основаны на многочисленных примерах из области биологии и экологии. Во-вторых, дисциплина «Математические методы в биологии и экологии» служит методической основой выполнения студентами курсовых и дипломных научных работ. Адекватное применение методов статистической обработки научных данных является необходимым условием успешной защиты выпускных квалификационных

работ, это повышает доказательность выводов и общий уровень научной работы студентов. Относительно сложные математические методы применяются учащимися в научных работах довольно редко, с другой стороны, такие элементарные статистические процедуры, как расчет среднего значения, показателей вариации, стандартной ошибки, доверительного интервала, построение статистических графиков и таблиц, используются повсеместно. Первые представления о способах расчета разнообразных описательных статистик и применении графических методов анализа данных с использованием различных пакетов статистических программ студенты получают на занятиях по дисциплине «Математические методы в биологии и экологии». В дальнейшем это значительно облегчает работу над оформлением и представлением результатов научного исследования. В-третьих, курс является хорошей основой для подготовки будущих научных сотрудников университетов и институтов. Общеизвестно, что биологические факультеты университетов, в том числе ЯрГУ, являются основными «поставщиками» высококвалифицированных специалистов для исследовательских институтов биологического и экологического профиля в нашей стране. Представить научного сотрудника, не владеющего математическими методами анализа данных, при современном уровне развития биологии и экологии довольно сложно. С другой стороны, следует признать, что в большинстве научных работ в нашей стране, в том числе публикуемых в рецензируемых журналах, совсем не применяются столь необходимые методы количественного анализа. Их применение заменяется аргументами типа «мы считаем», «по нашему мнению», «вероятно», при этом желаемое часто искажает действительное. Ещё более опасна ситуация, при которой математические методы обработки данных в научных работах используются некорректно, это приводит к ложным выводам. Огромное количество подобных ошибок в научных статьях и диссертациях читатель может найти на сайте www.biometrika.tomsk.ru в разделе «Кунсткамера». Предостеречь студентов от распространенных ошибок при количественной обработке данных ещё одна из задач данной книги.

Внедрение в современных условиях компьютерных технологий в процесс математического анализа данных является существенным условием. Поэтому первый совет автора пособия студентам и

начинающим исследователям – забыть о вычислениях «в столбик» или на калькуляторах, а проводить обработку данных только на персональном компьютере с использованием специализированных программных пакетов для статистического анализа.

Учебное пособие не содержит какой-либо существенно новой информации по количественным методам обработки данных в сравнении с другими аналогичными работами. Однако автор предпринял скромную попытку изложить материал, исходя из предпосылки заинтересованности большинства биологов и экологов в получении конкретных результатов количественной обработки имеющихся данных и их правильной интерпретации. При этом суть базовых понятий биометрии излагается, как хочется думать автору, в максимально подробной и простой форме. Поэтому при чтении книги специальной математической подготовки *не требуется!*

Автор сознает, что пособие не лишено недостатков и даже ошибочных суждений, поэтому будет благодарен специалистам по статистике за конструктивные критические замечания, направленные на улучшение пособия, которые можно отправлять на электронный адрес: Sidelev@mail.ru.

Глава 1. Общие вопросы применения количественных методов в биологии и экологии

1.1. Роль статистических методов в биологии и экологии

Значимость статистических методов в биологии и экологии определяется как самим характером современных исследований в этих областях, так и естественными свойствами объектов изучения.

Основной метод современной биологии и экологии – количественный, качественная интерпретация изучаемых явлений и процессов давно перестала быть достаточным и надежным инструментом для подтверждения или опровержения выдвигаемых гипотез, доказательства теоретических положений, установления причинно-следственных зависимостей, определения влияния факторов среды на свойства живых систем. Большинство

современных биологических и экологических исследований имеет дело с «лавиной» чисел, через которые выражаются данные о размерах, весе, возрасте, численности, биомассе, плодovitости организмов, продуктивности экосистем, урожайности сортов, концентрации веществ, соотношении между признаками, дозами факторов, различными количественными показателями и числовыми характеристиками.

На этот кажущийся первоначально хаотичным набор первичной числовой информации накладываются свойства самих объектов изучения, усиливающие разброс данных, в частности широкая изменчивость живых систем. Современная статистика оказывается столь полезной при обработке численных данных в биологии и экологии именно потому, что она основана на признании этой изменчивости и обладает мощными средствами её учета. В итоге, в кажущемся хаосе полученных цифр вдруг открываются конкретные закономерности, которые требуют объективной оценки. Подтверждение существования закономерного в видимом хаосе изменчивости достигается посредством использования методов статистического анализа. Применение прикладных методов статистики к сложным живым системам способствовало появлению нового направления в биологических науках и математике, которое получило название «*биометрия*». Кроме данного общепризнанного термина, использовались и используются другие – *биометрика*, *вариационная статистика*, *биологическая статистика*, *биома-тематика*, в последнее время *компьютерная биометрия*. Однако какие бы термины и громкие словосочетания ни применялись, суть данного научного направления остается фактически одной и той же – статистическая обработка результатов наблюдений и экспериментов в биологических науках (к коим относится и экология) с целью отделения закономерного от случайного, оценки разнообразных связей и зависимостей между биологическими явлениями, поиска причин, определения влияния фактора и т. д. Некоторые исследователи определяют биометрию как направление, опирающееся на *индуктивный* подход, идущий от конкретных эмпирических данных и фактов к теоретическим обобщениям. Путь от «эмпирики» к общим теориям «обслуживается» биометрическими методами, поэтому биометрию принято считать средством эмпирического познания природы. В основе биометрии лежат такие разделы математики, как теория вероятностей и матема-

тическая статистика. Другой путь называется *дедуктивным подходом*, при котором на первое место выдвигаются математические модели, основанные на теоретических обобщениях, с последующей проверкой моделей опытом. Этот путь «обслуживается» так называемой *математической биологией*, исследующей теоретические проблемы с помощью математического моделирования.

Характерной особенностью биометрии является применимость её методов не к единичным фактам, а только к их совокупностям, к массовым явлениям. Именно в сфере массовых случайных явлений обнаруживаются закономерности, не свойственные единичным объектам. В этом плане область приложения статистических методов в биологии и экологии очень значительна, так как многие экологические и биологические явления массовы по своей природе – в них участвуют не одна клетка, не одна особь, не одна бактерия, не один вид или популяция, а их совокупности, взаимодействующие между собой. Осуществление событий в таких совокупностях может быть оценено вероятностями. Такие проблемы, как изменчивость морфологических, физиологических, экологических признаков животных и растений, возрастная изменчивость органов у человека, установление влияния экологических факторов, количественный учет организмов, классификационные построения в систематике, изучение наследственности в генетике, индивидуальный рост организмов, популяционная динамика численности, особенности сукцессии экосистем, могут изучаться лишь с помощью математических и математико-статистических методов. С другой стороны, не всякое исследование в биологии и экологии должно и может опираться на биометрию, многие великие открытия были сделаны без использования количественных методов анализа. Но в тех областях биологических наук, где исследования проводятся на основе измерений и подсчетов, игнорирование статистической обработки полученного исследователем материала может привести к мало убедительным или даже ошибочным выводам. Напротив, корректное применение биометрических методов увеличивает доказательность сделанных заключений, помогает правильно планировать эксперименты, выявлять скрытые закономерности и правильно их интерпретировать, устанавливая причины наблюдаемых явлений, отделять их от следствий, выделять из множества воздействующих на явление факторов наиболее

важные, измерять силу их влияния, дает возможность получить точную количественную характеристику изменчивости исследуемых показателей, оценить достоверность проверяемой гипотезы, определить степень различий между признаками.

Несмотря на ценность применения методов статистики в биологии и экологии, существуют некоторые опасности, от которых следует предостеречь студентов и начинающих исследователей.

Первая из них – это механическое использование количественных методов анализа в исследованиях, без понимания их сути и приложимости к тем или иным биологическим явлениям и экологическим процессам. Очень важно знать и учитывать особенности и условия применения тех или иных статистических процедур, поскольку любой метод имеет свои ограничения. Без учета этих ограничений применение соответствующего метода становится математически неправомерно, это приводит к фальсификации результатов и выводов научной работы, к отклонению проверяемой гипотезы там, где на самом деле её нужно было бы принять, к установлению влияния фактора, который в реальности не влияет, к подтверждению не существующих связей между элементами системы. Описанные фальсификации могут возникать при формальном применении биометрических методов с целью создать лишь видимость строгой научности в той или иной исследовательской работе.

Вторая опасность связана с широко распространенным мнением о том, что математическая обработка данных может если не полностью учесть, то свести к минимуму те технические, организационные и методические ошибки, которые возникли при проведении исследования. На это часто надеются недобросовестные исследователи. Данное мнение глубоко ошибочно, статистические методы можно с равным успехом применять как к верным данным, так и к неправильно полученным. В данном случае работает принцип: «что посеешь, то и пожнешь». Поэтому биометрию можно сравнить с жерновом, «который всякую засыпку смелет, но ценность помола определяется исключительно ценностью засыпанного» (Лакин, 1990).

Наконец, осталось дать краткую **историческую справку**. В биологии и экологии использование математики началось значительно позже, нежели в физике и химии. Биологические науки долгое время развивались на основе только качественного

анализа явлений. Необходимость количественного анализа стала ясно осознанной только в конце XIX века. *Френсис Гальтон* (1899) разработал основы новой науки, названной им «биометрия». Ф. Гальтону принадлежит первая попытка применить статистические методы к решению проблемы наследственности и изменчивости организмов. Достойным продолжателем исследований Ф. Гальтона явился его ученик *Карл Пирсон*. Он создал математический аппарат биометрии, развил учение о разных типах кривых распределения, разработал критерий χ^2 («хи квадрат»), ввёл в биометрию такие показатели, как стандартное отклонение, коэффициент вариации. Следует отметить, что исследования Гальтона и Пирсона поначалу не получили признания у научной общественности и их статьи даже отказывались печатать в ведущих научных изданиях. Поэтому в 1901 г. Пирсон был вынужден организовать выпуск собственного журнала «*Biometrika*», который существует до сих пор и считается наиболее авторитетным изданием в своей области. Одной из причин недоверия к первым биометрическим работам было то, что биометрики акцентировали своё внимание на многочисленных рядах данных и фактически не интересовались анализом «малых выборок». Поэтому, даже приняв новый подход Гальтона и Пирсона, большинство исследователей не смогло бы использовать его на практике. Данную проблему разрешил англичанин *Вильям Госсет*, обосновав теорию малой выборки и представление о том, что даже для небольшого количества данных можно успешно использовать статистические методы. В. Госсет в 1908 г. под псевдонимом Student (Стьюдент) опубликовал свою известную работу «Вероятная ошибка средней», где описал разработанный им **t-критерий**. Дальнейшее развитие теория малой выборки получила в трудах выдающегося английского статистика *Рональда Фишера*. Его научные работы по праву можно считать вершиной классической и фундаментом современной биометрии. Он основатель дисперсионного анализа и статистической теории планирования экспериментов. Ценный вклад в развитие и пропаганду методов биометрии внесли и отечественные ученые: С. С. Четвериков, Ю. А. Филипченко, П. В. Терентьев, В. И. Василевич, Л. А. Животовский, А. А. Любищев, Н. А. Плохинский, Ю. А. Песенко, Н. С. Ростова, П. Ф. Рокицкий.

1.2. Программное обеспечение анализа данных

Программное обеспечение анализа данных можно условно разделить на пакеты общего назначения (MS Excel) и специальные программные продукты, которые, в свою очередь, делятся на математические программы (Mathematica, Matlab, Maple, Mathcad), статистические программы (Statistica, StatGraphics, SPSS, Stadia, Biostat, Systat, Attestat) и пакеты научной графики.

Наиболее доступными в вузах и широко применяемыми в научно-исследовательских организациях биологического и экологического профиля являются табличный процессор MS EXCEL и пакет STATISTICA. Хорошей альтернативой популярным и дорогостоящим зарубежным пакетам статистического анализа (Statistica, StatGraphics, SPSS) является удобная в работе русскоязычная статистическая программа ATTESTAT, доступная для бесплатного скачивания в сети Internet (<http://attestatsoft.narod.ru>, разработчик Игорь Гайдышев).

Табличный процессор MS EXCEL

Основными средствами анализа данных в MS EXCEL являются статистические и математические функции библиотеки встроенных функций (**Мастер функций**), статистические процедуры надстройки **Пакет анализа** и специальный инструмент для проведения графического анализа – **Мастер диаграмм**.

Функции – это заранее определенные формулы, с помощью которых можно быстро выполнять вычисления по заданным величинам, называемым аргументами. Список аргументов может состоять из чисел, текста, логических величин (например, ИСТИНА или ЛОЖЬ), массивов и др. Кроме того, аргументы могут быть как константами, так и формулами. Эти формулы, в свою очередь, могут содержать другие функции (Тихонов, 2003). Для доступа к функциям необходимо в меню **Вставка** выбрать указателем мыши строку **Функция**, откроется диалоговое окно, в котором можно выбрать необходимую функцию (рис. 1.1).

Структура функции в MS EXCEL начинается с указания имени функции, затем вводится открывающая скобка, указываются аргументы, отделяемые точками с запятыми, а затем – закрывающая скобка. Перед именем функции вводится знак равенства

(=) (Тихонов, 2003). Имена некоторых функций и их краткая характеристика приведены ниже (табл. 1.1).

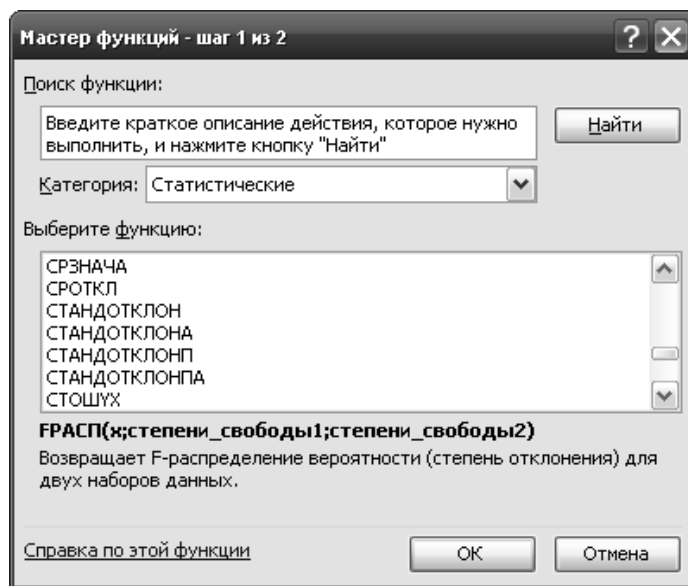


Рис. 1.1. Диалоговое окно «Мастер функций»

Таблица 1.1

***Некоторые математические и статистические функции
табличного процессора MS EXCEL***

<i>Категория</i>	<i>Название</i>	<i>Выполняемое действие</i>	<i>Список аргументов</i>
Математические	LN	Рассчитывает натуральный логарифм числа.	(число)
Математические	LOG	Рассчитывает логарифм числа по заданному основанию.	(число; основание)
Математические	LOG10	Рассчитывает десятичный логарифм числа.	(число)
Математические	КОРЕНЬ	Рассчитывает положительное значение квадратного корня.	(число)
Математические	СТЕПЕНЬ	Рассчитывает результат возведения в степень.	(число; степень)
Математические	СУММ	Суммирует числа.	(диапазон чисел)
Статистические	ДОВЕРИТ	Рассчитывает доверительный интервал для среднего генеральной совокупности.	(критический уровень значимости;

			стандартное отклонение; количество чисел в диапазоне)
Статистические	МАКС	Определяет наибольшее значение в выборке.	(диапазон чисел)
Статистические	МИН	Определяет наименьшее значение в выборке.	(диапазон чисел)
Статистические	МЕДИАНА	Определяет медиану.	(диапазон чисел)
Статистические	МОДА	Определяет наиболее часто встречающееся или повторяющееся значение.	(диапазон чисел)
Статистические	СКОС	Рассчитывает коэффициент асимметрии распределения.	(диапазон чисел)
Статистические	ЭКСЦЕСС	Рассчитывает коэффициент эксцесса распределения.	(диапазон чисел)
Статистические	СРГЕОМ	Рассчитывает среднее геометрическое.	(диапазон чисел)
Статистические	СРЗНАЧ	Рассчитывает среднее арифметическое.	(диапазон чисел)
Статистические	СТАНДОТКЛОН	Рассчитывает стандартное отклонение по выборке.	(диапазон чисел)
Статистические	ДИСП	Рассчитывает дисперсию по выборке.	(диапазон чисел)

В электронных таблицах также имеется пакет **Анализ данных**, вызываемый из меню **Сервис** (рис. 1.2). При отсутствии названия пакета в списке следует использовать пункт «Надстройки».

Статистические процедуры **Пакета анализа** обладают большими возможностями, чем статистические функции. С помощью данного модуля можно решать более сложные задачи обработки данных, применение отдельных процедур пакета для решения практических задач разбирается в некоторых разделах пособия. Однако следует отметить, что в **Пакет анализа** входят только параметрические методы, основанные на предположении о нормальности распределения изучаемых признаков, что ограничивает использование MS EXCEL по сравнению со специализированными программами.

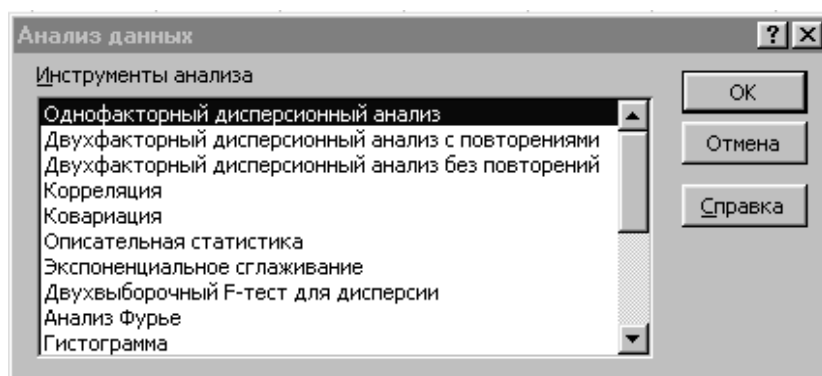


Рис. 1.2. Общий вид меню пакета «Анализ данных»

Наконец, в MS EXCEL есть возможность проводить графический анализ данных и создавать диаграммы и графики довольно приемлемого качества. Для этого имеется специальное средство – Мастер диаграмм (рис. 1.3), под руководством которого пользователь проходит 4 этапа процесса построения графика. Доступ к Мастеру диаграмм осуществляется через меню Вставка.

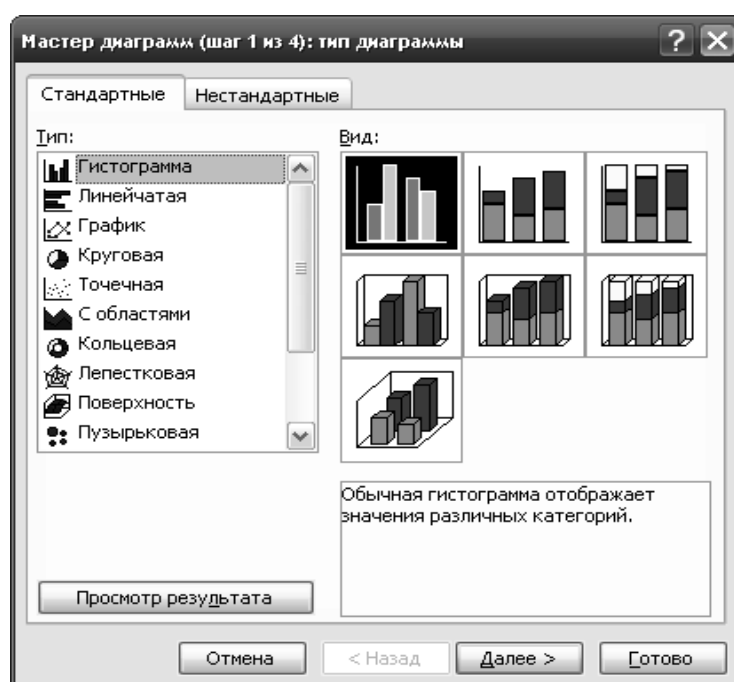


Рис. 1.3. Диалоговое окно «Мастер диаграмм»

Пакет прикладных программ STATISTICA

STATISTICA представляет собой интегрированную систему статистического анализа и обработки данных. Система состоит из следующих основных компонентов:

- многофункциональной системы для работы с данными, которая включает в себя электронные таблицы для ввода и задания исходных данных, а также специальные таблицы (*Scroolsheet*) для вывода численных результатов анализа. Для сложной (специализированной) обработки данных в STATISTICA имеется модуль «Управление данными»;

- графической системы для визуализации данных и результатов статистического анализа;

- набора статистических модулей, в которых собраны группы логически связанных между собой статистических процедур. В любом конкретном модуле можно выполнить определенный способ статистической обработки, не обращаясь к процедурам из других модулей. Каждый модуль является полноценным *Windows*-приложением. Поэтому пользователь имеет возможность одновременной работы как с одним, так и с несколькими модулями. Переключаться между ними можно как между обычными *Windows*-приложениями. Все основные операции при работе с данными и графические возможности доступны в любом статистическом модуле и на любом шаге анализа;

- специального инструментария для подготовки отчетов. При помощи текстового редактора, встроенного в систему, можно готовить полноценные отчеты. В STATISTICA также имеется возможность автоматического создания отчетов.

Данные в STATISTICA организованы в виде электронной таблицы – *Spreadsheet*. Они могут содержать как численную, так и текстовую информацию (рис. 1.4). Электронные таблицы в STATISTICA поддерживают различные типы операций с данными: операции с использованием *Буфера обмена Windows*; операции с выделенными блоками значений (аналогично MS EXCEL), в том числе и с использованием метода *Drag-and-Drop* – «Перетащить и опустить», автозаполнение блоков и т. д. (Тихонов, 2003).

Vars Cases		Data: Children.STA 21v * 308c								
TEXT	VALU	1	2	3	4	5	6	7	8	9
		ПОЛ	ВОЗРАСТ	РОСТ	ВЕС	ОГК1	ОГК2	ОГК3	ДИН1	ДИН2
1	f	f	6	128,5	24,0	67,0	63,0	59,0	10,0	5,0
2	f	f	6	135,0	25,5	61,0	57,5	54,0	15,5	10,0
3	f	f	6	126,0	20,3	63,0	59,0	55,0	9,0	9,0
4	f	f	6	124,0	20,2	60,0	56,0	53,0	15,0	6,0
5	f	f	6	128,5	20,4	60,0	56,0	53,0	10,0	10,0
6	f	f	6	124,6	20,5	68,0	62,0	57,0	6,0	10,0
7	f	f	6	124,0	19,8	58,0	57,0	56,0	3,0	4,0
8	f	f	6	124,0	20,0	58,0	57,0	56,0	6,0	4,0
9	f	f	6	130,0	25,6	62,0	60,0	60,0	3,0	6,0
10	f	f	6	125,0	27,2	65,0	63,0	62,0	15,0	13,0
11	f	f	6	124,0	21,1	60,0	58,0	56,0	5,0	9,0
12	f	f	6	124,0	21,1	59,0	57,5	57,0	10,0	12,0
13	f	f	6	122,0	23,8	62,0	60,0	59,0	10,0	12,0
14	f	f	6	112,0	17,0	56,0	54,0	53,0	9,0	5,0
15	f	f	6	129,0	24,0	59,0	57,0	56,0	14,0	11,0
16	f	f	6	122,5	24,6	63,0	62,0	61,0	5,0	5,0

Рис. 1.4. Электронная таблица системы STATISTICA.
 Кнопка **Vars** обеспечивает управление столбцами таблицы
 (переменными – **Variables**), **Cases** – строками (наблюдениями)
 (по: Тихонов, 2003)

Ввести данные в электронную таблицу можно одним из следующих способов:

- непосредственно ввести их в электронную таблицу с клавиатуры;
- воспользоваться данными, подготовленными в другом приложении, применяя копирование данных через *Буфер обмена*.

Численные результаты статистического анализа в системе STATISTICA выводятся в виде специальных электронных таблиц. Они могут содержать любую информацию (как численную, так и текстовую), от короткой строчки до мегабайтов результатов. Обычно даже в результате простейшего статистического анализа получается на выходе большое количество численной и графической информации. В системе STATISTICA эта информация выводится в виде последовательности таблиц и графиков (Тихонов, 2003). STATISTICA содержит большое количество инструментов для удобного просмотра результатов статистического анализа и их визуализации. К примеру, статистически значимые результаты выделяются в таблицах красным цветом.

Прямой доступ ко всем статистическим и графическим методам обработки данных осуществляется через меню **Statistics** (Статистика, Анализ) и меню **Graphs** (Графики) (рис. 1.5).

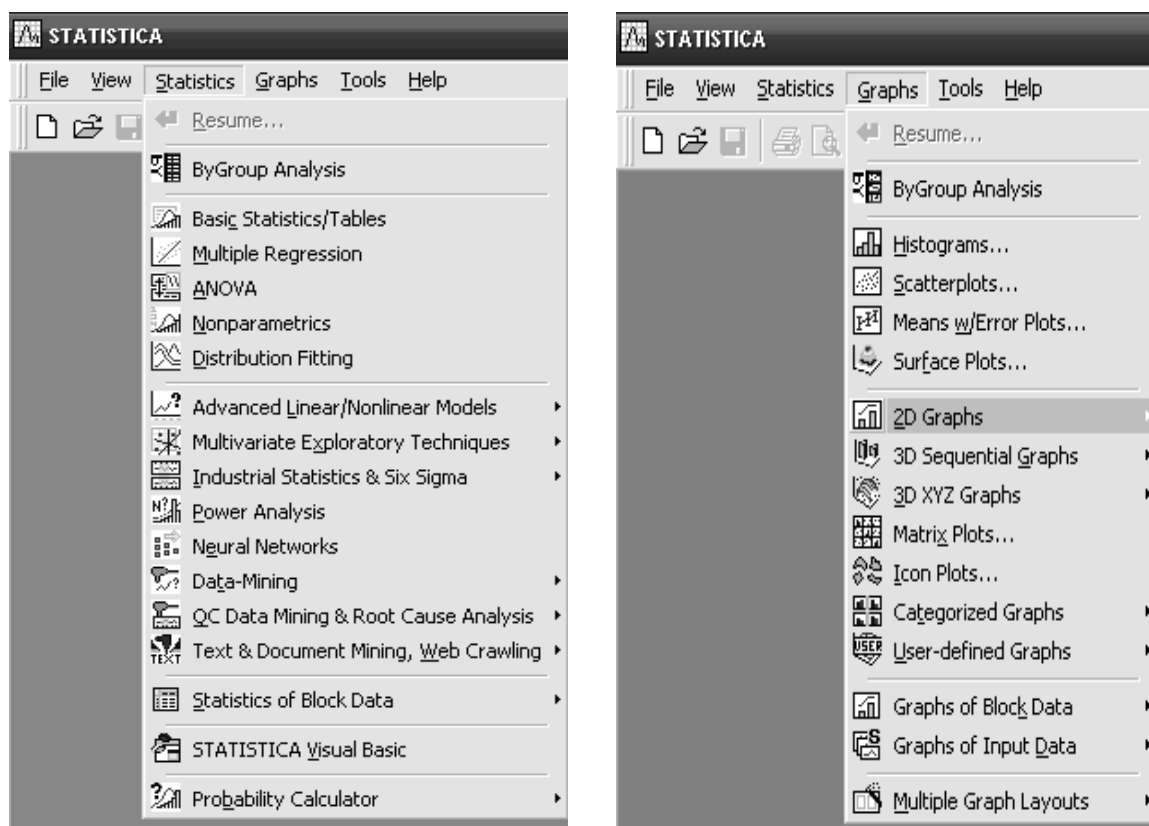


Рис. 1.5. Набор статистических и графических модулей анализа данных в меню «Statistics» и «Graphs» пакета STATISTICA 7

Статистические процедуры системы STATISTICA сгруппированы в нескольких специализированных статистических модулях. В каждом модуле можно выполнить определенный способ обработки. В системе реализованы процедуры непараметрических методов анализа и все другие из известных типов статистической обработки данных.

Система STATISTICA обладает широкими графическими возможностями, включая большое количество разнообразных категорий и типов графиков (научные, деловые, трехмерные и двухмерные графики в различных системах координат, гистограммы, матричные, категоризованные графики и др.).

В системе имеется возможность создания отчета, в окно которого может быть выведена вся информация. Отчет – это документ, который может содержать любую текстовую или графическую

информацию. При этом любая таблица *Scrollsheet* или график могут быть автоматически направлены в отчет (Тихонов, 2003).

Статистическая обработка данных в системе STATISTICA обычно состоит из следующих основных шагов:

- ввод исходных данных в электронную таблицу системы STATISTICA;
- визуализация данных при помощи того или иного типа графиков;
- статистический анализ при помощи некоторого статистического метода;
- вывод численных, текстовых и графических результатов на рабочее пространство системы и в файл с отчетом;
- анализ и интерпретация результатов.

Программа анализа данных ATTESTAT, версия 12.5

Программное обеспечение ATTESTAT предназначено для математико-статистического анализа данных. Программа выполнена в виде надстройки к популярным электронным таблицам MS EXCEL. Поэтому для работы с программным обеспечением ATTESTAT после его установки просто необходимо запустить табличный процессор MS EXCEL и зайти в добавленный программным обеспечением новый пункт меню *AtteStat*. На экране появится диалоговое окно программы с набором модулей (рис. 1.6.).

Конструктивно программа состоит из функционально независимых модулей, объединённых общим интегратором. Алгоритм работы во всех модулях одинаков, разберем его на примере модуля **Описательная статистика** (см. рис. 2.7 раздела 2.4).

1. *Первый шаг*: интервал переменной – необходимо при помощи указателя мыши ввести интервал ячеек с исходными данными из электронной таблицы MS EXCEL.

2. *Второй шаг*: интервал вывода – следует указать ячейку в электронной таблице MS EXCEL, начиная с которой будут выведены вычисленные статистические показатели.

3. *Третий шаг*: необходимо поставить указателем мыши флажки напротив тех статистических показателей, рассчитать которые необходимо пользователю.

4. *Четвертый шаг*: для вывода результатов анализа надо нажать кнопку **Выполнить расчет**.

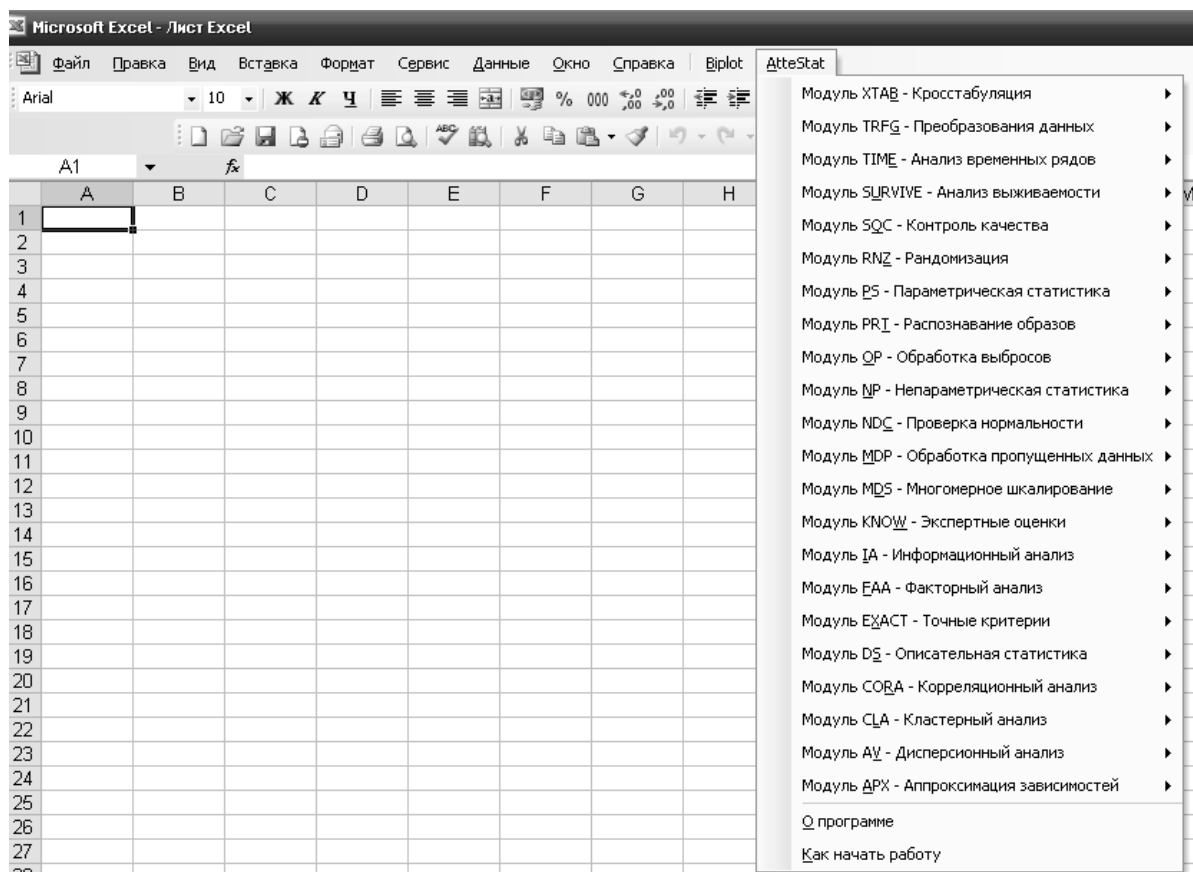


Рис. 1.6. Набор статистических модулей программы ATTESTAT

Программу ATTESTAT выгодно отличает компактное и легко интерпретируемое представление результатов анализа, широкий набор методов обработки данных (в некоторых случаях этот набор значительно превышает таковой в популярных зарубежных программах), удобный русскоязычный интерфейс, наличие справочной системы по статистическим методам анализа и пошаговой инструкции работы с программой.

1.3. Несколько слов о терминологии

Введение данного раздела обусловлено необходимостью определить те основные термины и понятия, которыми принято оперировать при статистической обработке биологических и экологических данных. В биометрии до сих пор нет единой терминологической системы, и в работах разных авторов используются различные термины и разные символы для обозначения одних и тех же статистических показателей. В данном пособии зачастую будет даваться несколько терминов для описания одного и того же статистического понятия, а также их

англоязычный аналог. Это поможет студентам в работе с нерусифицированными компьютерными программами. Теперь же перейдем к некоторым базовым понятиям, без которых дальнейшее понимание материала может оказаться затруднительным.

Статистическая совокупность – множество относительно однородных, но индивидуально различимых единиц, объединенных для группового изучения (Лакин, 1990); всякое множество отдельных отличающихся друг от друга и в то же время сходных в некоторых существенных отношениях объектов, сумма наблюдений и измерений (Рокицкий, 1973).

Примеры: популяции животных, виды в экосистеме, стадо коров, растения на опытных делянках, данные о численности и биомассе видов за определенный промежуток времени, база данных по концентрациям загрязняющих веществ в водных объектах.

Единица наблюдения (совокупности) – элемент статистической совокупности (Лакин, 1990).

Примеры: отдельная особь популяции, определенный вид в экосистеме, каждая корова в стаде, каждое отдельное растение на опытной делянке, отдельное значение численности вида или концентрации загрязняющего вещества.

Объем совокупности – число единиц наблюдения, входящих в статистическую совокупность (Рокицкий, 1973).

Примеры: число коров в стаде, число растений, отобранных для эксперимента, количество проб, в которых измерена концентрация загрязняющего вещества за определенный промежуток времени, количество животных, отловленных для изучения.

Признак – свойство, проявлением которого один объект отличается от другого (Лакин, 1990); характеристика объекта исследования (Реброва, 2002); любая информация о наблюдаемом объекте, выраженная качественно или количественно определенная (Ивантер, Коросов, 2003).

Примеры: цвет глаз, количество зерен в колосе, длина и вес особи, число видов в разных сообществах, процент жира в молоке определенной коровы, концентрация хлорофилла «а» в листьях растений.

В качестве синонимов в пособии используются такие термины, как **показатель, характеристика, величина**, употребление которых более приемлемо в отношении экологических объектов. Кроме того, поскольку признаки могут принимать

различные значения, то их принято называть переменными величинами или просто **переменными** (variables). Каждая единица наблюдения, таким образом, может характеризоваться определенными признаками, однако встречаются случаи, когда понятия «единица наблюдения» и «признак» (показатель, характеристика, величина переменная) совпадают.

Примеры: такие показатели, как значения численности вида или концентрации загрязняющего вещества, полученные исследователем за некий промежуток времени, одновременно будут являться и единицами наблюдения, составляя статистическую совокупность.

Варианта – отдельное числовое значение признака (Рокицкий, 1973; Лакин, 1990).

1.4. Характер биологических и экологических данных

Характерным свойством признаков биологических и экологических объектов является варьирование их значений в определенных пределах при переходе от одной единицы наблюдения к другой. Подобная особенность в статистике обозначается терминами **вариация, дисперсия, вариабельность, рассеяние вариантов, разброс, изменчивость.**

Примеры: особи одной популяции различаются по массе и размерам тела; в листьях растений одного вида или даже в разных листьях одного растения может содержаться различное количество хлорофилла «а»; изменчивость окраски у животных одного вида; изменение плотности популяции, урожайности сортов, плодовитости организмов во времени и пространстве.

Подобная вариабельность обусловлена влиянием на изучаемые объекты многочисленных причин (генетические различия особей, влияние факторов среды). Кроме того, на величине признаков сказываются погрешности (ошибки) измерений – разница между результатами измерения и действительно существующими значениями измеряемой величины (Лакин, 1990). Погрешности измерений принято делить на 2 группы (Лакин, 1990):

1. *Систематические* (смещение оценок) – неслучайные, постоянно повторяющиеся. В том числе:

1.1. *Технические*. Возникают из-за неточности измерительных приборов и инструментов (прибор не откалиброван или имеет техническую неисправность, использование термометра, дающего ошибку на полградуса в положительную сторону) (Лакин, 1990).

1.2. *Личные*. Возникают из-за личных качеств исследователя, его навыков и мастерства в работе (систематическое применение одной и той же ошибочной формулы расчета концентрации вещества, постоянное недоливание воды до метки при приготовлении растворов) (Лакин, 1990).

2. *Случайные* (неточность оценок) – возникают от целого ряда других, не поддающихся регулированию и неустранимых причин (лаборант при снятии показаний перепутал значение на шкале прибора) (Лакин, 1990).

Систематические ошибки можно свести до минимума, постоянно проверяя точность измерений прибора, совершенствуя технические средства, повышая квалификацию. Случайные ошибки, как независимые от воли исследователя, остаются и сказываются на результатах исследований, обуславливая определенную долю вариации признаков биологических и экологических объектов.

Биологические и экологические данные, пригодные для математической обработки, могут быть представлены в различной форме. В математической статистике различают три типа признаков (Урбах, 1964; Рокицкий, 1973; Терентьев, Ростова, 1977; Лакин, 1990; Реброва, 2002; Ивантер, Коросов, 2003, 2005):

1. Количественные признаки являются числовыми, могут быть упорядочены, и для них имеют смысл различные вычисления, например средних величин и показателей вариации. Количественные признаки делятся на *счетные* (число ветвистых лучей в спинном плавнике рыбы, число зерен в колосьях, яйценоскость, плотность популяции) – варьируют прерывисто (дискретно), их числовые значения выражаются только целыми числами – и *мерные* (размер, вес особи, температура тела, артериальное давление) – варьируют непрерывно, их величина может принимать в определенных пределах (от – до) любые числовые значения.

2. Качественные признаки (пол животного, вид растения, цвет глаз) являются нечисловыми, они означают принадлежность к некоторым классам и не могут быть упорядочены или непосредственно использованы в вычислениях. Значения таких признаков, как правило, выражаются словами, символами, зна-

ками (♀, ♂, бурый, синий, зеленый, +, -). Простейший способ перевода качественных данных в количественные – это подсчет числа единиц наблюдения (частота встречаемости), у которых отмечается тот или иной качественный признак. В дальнейшем это дает возможность обрабатывать подобные признаки с помощью количественных статистических методов.

3. Порядковые признаки занимают промежуточное положение: их значения упорядочены (стадия развития животного или растения, уровень эвтрофикации водоема, зоны загрязнения вокруг промышленного предприятия), но не могут быть с уверенностью измерены и сопоставлены количественно. Они в большей или меньшей степени обладают качеством, выраженным данной переменной. Однако они не позволяют сказать, «на сколько больше» или «на сколько меньше». Так, ранжируя зоны загрязнения в порядке возрастания степени этого загрязнения, как 1, 2, 3 ...10, можно с уверенностью утверждать, что 10-я зона загрязнения сильнее, чем 5-я зона, но вовсе не в два раза.

Примечание. Для различных типов переменных применяются разные методы статистического анализа!!! При планировании исследований важно понимать, что порядковые или качественные данные можно статистически исследовать только с помощью непараметрических приемов, тогда как для количественных признаков можно использовать, кроме того, точные и высокоэффективные параметрические методы. В целом, возможности статистической обработки порядковых и качественных данных значительно ограничены.

1.5. Выборочный метод исследования

Основной методологией подавляющего большинства биологических и экологических исследований является количественное обследование лишь определенной части элементов наблюдения выбранного объекта изучения. Что представляет собой эта часть элементов наблюдения, как правильно её формировать и по каким причинам биологи и экологи вынуждены прибегать к подобному методу исследования, постараемся выяснить в данном разделе.

Генеральная совокупность и выборка. Если исследование охватывает все единицы наблюдения статистической совокупности без единого исключения, то оно называется *сплошным или*

полным (изучение всех особей биологической популяции, учет всех видов растений и животных в экосистеме). Если ограничиваются обследованием лишь некоторой части статистической совокупности, то исследование называется *частичным или выборочным*. В соответствии с этим в математической статистике принято делить статистическую совокупность на генеральную и выборочную. Поскольку эти понятия являются ключевой идеей математической статистики, рассмотрим их подробнее. Ниже приводится ряд определений.

1. «Совокупность, из которой отбирают определенную часть её элементов для совместного изучения, называется *генеральной совокупностью*. Отобранная часть генеральной совокупности для изучения называется *выборочной совокупностью или выборкой*» (Лакин, 1990).

2. «В подавляющем большинстве случаев исследование ведется выборочным методом, т. е. из всего количества существующих в природе объектов или из теоретически возможного бесконечного множества опытов (*генеральная совокупность*) учитывается лишь какая-то часть (*выборка*)» (Терентьев, Ростова, 1977).

3. «*Генеральная совокупность* – это вся подлежащая изучению совокупность данных объектов. В пределе она рассматривается как состоящая из бесконечно большого количества отдельных единиц. Та часть объектов, которая подвергается исследованию, называется *выборочной совокупностью* или просто *выборкой*» (Рокицкий, 1973).

4. «Всю совокупность интересующих нас объектов исследования мы будем называть *генеральной совокупностью*, а малую, детально изучаемую её часть – *выборочной*» (Поморский, 1935).

5. «Множество объектов, конечное или бесконечное, относительно которого делаются статистические выводы, носит название *генеральной совокупности*. Этот термин приобретает смысл в сочетании с понятием о *выборке*, т. е. части этого множества» (Владимирский, 1983).

6. «...обычно изучается лишь часть популяции, которую принято называть *выборкой* из *генеральной совокупности* – совокупности всех экземпляров, или особей, или членов данной совокупности, которые вообще в принципе могут относиться к этой совокупности. ...в экспериментальной биологии почти всегда имеют дело с *выборками* – *генеральной совокупностью*

здесь обычно является бесконечное множество однотипных экспериментов, которые в принципе можно было бы провести» (Урбах, 1964).

7. «Термин "*выборка*" указывает на процесс выбора части из чего-то большего, в данном случае – на процесс получения ограниченного количества значений из генеральной совокупности. *Генеральная совокупность* – это множество всех вариантов определенного типа (выборка бесконечного размера). Чаще всего получить все возможные значения в принципе невозможно. Поэтому судить о генеральной совокупности приходится, исследуя выборки – по части составлять представление о целом» (Ивантер, Коросов, 2005).

8. «*Выборочной совокупностью* или просто *выборкой* называют совокупность случайно отобранных объектов. *Генеральной совокупностью* называют совокупность объектов, из которых производится выборка» (Гмурман, 2001).

Анализируя приведенные определения, нужно выделить ряд важных моментов:

1. Объем выборки (обозначается буквой n) не может превышать объем генеральной совокупности (обозначается буквой N).

2. Объем генеральной совокупности часто представляется теоретически бесконечным, но на практике он имеет конечные размеры. Таким образом, n и N могут значительно различаться в зависимости от целей и объектов исследования.

Пример: в качестве генеральной совокупности можно рассматривать всех особей изучаемого вида или особей этого же вида, обитающих на конкретной территории (географическая популяция), в этом случае объем генеральной совокупности будет меньше; в генеральную совокупность могут входить несколько мелководных озер региона, или одно мелководное озеро может рассматриваться как генеральная совокупность.

3. Совокупность единиц наблюдения в выборку должна отбираться определенным образом, а именно случайно.

4. Любой исследователь стремится охарактеризовать объект изучения в целом, поэтому выборка сама по себе не должна представлять интереса для исследователя, она служит для оценки генеральной совокупности, из которой извлечена.

Закономерно возникает вопрос: зачем биологу (экологу) извлекать выборки из генеральной совокупности, если необхо-

можно охарактеризовать объект изучения в целом? Для этого более целесообразным будет полное исследование генеральной совокупности, что позволит получить исчерпывающую информацию об объекте изучения. Тем более что при этом будут отсутствовать ошибки (о них мы поговорим в главе 4), неизбежно возникающие в процессе отбора выборок из генеральной совокупности.

Дело в том, что сплошные исследования в биологии и экологии, такие как, к примеру, оценка рыбопродуктивности путем тотального вылова рыбы, имеют ряд существенных недостатков и ограничений, которые преобразуются в преимущества при проведении выборочного исследования.

Причины применения выборочного метода исследования:

1. Экономия времени, материальных и кадровых ресурсов при проведении исследования, поскольку изучается лишь часть генеральной совокупности.

2. Возможность изучать объекты, сплошное обследование которых практически невозможно или нецелесообразно.

Пример: невозможен полный учет бактерио-, фито- или зоопланктона даже небольшого водоема, или фактически нереально определить все виды растений, животных и микроорганизмов в экосистеме, невыполнима задача отлова для изучения всех особей из популяции какого-либо вида насекомого, нецелесообразно высеивать всю партию семян, чтобы определить их всхожесть, или не имеет смысла отбирать 100 проб в одном и том же створе, чтобы оценить экологическое состояние участка реки.

По этим причинам биологи и экологи практически всегда вынуждены иметь дело с выборками, при этом от того, каким образом была взята выборка из генеральной совокупности, будет зависеть конечный результат исследования.

Пример: может ли исследователь получить правильное представление о состоянии планктона озера в целом, анализируя пробы воды, отобранные на одной станции? Если исследователь для изучения отбирает растения только в центре луга и не отбирает их по краям, можно ли в итоге результаты подобного исследования переносить на всю луговую экосистему? Часто на физиологических кафедрах университетов студенты, занимаясь научной работой, ведут отбор испытуемых на биологических факультетах. При этом выборки создаются из знакомых, сокурсников и друзей. Могут ли такие выборки корректно отражать свойства генеральной совокупности (все студенты биологического факультета)?

Математическая статистика на все поставленные вопросы отвечает отрицательно. Для того чтобы лишь по части генеральной совокупности, которая изучена, можно было правильно судить о всей генеральной совокупности, выборка должна быть *репрезентативной*, иначе представительной. Репрезентативность выборки означает равную вероятность для всех единиц наблюдения генеральной совокупности быть представленными в составе выборки, другими словами, в выборке должны быть представлены все возможные варианты изучаемой переменной в тех же пропорциях, что и в генеральной совокупности. Во всех приведенных примерах данное требование, очевидно, не было выполнено.

Каким образом можно достичь равной возможности для всех единиц наблюдения попасть в выборку? Для этого при планировании исследований необходимо соблюдать принцип *рандомизации* (от англ. random – случай) – случайный отбор элементов из генеральной совокупности, исключающий систематические ошибки. Рассмотрим наиболее часто употребляемые в биологии и экологии способы формирования выборочных совокупностей (Бейли, 1962; Урбах, 1964; Василевич, 1969; Плохинский, 1970; Апостолов, Ивашов, 1981; Шмидт, 1984; Лакин, 1990).

Способы отбора выборок из генеральной совокупности

I. *Повторный отбор* – производят по схеме возвращения учтенных единиц в генеральную совокупность, так что одна и та же единица может попасть в выборку повторно (отлов окольцованных птиц, суточный количественный учет животных на пробной площадке, повторное использование объектов для экспериментов). Подобный отбор не влияет на состав генеральной совокупности, и возможность каждой единицы попасть в выборку не меняется.

II. *Бесповторный отбор* – учтенные единицы не возвращаются в генеральную совокупность, каждая отобранная единица регистрируется только один раз (отлов животных для изучения питания, отбор почвенных или водных проб). Этот отбор влияет на состав генеральной совокупности и возможность каждой единицы попасть в выборку меняется.

Оба способа отбора (повторный и бесповторный) делятся на 2 типа:

1. Отбор, не требующий расчленения генеральной совокупности на части, – *простой случайный отбор* – элементы извлека-

ются случайным образом непосредственно из генеральной совокупности (отбор животных для эксперимента, отлов животных из популяции, выборка из рабочих вредного производства). Человек весьма несовершенное «орудие» случайного отбора. В психике каждого из нас, даже при кажущейся беспристрастности, заложено подсознательное предпочтение определенного облика объекта, а значит и стремление к тенденциозному субъективному отбору элементов наблюдения. Независимо от наших попыток соблюдать максимальное благоразумие и честность при формировании выборок всегда имеется определенная вероятность того, что появится невольная систематическая ошибка. На практике для осуществления случайного отбора применяют метод случайных чисел: для этого можно воспользоваться либо таблицами случайных чисел, либо соответствующими модулями (процедурами) в статистических программах.

Пример: для проведения эксперимента из 100 подопытных мышей необходимо отобрать 10 особей. Конечно, исследователь может отобрать тех животных, которые первыми выбегут из клетки после открывания дверцы. Но в этом случае он должен будет понимать, что его отбор является довольно субъективным и не лишен скрытых систематических ошибок. В результате этого выборка может неадекватно отражать свойства генеральной совокупности или опытная и контрольная группы животных могут оказаться изначально неоднородными, что будет влиять на результаты эксперимента. Чтобы произвести действительно простой случайный отбор, необходимо создать равную вероятность для всех мышей быть включенными в выборку. Для этого исследователь предварительно может пронумеровать (от 1 до 100) всех животных (генеральная совокупность) и для отбора воспользоваться таблицей случайных чисел (табл. 1.2).

Таблица 1.2

Случайные числа (по: Лакин, 1990)

3393	6270	4228	6069	9407	1865	8549	3217	2351	8410
9108	2330	2157	7416	0388	6173	1703	8132	9065	6717
7891	3590	2502	5945	3402	0491	4328	2365	6175	7695
9085	6307	6910	9174	1753	1797	9229	3422	9861	8357
2638	2908	6368	0398	5495	3283	0031	5955	6544	3883

Случайные числа – это последовательность чисел, выбранных из некоторой генеральной совокупности чисел при помощи какого-нибудь случайного процесса (жеребьевка). Из таблицы необходимо отобрать 10 чисел, не превышающих значение 100 (т. к. в нашем случае $N = 100$). Просматривание таблицы можно начинать в любом месте и вести в произвольном направлении. Допустим, мы начнем с первого столбца таблицы и будем двигаться последовательно сверху вниз, учитывая 2 последних цифры четырехзначных чисел. В итоге мышей под номерами 93, 8, 91, 85, 38, 70, 30, 90, 7, 28 необходимо будет отобрать для эксперимента.

Более удобный способ составления случайных выборок исследователь может найти, открыв табличный процессор MS EXCEL: в меню **Сервис** нужно выделить строку **Анализ данных**, найти процедуру **Выборка** и щелкнуть на кнопку **ОК** (рис. 1.7).

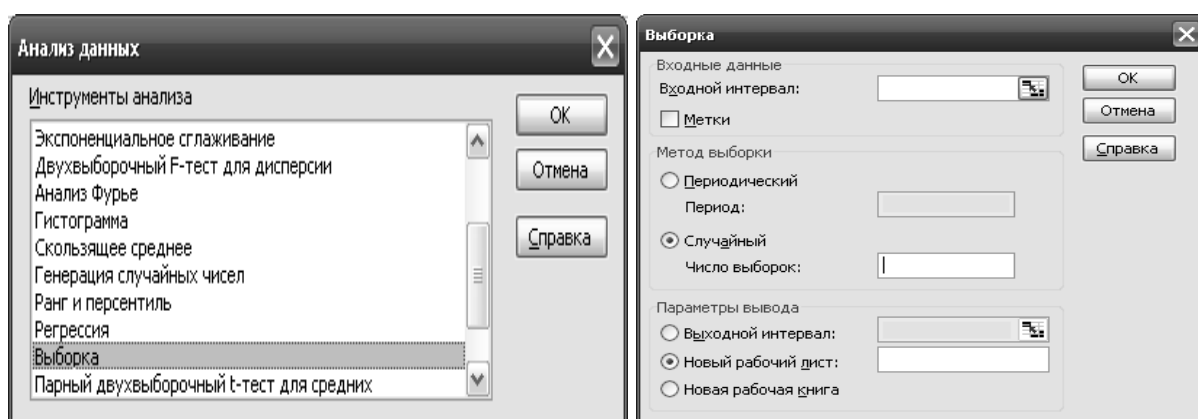


Рис. 1.7. Общий вид меню пакета «Анализ данных» и диалоговое окно процедуры «Выборка»

В диалоговом окне процедуры **Выборка** необходимо лишь установить **Случайный** метод выборки, мышкой указать **Входной интервал** (столбец электронной таблицы с порядковыми номерами животных) и ввести объем выборки в поле **Число выборок** (в нашем примере 10). Необходимо отметить, что данная процедура реализует повторную случайную выборку, поэтому заранее полезно указывать немного больший объем выборки, чем требуется отобрать для исключения повторяющихся объектов.

Не менее удобным для целей формирования выборок является модуль Рандомизация статистической программы ATTESTAT (рис. 1.8).

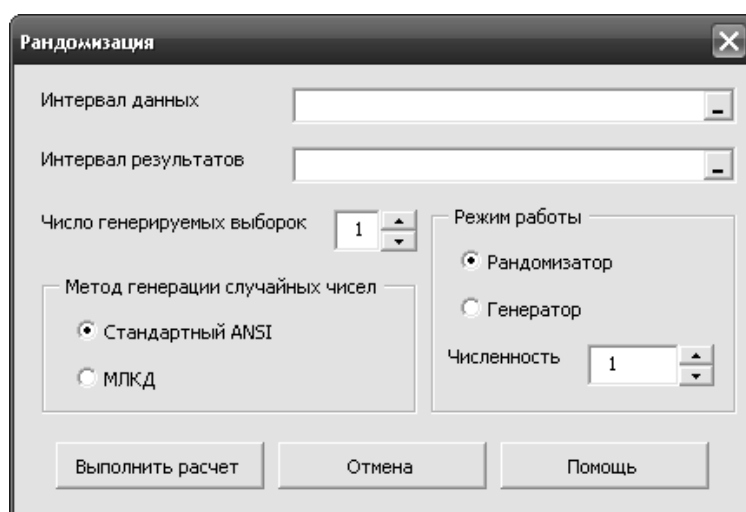


Рис. 1.8. Диалоговое окно модуля «Рандомизация»

К примеру, с помощью данного модуля можно случайным образом распределять объекты в контрольные и опытные группы.

Простой случайный отбор имеет преимущества в тех случаях, когда объем генеральной совокупности не слишком велик. Но каким образом можно создать равную вероятность попадания в выборку единиц наблюдения, если исследователь имеет дело, например, с насекомыми или планктонными животными или его целью является экологическое исследование обширного географического района? Предварительно отловить и пронумеровать всех особей для того, чтобы произвести простой случайный отбор, явно не получится. В этом случае можно использовать отбор второго типа.

2. Отбор, при котором генеральная совокупность разбивается на части:

А. *Серийный отбор* – генеральную совокупность предварительно делят на классы (серии), затем из общего количества серий случайным способом отбирают несколько серий для сплошного изучения. Подобный вариант отбора следует применять при работе с относительно однородными объектами исследования (агроэкосистемы, группы испытуемых одинакового возраста, небольшие и простые по морфологии водоемы).

Пример: необходимо составить выборку из жуков люцернового поля. Предварительно нужно разбить всю площадь поля на небольшие пробные площадки, присвоить им порядковые номера и с помощью уже известных нам способов случайным образом отобрать ряд пробных площадок, на которых произвести сплошное обследование по учету жуков.

Б. Типический отбор – генеральная совокупность делится на несколько классов (типических групп), а затем случайным образом делается выборка из каждой отдельной типической группы (т. е. в отличие от серийного отбора сплошного изучения каждой типической группы не производится). Используется этот способ с успехом в тех случаях, когда исследуемые объекты неравномерно распределены в определенном объеме или на определенной территории, что и встречается наиболее часто в природных условиях.

Пример: необходимо установить размерно-возрастную структуру популяции ящерицы живородящей на определенной территории, включающей различного типа лесные и луговые участки, вырубки, дороги и т. д. Для получения репрезентативной выборки всю территорию нужно поделить на ряд типических групп (участки леса, луг, опушки, вырубка, обочины дорог), в наибольшей степени различающихся между собой. Затем каждую типическую группу можно разбить на ряд пробных площадок, случайным способом выбрать в каждой группе несколько таких площадок, на которых и произвести отлов животных. Таким образом, общий объем выборки будет включать особи из каждой типической группы, что обеспечит репрезентативное описание всей генеральной совокупности.

В. Механический отбор – генеральная совокупность «механически» делится на столько групп, сколько объектов должно войти в выборку, а из каждой группы отбирается один объект.

Пример: при обследовании посева ржи на урожайность намечено отобрать 100 колосьев. Следовательно, поле ржи можно разбить на 100 равных делянок и с каждой случайным образом отобрать по 1 растению; при изучении питания можно отлавливать каждый пятый, десятый и т. п. экземпляр животного данного вида на маршруте. В последнем случае, однако, нужно следить за тем, чтобы алгоритм составления выборки не совпадал с каким-либо периодическим процессом в природе, способным повлиять на репрезентативность выборки.

Примечание 1. Несмотря на то, что методы математической статистики построены на предположении случайности формирования выборочной совокупности, на практике не исключены ситуации, когда именно в силу случайности в выборку могут попасть экземпляры с преимущественно крайними вариантами признаков (нерепрезентативная выборка) (Шмидт, 1984). Исследователь должен понимать, что вероятность подобного будет тем выше, чем малочисленнее формируемая выборка.

Примечание 2. Следует признать, что описанные в данном разделе методы выборочного исследования в практике биологических и экологических работ применяются далеко не всегда. Причинами этого могут быть неосведомленность исследователей, дань традициям, иногда трудности технической реализации того или иного способа случайного составления выборки, недостаток времени и т. д. В конечном итоге изучается совсем не то, что декларируется в названиях статей на страницах рецензируемых журналов. К примеру, часто исследователи приводят результаты изучения водных экосистем, при этом имея в наличии данные лишь с одной, так называемой «стандартной», станции (точки) отбора проб. Теория выборочного исследования редко используется при изучении бентоса с его крайне неравномерным распределением в донных биотопах (Баканов и др., 2001). Многие методы количественного учета животных и растений на практике применяют, игнорируя описанные в разделе принципы. При этом данные, полученные по нерепрезентативной выборке, никаким образом нельзя будет «исправить» или «подправить» в процессе их статистической обработки. Какие бы мощные методы математической обработки ни применялись, получить адекватное представление о генеральной совокупности не удастся.

Глава 2. Приемы первичной статистической обработки данных

Объектами исследований биологов и экологов могут быть системы различного уровня (клетка, орган, организм, популяция, биоценоз, экосистема) и разнообразные биологические (экологические) процессы и явления (размножение, питание, динамика численности популяций, сукцессия экосистем). Для изучения этих объектов необходимо получить, обработать и проанализировать соответствующие данные. *Данные* – это исходная информация об объекте исследования, полученная путем наблюдения или эксперимента и представленная в форме, пригодной для постоянного хранения, передачи, обработки и анализа (например, набор конкретных чисел).

В биологических и экологических исследованиях принято регистрировать первичные данные в специальных журналах, дневниках, бланках, ведомостях.

Пример:

1. Учетная (маршрутная) ведомость встречаемости птиц.
2. Бланки обработки гидробиологических проб.
3. Ихтиологический журнал траловой съемки озера.
4. Лабораторный журнал обработки проб.

Зафиксированные в подобных документах учета сведения об изучаемом объекте представляют собой беспорядочную массу фактического материала, выраженного, как правило, в числовой, балльной, текстовой, знаковой формах. В современных условиях внедрения в научные исследования компьютерных технологий следующим обязательным этапом является ввод этих данных в одну из программ статистического анализа, к примеру в электронную таблицу MS EXCEL или пакет STATISTICA. Форма организации первичных данных в электронных таблицах, естественно, будет различаться в зависимости от цели статистической обработки.

Пример: для вычисления описательных статистик необходим ввод всех числовых значений исследуемого показателя в один столбец электронной таблицы, а для проведения дисперсионного анализа требуется ввод в соседний столбец некой группирующей переменной или разбивка этих значений по нескольким столбцам.

Фактически уже на этом этапе исследователь приступает к простейшей статистической обработке собранных данных для выявления скрытых в первичной информации закономерностей. Овладение приемами статистической обработки первичных данных необходимо в первую очередь для последующего освоения более сложных статистических методов, поскольку эти приемы часто являются важными этапами к проведению дальнейшего количественного анализа.

Пример: построение статистических рядов является первым этапом осуществления анализа временных рядов, корреляционного и регрессионного анализов; графическое представление данных – неотъемлемая часть кластерного анализа; расчет ряда описательных статистик (среднего значения или дисперсии) необходим при проверке статистических гипотез.

Статистическая обработка первичных данных имеет и самостоятельное значение. Так, построение вариационных рядов и кривых может дать исследователю ценную информацию о законе распределения изучаемого признака или показателя, в дальнейшем это может помочь как в выборе корректных методов математической обработки, так и в определении факторов, вызывающих подобное вариационное распределение; вычисление средних значений и показателей вариации само по себе является важной характеристикой объекта исследований.

Методы, рассматриваемые в данной главе, относятся, пожалуй, к наиболее популярным и часто используемым формам статистической обработки данных и в научных работах студентов вузов, и в подавляющем большинстве статей, публикуемых в ведущих научных журналах. Такие элементарные статистические процедуры, как расчет среднего значения, показателей вариации, построение статистических графиков и таблиц, приходится применять биологам и экологам, вероятно, в 95% случаев. Проще найти множество работ, где не применяются необходимые относительно сложные методы статистической обработки данных (например, корреляционный или дисперсионный анализы), чем работы, в которых не представлены расчеты средних значений исследуемых показателей или графический анализ.

2.1. Статистические ряды

Математическая обработка собранных данных часто (но далеко не всегда!) начинается с построения так называемых *статистических рядов*, представляющих собой набор числовых значений признака, расположенных в определенном порядке.

Рассмотрим более подробно типы статистических рядов.

1. Ранжированный ряд – одинарный ряд, в котором значения признака располагаются в возрастающем (или убывающем) порядке.

Пример:

34342543345

23333444455 – ранжированный ряд.

Значение ряда: можно определить размах изменчивости признака (от 2 до 5), наиболее часто встречающееся значение (3 и 4), подготовительный этап для построения вариационного ряда.

2. Вариационный ряд (ряд распределения) – двойной ряд чисел, отражающий соотношение ранжированных значений признака с частотой их встречаемости в данной выборке.

Пример:

23333444455 – ранжированный ряд,

2345 – значение признака,

1442 – частота встречаемости.

Значение ряда: позволяет определить закономерность варьирования (закон распределения) изучаемого признака.

В зависимости от того, в каком диапазоне и как варьирует признак – дискретно или непрерывно, – статистическая совокупность может распределяться в *безынтервальный* или *интервальный* вариационные ряды. Тип вариационного ряда можно определить по формуле (Лакин, 1990):

$$\lambda = x_{\max} - x_{\min} / k,$$

где λ – ширина классового интервала,

x_{\max} ; x_{\min} – максимальное и минимальное значение выборки,

k – число классов, на которые следует разбить вариацию признака, рассчитывается по формуле Стерджеса:

$$k = 1 + 3.32 \cdot \lg(n), \text{ где } n \text{ – объем выборки.}$$

Таким образом, если $\lambda = 1$ или $\lambda \approx 1$, то строится безынтервальный ряд, если $\lambda \neq 1$, то строится интервальный ряд.

Если признак варьирует дискретно и в узких границах ($\lambda = 1$ или $\lambda \approx 1$), то строится безынтервальный вариационный ряд. Рассмотрим данные о количестве птенцов в гнездах древесной ласточки *Tachycineta bicolor* (Рокицкий, 1973):

4 6 6 4 5 5 5 5 5 5 1 4 5 4 5 4 5 5 7 4 6 6 5 6 4 4 5 6 5 5 4 2 6 4 6 2 5 6 5 5 4

Данный признак является дискретным и $\lambda \approx 1$, значит достаточно подсчитать встречаемость конкретных значений, не разбивая их на классовые интервалы. Искомый безынтервальный вариационный ряд будет выглядеть следующим образом:

<i>Количество птенцов</i>	<i>Частота встречаемости</i>
1	1
2	2
4	11
5	18
6	9
7	1

Интервальный вариационный ряд применяется, если изучаемый признак изменяется непрерывно ($\lambda \neq 1$) или значения дискретного признака, варьирующего в широких пределах, имеют малую повторяемость. В воде мелководного озера Неро (Ярославская область) в течение года были измерены концентрации общего фосфора (в мкг/л):

46	41	153	98	140	95	208	88	65	108
60	41	179	320	176	118	191	108	62	91
90	66	189	274	170	95	62	108	45	58
90	83	202	134	166	82	117	62	91	37
80	45	111	83	120	108	91	241	90	66
163	110	117	91	180	104	91	134	92	83

Для построения интервального вариационного ряда сначала весь диапазон изменчивости концентраций общего фосфора разбивается на серию равных классовых интервалов, затем подсчитывается, сколько вариантов попало в каждый интервал. В нашем

примере ширина классового интервала $\lambda = 41$, число классовых интервалов $k = 7$, соответственно вариационный ряд имеет вид:

<i>Классовые интервалы концентраций (мкг/л)</i>	<i>Частота встречаемости</i>
37–78	14
78.1–119	28
119.1–160	5
160.1–201	8
201.1–242	3
242.1–283	1
283.1–324	1

3. Временной ряд (ряд динамики) – двойной ряд чисел, отражающий варьирование вариант изучаемого признака во времени (по годам, месяцам, дням, часам).

Пример: сезонные изменения биомассы фитопланктона в озере можно охарактеризовать следующим временным рядом

2 11 6 1 20 30 10 2 – биомасса фитопланктона (мг/л),
 III IV V VI VII VIII IX X – месяцы.

4. Эмпирический ряд регрессии – двойной ряд чисел, отражающий связь между значениями сопряженных признаков.

Пример: в 2011 г. в районе биостанции «Улейма» студентами ЯрГУ были получены следующие данные о численности насекомых-опылителей на пробной площадке (X) и температуре воздуха в периоды учета насекомых (Y):

X: 17 19 59 114 94 78 78 64 78 48 35 36 5 5 11
 Y: 17 16.8 23.8 25.6 27 24.7 21.8 22.7 23.1 21.8 20.3 19 15 14.5 18.8

2.2. Графический анализ

Визуализация, или наглядное представление, результатов исследований является важным этапом при первичной математической обработке данных. Графическое осмысление фактов входит почти в каждую научную работу, и к нему следует прибегать, где только возможно и целесообразно. Построение графиков различных типов упрощает содержательный анализ количественных данных и во многих случаях является эффективным

средством контроля возможных ошибок при интерпретации результатов, полученных тем или иным статистическим методом.

В данном разделе мы начнем знакомство с возможностями графического анализа при математической обработке биологических и экологических материалов на примере наглядной иллюстрации закономерностей, заключенных в статистических рядах. Изложение иных способов визуализации результатов количественного анализа будет продолжено в последующих главах.

Графическое представление закономерностей варьирования количественных признаков осуществляется с помощью *вариационных кривых* (*полигон распределения частот*) (рис. 2.1 а) и *гистограмм распределения* (частот встречаемости значений признака) (рис. 2.1 б).

Вариационные кривые строятся для безынтервальных вариационных рядов в осях: значения признака (абсцисса) – частота встречаемости значений признака (ордината). Данный график представляют собой ряд точек, соединенных прямыми линиями, при этом каждая точка отражает частоту встречаемости конкретного значения дискретного признака. Анализ вариационной кривой на рис. 2.1 а обнаруживает характерную закономерность поведения количественного признака – число птенцов в гнездах древесной ласточки: высокие частоты встречаемости вариант наблюдаются в центре распределения, а низкие по периферии.

Весьма сходны с вариационными кривыми так называемые *гистограммы распределения частот* – столбчатые диаграммы, отражающие распределение частот встречаемости значений признака по отдельным классовым интервалам. Соответственно, в отличие от вариационной кривой на гистограмме распределения частот по оси абсцисс откладываются классовые интервалы. Подобные графики применяются для интервальных вариационных рядов. Возвращаясь к ранее описанному примеру, можно заключить, что закономерность варьирования концентраций общего фосфора значительно отличается от распределения количества птенцов в гнездах древесной ласточки: наблюдается смещение наиболее часто встречающихся концентраций фосфора в область меньших значений (рис. 2.1 б).

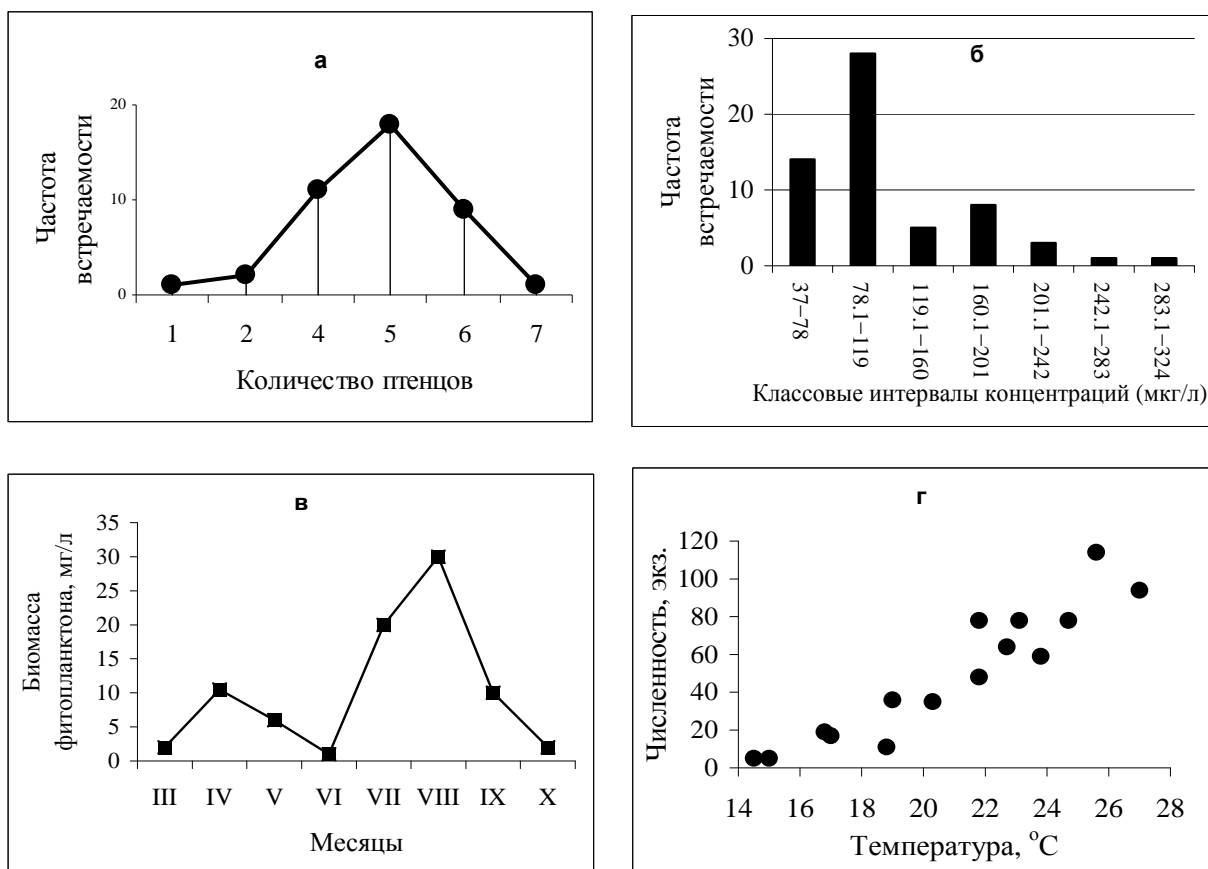


Рис. 2.1. Графическое представление закономерностей статистических рядов:
 а – вариационная кривая распределения количества птенцов в гнездах
 древесной ласточки *Tachycineta bicolor*; б – гистограмма распределения
 концентраций общего фосфора; в – сезонная динамика биомассы
 фитопланктона в озере; г – точечная диаграмма, отражающая связь
 температуры воздуха и численности насекомых-опылителей
 на пробной площадке

Табличный процессор MS EXCEL содержит процедуру автоматического построения из исходных данных одновременно вариационного ряда и гистограммы распределения частот этого ряда. Для этого в диалоговом окне **Анализ данных** надо выделить процедуру **Гистограмма** и нажать кнопку **ОК** (рис. 2.2). Для построения гистограммы распределения частот необходимо установить флажок **Вывод графика** (рис. 2.2).

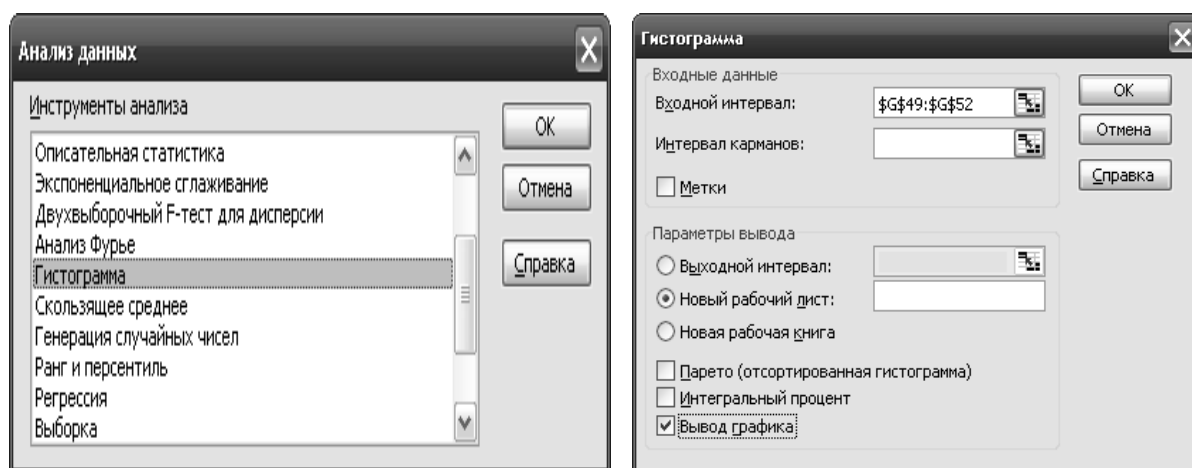


Рис. 2.2. Общий вид меню пакета «Анализ данных» и диалоговое окно процедуры «Гистограмма»

По данным рядов динамики строится график в осях: время (абсцисса) – значение признака (ордината) (рис. 2.1 в). Графический анализ сезонной динамики биомассы фитопланктона выявляет наличие весеннего и позднелетнего пика в обилии микроводорослей. Спад в развитии приходится на ранее лето, в гидробиологии этот период именуется стадией «чистой воды», что часто связано либо с биогенным лимитированием, либо с выеданием фитопланктона зоопланктоном.

На основе эмпирических рядов регрессии строится *точечная диаграмма* (*диаграмма рассеяния*), отражающая связь между парой признаков (показателей) (рис. 2.1 г). По оси абсцисс откладываются значения одного признака, по оси ординат – другого признака, сопряженного с первым. Таким образом, каждая точка на подобной диаграмме отражает значения пары признаков. Форма фигуры, создаваемой совокупностью точек на графике, является показателем связи двух признаков. Если между переменными существует сильная связь, то точки на графике образуют упорядоченную форму (например, близкую к прямой или кривой линии). Если переменные не связаны, то точки образуют «облако». Из рисунка 2.1 г видно, что точки образуют фигуру вытянутой формы, через которую в первом приближении можно провести прямую линию, при этом более высоким значениям температуры воздуха соответствуют более высокие численности насекомых-опылителей на пробной площадке. Это указывает на существование связи между двумя переменными.

Программное обеспечение графического анализа. Удобным средством проведения графического анализа является Мастер диаграмм в электронных таблицах MS EXCEL. В программном пакете STATISTICA предлагаются ещё более разнообразные графические методы, с помощью которых исследователь может запрашивать или самостоятельно организовывать построение графиков (рис. 2.3). Доступ к графическим средствам осуществляется через верхнее меню и команду Graphs (графики).

Так, программа STATISTICA дает возможность анализировать данные в трехмерном пространстве, для этого используются многообразные трехмерные графики (3D Graphs). Можно одновременно посмотреть, каким образом могут быть связаны между собой несколько переменных: к примеру, численность насекомых-опылителей на пробной площадке, температура воздуха и атмосферное давление. Это позволяет сделать трехмерная диаграмма рассеяния (3D XYZ Graphs), где каждая точка отображает значения 3-х переменных (рис. 2.4).

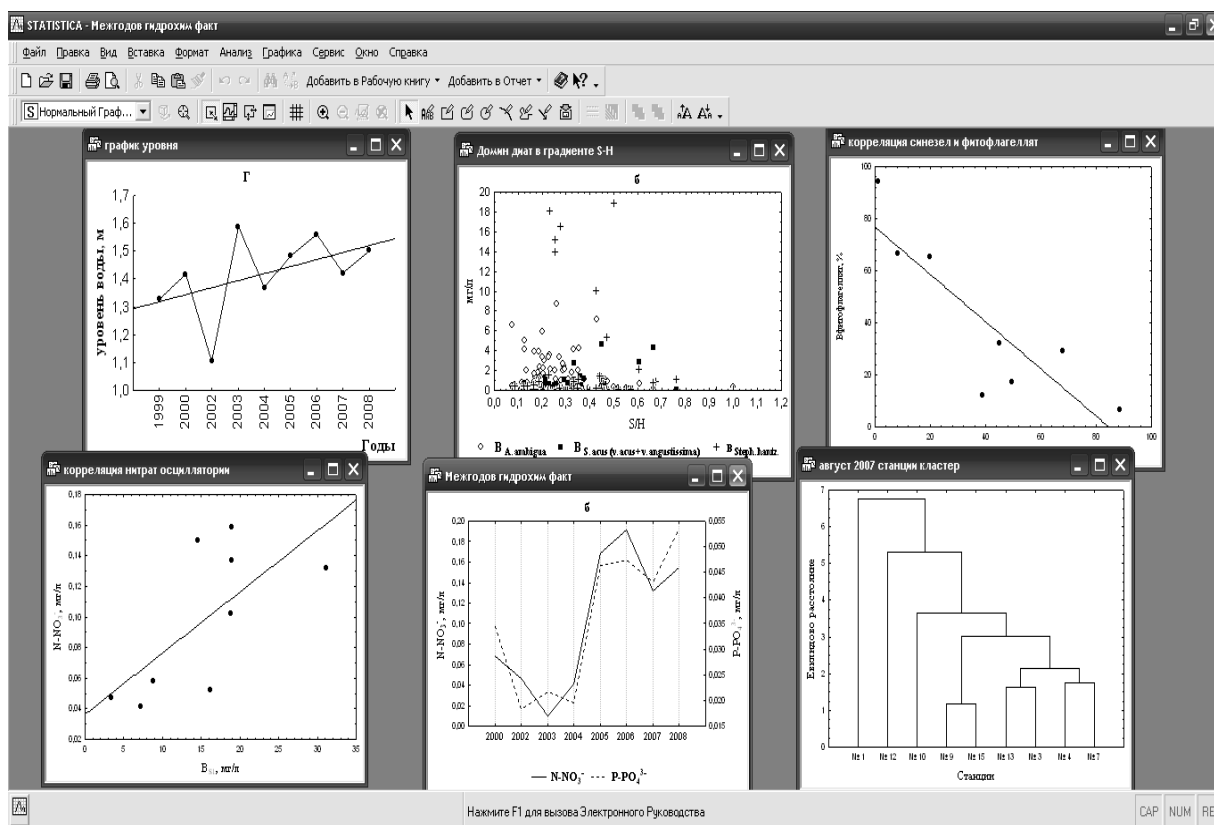


Рис. 2.3. Графическое представление данных в пакете STATISTICA

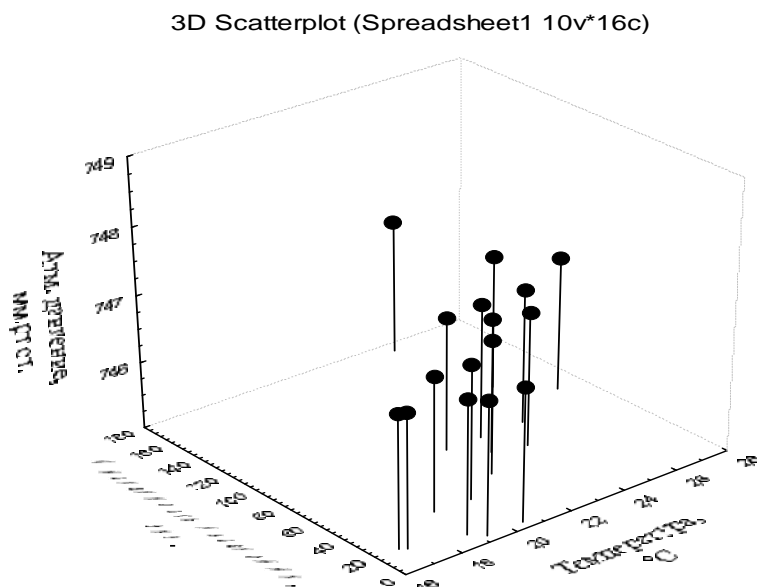


Рис. 2.4. Трехмерная диаграмма рассеяния, построенная в программе STATISTICA

2.3. Таблицы

Табличный способ представления данных является не менее важным, наряду с графическим анализом, средством первичного упорядочения, систематизации и группировки результатов исследований (Терентьев, Ростова, 1977; Лакин, 1990).

Статистические таблицы внешне представляют пересечения вертикальных граф и горизонтальных строк, которые образуют клетки, предназначенные для записи в них статистического материала. Статистическая таблица от других табличных форм отличается тем, что она должна содержать результаты подсчета (обработки) исходных количественных данных. Важным требованием является отсутствие перегруженности таблицы «излишней» числовой информацией, которая может затушевывать основную количественную закономерность. В задачи данного пособия не входит подробное изложение структуры, типов и правил оформления статистических таблиц, поэтому остановимся лишь на нескольких наглядных примерах.

Таблица 2.1

Спектр питания ящерицы живородящей в разных биотопах

Объект питания	Тип биотопа		Всего	
	Вырубка	Хвойный лес	шт.	%
Пауки	10	4	14	38
Двукрылые	9	2	11	30
Жуки	3	5	8	22
Саранчовые	3	-	3	8
Дождевые черви	1	-	1	2
Всего, шт.	26	11		

Таблица 2.2

Значение различных групп кормов в питании некоторых земноводных (по: Банников, Денисова, 1956)

Виды амфибий	% наземных кормов	% кормов, могущих быть добытыми и в воде, и на суше	% водных кормов
Серая жаба	100	0	0
Травяная лягушка	94.2	3.6	2.2
Прудовая лягушка	78.6	4.2	17.2
Озерная лягушка	67.5	9.1	23.4

Таблица 2.3

Биомасса бентоса ($г/м^2$) в озерах разных природных зон (по: Китаев, 1984)

Биомасса бентоса			
Тундра	Северная тайга	Средняя тайга	Смешанный лес
5.00	4.04	5.17	3.73
3.21	3.85	3.92	7.56
3.75	4.02	5.10	8.74
3.85	3.32	3.80	8.75
5.39	2.23	3.77	15.39
3.59	2.11	2.56	11.85
8.44	1.87	1.94	9.84

Таблица 2.4

**Биомасса бентоса ($г/м^2$) в зависимости от площади озера
и природной зоны (по: Китаев, 1984)**

Площадь озера	Биомасса бентоса				Среднее значение
	Тундра	Северная тайга	Средняя тайга	Смешанный лес	
<10 га	5.00	4.04	5.17	3.73	4.5
10–50 га	3.21	3.85	3.92	7.56	4.6
50–100 га	3.75	4.02	5.10	8.74	5.4
100–500 га	3.85	3.32	3.80	8.75	4.9
500–1000 га	5.39	2.23	3.77	15.39	6.7
1000–5000 га	3.59	2.11	2.56	11.85	5.0
>5000 га	8.44	1.87	1.94	9.84	5.5
Среднее значение	4.7	3.1	3.8	9.4	

Что показывают приведенные таблицы? В таблице 2.1 группировка содержимого желудков ящериц, обитающих в разных биотопах, позволяет выдвинуть на основе выборочных данных по крайней мере 4 гипотезы:

1. Главными пищевыми компонентами ящериц на исследованных участках являются пауки (38%), двукрылые (30%) и жуки (22%).

2. Главными компонентами питания ящериц на вырубке выступают пауки (10 шт.) и двукрылые (9 шт.), а в хвойном лесу – жуки (5 шт.) и пауки (4 шт.).

3. На вырубке рацион разнообразнее (5 объектов), чем в хвойном лесу (3 объекта).

4. Интенсивность питания ящериц в хвойном лесу (11 шт.) меньше, чем на вырубке (26 шт.).

Упорядочение количественных данных о группах кормов в питании некоторых земноводных (в относительных единицах) в табличной форме не только показывает их соотношение в спектре питания разных видов амфибий, но и позволяет сравнить рассматриваемых животных по образу жизни и особенностям их местообитания (табл. 2.2). Главная количественная закономерность в ряду «серая жаба – травяная лягушка – прудовая лягушка – озерная лягушка» заключается в уменьшении доли наземных кормов и одновременном увеличении доли

водных кормов в питании этих земноводных. Известно, что жабы по сравнению с другими амфибиями более устойчивы к засушливым условиям и, являясь типичной лесной формой, около водоемов концентрируются только на период икрометания. Этим можно объяснить отсутствие водных форм в питании серой жабы. Травяные лягушки более гигрофильны, однако способны удаляться от водоемов после периода размножения на значительные расстояния, проводя всё лето на суше. По-видимому, в связи с этим травяные лягушки очень редко кормятся водными формами. И, наконец, зеленые лягушки (прудовая и озерная) всю жизнь проводят в воде или около воды, поэтому доля наземных кормов в их питании снижается и повышается процент водных форм. Более высокая доля наземных кормов у прудовых лягушек по сравнению с озерными может быть связана со способностью первых удаляться от водоемов на более значительные расстояния в поисках пищи (Банников, Денисова, 1956).

Способы группировки количественных данных, применяемых при выяснении причинно-следственных отношений между признаками, представлены в таблицах 2.3 и 2.4. В таблице 2.3 показано, каким образом изменяется биомасса бентоса в озерах, расположенных в разных природных зонах. При первичном анализе таблицы трудно определить, действительно ли биомасса бентоса зависит от природной зоны, в которой находятся озера. Для этого необходимо провести дисперсионный анализ, представленный в главе 6. Единственное, что бросается в глаза при анализе таблицы, – это более высокие биомассы бентоса в озерах, расположенных в зоне смешанных лесов. Ещё более сложный вариант группировки данных представлен в таблице 2.4. В данном случае вводится дополнительный фактор – площадь озер, и все количественные данные по биомассе бентоса группируются в «осях» двух факторов. При этом, в отличие от природной зоны, даже без проведения специального статистического анализа видно, что биомасса бентоса мало зависит от вариаций площади озер, поскольку при значительном увеличении площади от <10 га до >5000 га средние значения биомассы бентоса изменяются слабо ($4.5\text{--}6.7$ г/м²) (табл. 2.4). В заключение следует отметить, что единственный количественный показатель, представленный в таблице 2.4 и не оговоренный до сих пор, – это *среднее значение признака*. Этим мы займемся в следующем разделе.

2.4. Статистические характеристики выборочной совокупности, или как сжато описать данные

В предыдущих разделах были представлены элементарные приемы упорядочения и визуализации количественных данных, полученных с использованием выборочного метода исследования. Однако до сих пор речь в основном шла лишь о различных способах группировки исходных выборочных данных, без расчета каких-либо (отличающихся от исходных вариантов) обобщающих числовых показателей, способных характеризовать выборку целиком. Для более полного описания выборочной совокупности используются специально разработанные статистические характеристики – *средние значения* и *показатели вариации*. При изучении биологических и экологических объектов расчет выборочных характеристик составляет основу первичной математической обработки данных.

Средние величины

Необходимость определения средней величины какого-либо количественного признака обычно возникает тогда, когда исследователю предстоит сравнить между собой выборки по степени выраженности данного признака. Рассмотрим, к примеру, данные о росте девочек и мальчиков дошкольного возраста.

<i>Рост 6-летних девочек, см</i>	<i>Рост 6-летних мальчиков, см</i>
128.5	111
135	112
126	125
124	116
128.5	120
124.6	127
124	119
124	127
130	135
125	116

Возникает вопрос: отличаются ли в выборках девочки 6-летнего возраста от мальчиков этого же возраста по росту? Уже интуитивно ясно, что пытаться ответить на данный вопрос, сравнивая между собой отдельных детей по росту, было бы нецелесообразным. Во-первых, в силу изменчивости признака в выборке есть девочки, которые как выше, так и ниже отдельных мальчиков, точно так же встречаются и мальчики, которые выше или ниже некоторых девочек. Во-вторых, как правило, получаемые исследователями реальные выборки часто бывают значительно многочисленней той, что рассматривается в данном примере. Уже одно это будет затруднять сопоставление отдельных значений признака 2-х выборок между собой. Таким образом, чтобы сравнить выборки мальчиков и девочек по росту, необходимо учитывать все значения признака одновременно, т. е. рассматривать выборку из мальчиков и выборку из девочек в целом. Сделать это можно, вычисляя средние значения изучаемого признака.

Средние величины принято разделять на *степенные* и *структурные*.

I. Степенные средние величины. Существует несколько видов степенных средних (средняя арифметическая, средняя геометрическая, средняя квадратичная, средняя кубическая), но в практике биологических и экологических исследований наибольшее значение имеет средняя арифметическая – величина, вокруг которой «концентрируются» отдельные значения признака.

1. Средняя арифметическая – это отношение суммы отдельных значений признака (X_i) в выборке к их числу (объему выборки, n). Если средняя арифметическая рассчитывается на основе данных выборки (выборочное среднее значение), то её обозначают символами с чертой наверху – \bar{X} , \bar{Y} и т. д. или M . Если среднюю арифметическую получают при изучении всей генеральной совокупности (генеральное среднее значение), то используют символ μ (читается как «мю»).

Общая формула для определения средней арифметической имеет вид:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

Рассчитав по данной формуле средние значения роста в выборках, для девочек получим величину **127** см, а для мальчиков – **121** см. В итоге можно утверждать, что в полученных выборках девочки в среднем выше мальчиков.

Значение средней арифметической:

- обладает способностью характеризовать целую группу однородных единиц наблюдения одним числом;

- является центром вариационного распределения, вокруг которого группируются отдельные значения выборочной совокупности, взаимопогашаются и отменяются случайные колебания от центральной тенденции;

- позволяет легко и быстро производить сравнительный анализ выборок разного объема.

Кратко рассмотрим другие степенные средние значения, получившие в биологии и экологии в целом меньшее распространение, по-видимому, из-за большей сложности вычислений и отсутствия необходимости в большинстве исследований расчета более точных средних показателей, нежели средняя арифметическая.

2. Средняя квадратическая – применяется для более точной характеристики мер площади, т. е. когда изучаются признаки, выраженные в единицах площади (см², м²), или для того, чтобы вычислить среднее арифметическое значение площади на основании замеров линейного показателя (диаметр), характеризующего эту площадь. В последнем случае определяют среднюю квадратическую для линейного показателя.

Ее можно использовать при расчете среднего диаметра эритроцитов, величины листовой пластинки у растений, размеров колоний микробов, площади поверхности покровов тела и т. д.

Средняя квадратичная равняется корню квадратному из суммы квадратов отдельных значений признака, отнесенной к их общему числу (объему выборки), и рассчитывается по формуле:

$$\bar{X}_q = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}.$$

3. Средняя кубическая используется в качестве характеристики объемных признаков или для того, чтобы вычислить среднее

арифметическое значение объема на основании замеров линейного показателя (диаметр), характеризующего этот объем. В последнем случае определяют среднюю кубическую для линейного показателя. Средняя кубическая равняется корню кубическому из суммы кубов отдельных значений признака, отнесенной к их общему числу (объему выборки), и рассчитывается по формуле:

$$\bar{X}_Q = \sqrt[3]{\frac{\sum_{i=1}^n X_i^3}{n}}.$$

Средняя кубическая может быть полезной при расчетах среднего размера клеток микроскопических водорослей (мкм³), определении среднего суммарного объема (биомассы) бактерио- и фитопланктона и т. д.

4. Средняя геометрическая – используется при исследовании средней скорости прироста какой-то величины с течением времени, характеризует процесс. Средняя геометрическая обычно применяется при анализе признаков, величина которых во времени изменяется по закону геометрической прогрессии. Сюда относятся изменение веса тела в начальном периоде роста организма или рост численности популяции в естественных условиях.

Скорость прироста часто выражают в относительных величинах. Относительную скорость роста можно вычислить по формуле Ч. Майнота:

$$V = \frac{t_2 - t_1}{t_1} \cdot 100\%,$$

где t_1 и t_2 – значения признака в начале и конце исследуемого отрезка времени.

Если вычислены величины относительной скорости роста $V_1, V_2, V_3 \dots V_n$ для последовательных равных промежутков времени, то средняя относительная скорость роста (средняя геометрическая) для всего периода исследования вычисляется по следующей формуле:

$$\bar{X}_g = \sqrt[n]{V_1 \cdot V_2 \cdot V_3 \dots V_n}.$$

II. Структурные (нестепенные) средние величины характеризуют структуру распределения признака.

1. Медиана (Me) – значение признака, относительно которого ранжированный ряд делится на 2 равные части: в обе стороны от медианы располагается одинаковое число вариантов.

2. Мода (Mo) – значение признака, наиболее часто встречающееся в выборочной совокупности. Класс с наибольшей частотой называется *модальным*. На гистограмме распределения частот моде соответствует самый высокий столбец, на вариационной кривой – самая высокая точка.

Пример: вернемся к данным о количестве птенцов в гнездах древесной ласточки *Tachycineta bicolor* (Рокицкий, 1973):

4 6 6 4 5 5 5 5 5 5 1 4 5 4 5 4 5 5 7 4 6 6 5 6 4 4 5 6
1 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 7 – ранжированный ряд.

Расположив данные в ранжированный ряд, легко можно найти медиану, т. е. значение признака, разделяющее ряд на 2 равные части ($Me = 5$) и наиболее часто встречающееся значение признака ($Mo = 5$).

Значение структурных средних

– Эти величины менее чувствительны по сравнению со средней арифметической к крайним членам (наиболее низким и наиболее высоким значениям) выборочной совокупности, которые бывают для неё как раз менее характерными.

2 3 3 4 4 4 4 5 5 5 5 5 5 5 5 60 – значения признака
 \bar{X} (со значением 60) = 7.6, \bar{X} (без значения 60) = 4.3, $Mo = 5$, $Me = 5$.

Так, наличие в выборке лишь одного значения признака (60), резко отклоняющегося от всех остальных, приводит к значительному смещению средней арифметической. Структурные средние в данном случае являются более устойчивыми характеристиками выборки.

– Если исследователь имеет дело с качественными данными, часто структурные средние оказываются единственно возможной количественной характеристикой «центра» (средней величины)

признака. Таковой может быть мода. Достаточно подсчитать частоту встречаемости того или иного качественного признака в популяции (серая, белая, черная окраска особей), при этом мода будет указывать, к примеру, на наиболее типичный («средний» для популяции) тип окраски.

Примечание. По рекомендациям некоторых авторов (Гланц, 1999; Платонов, 2000), при наличии нормального распределения значений признака расчет средней арифметической является лучшей характеристикой выборки. Напротив, когда значения признака распределены несимметрично относительно среднего (сильно отклоняются от нормального распределения), среднее выборочное значение лучше описывать с помощью медианы. Понятие закона нормального распределения случайной величины будет дано в главе 3.

Показатели вариации

Средние величины не являются универсальными характеристиками варьирующих признаков. При одинаковых средних значениях признаки могут различаться по степени и характеру варьирования.

Пример:

Выборка 1: 1 2 3 4 5 $\bar{x}_1 = 3$

Выборка 2: 3 3 3 3 3 $\bar{x}_2 = 3$

Таким образом, для полной количественной характеристики любого признака (показателя) на основе выборочной совокупности его значений, помимо средней величины, необходимо учитывать степень отклонения от неё вариантов, а также знать существенные черты варьирования признака.

Для этих целей разработаны разные *показатели вариации*, которые находят широкое применение в биологии и экологии.

Вариацию признаков можно оценить с помощью следующих количественных характеристик:

1. *Лимиты (пределы вариации)* – минимальное и максимальное значение признака в выборочной совокупности. Указывают границы варьирования признака. Обозначаются как *lim*.

2. *Размах вариации* – разность между максимальным и минимальным значением признака. Обозначается буквой *R*.

Чем сильнее варьирует признак, тем больше показатели пределов и размаха вариации, и наоборот.

Пример: диаметры (мм) колоний 2 штаммов бактерий составили:

$$X_1 \ 2.0 \ 2.2 \ 2.4 \ 2.6 \ 2.8 \ \bar{x}_1 = 2.4, \text{ lim} = 2.0-2.8, R = 0.8$$

$$X_2 \ 1.6 \ 2.0 \ 2.4 \ 2.8 \ 3.2 \ \bar{x}_2 = 2.4, \text{ lim} = 1.6-3.2, R = 1.6$$

Из примера видно, что вариабельность диаметра колоний 2-го штамма бактерий больше. Однако применение этих 2-х показателей в биологии и экологии для оценки вариации признаков имеет ограниченное значение, поскольку они зачастую не отражают сам характер варьирования признаков.

Пример: рассмотрим 2 выборочные совокупности:

$$X_1 \ 100 \ 110 \ 120 \ 130 \ 140 \ 150 \ 160 \ 170 \ 180 \ 190 \ \bar{x}_1 = 145, \text{ lim} = 100-190, R = 90$$

$$X_2 \ 100 \ 145 \ 145 \ 145 \ 145 \ 145 \ 145 \ 145 \ 145 \ 190 \ \bar{x}_2 = 145, \text{ lim} = 100-190, R = 90$$

Лимиты и размах вариации имеют одинаковые значения в обеих выборках, однако если внимательно присмотреться, то сам характер варьирования значений в каждой из выборок существенно различается. Если в первой выборке все варианты отличаются друг от друга, то во второй выборке из 10 вариантов 8 имеют одинаковые значения. Таким образом, в первой выборке рассеяние вариантов больше, чем во второй, но это никак не сказывается на лимитах и размахе вариации.

Очевидно, чтобы преодолеть отмеченные недостатки, необходимо учитывать не только крайние значения признака (лимиты), но и все варианты в выборке. Наиболее рациональный путь заключается в определении отклонений каждого отдельного значения признака от средней величины – $(x_i - \bar{x})$, затем все полученные отклонения можно просуммировать и разделить на объем выборки. В итоге мы получим некое среднее линейное отклонение, которое будет тем больше, чем значительнее каждая варианта будет отклоняться от среднего значения. Таким образом, с помощью этого показателя можно было бы сравнивать разные выборки по степени варьирования признака и одновременно учитывать внутренние черты вариации (степень отличия каждой варианты). Обратимся к предыдущему примеру и рассчитаем среднее линейное отклонение для каждой из выборок:

$$\begin{array}{l} (X_i - \bar{X}_1): \quad -45 \quad -35 \quad -25 \quad -15 \quad -5 \quad +5 \quad +15 \quad +25 \quad +35 \quad +45 \\ (X_i - \bar{X}_2): \quad -45 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad +45 \end{array}$$

Попытавшись просуммировать все полученные отклонения, мы натолкнемся на весьма существенное затруднение, которое легко будет увидеть, приступив к вычислению: сумма отклонений в обеих выборках будет равна 0.

И это не случайная игра чисел – данное затруднение будет возникать всякий раз для любой другой выборки при суммировании отклонений вариант от средней арифметической. Один из математических приемов избавления от отрицательных значений полученных отклонений – возведение их в квадрат. Так мы подходим к одному из ключевых понятий биометрии и показателей вариации.

3. Дисперсия (σ^2 , S^2) – это отношение суммы квадратов отклонений отдельных значений признака от средней арифметической к объему выборки за вычетом единицы:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

Из двух обозначений дисперсии пока будем применять символ S^2 , он используется, если дисперсия рассчитывается по выборочным данным. В числителе данной формулы суммируются не отдельные отклонения, а квадраты отклонений – таким образом мы избегаемся от нулевой суммы.

$$\begin{array}{l} (X_i - \bar{X}_1)^2: \quad -45^2 \quad -35^2 \quad -25^2 \quad -15^2 \quad -5^2 \quad +5^2 \quad +15^2 \quad +25^2 \quad +35^2 \quad +45^2 \\ (X_i - \bar{X}_2)^2: \quad -45^2 \quad 0^2 \quad 0^2 \quad 0^2 \quad 0^2 \quad 0^2 \quad 0^2 \quad 0^2 \quad 0^2 \quad +45^2 \end{array}$$

$(X_i - \bar{X}_1)^2:$	2025	1225	625	225	25	25	225	625	1225	2025	$\Sigma =$	8250
$(X_i - \bar{X}_2)^2:$	2025	0	0	0	0	0	0	0	0	2025	$\Sigma =$	4050

В итоге получаем:

$$S^2_1 = \frac{8250}{10-1} = 916.7 \quad S^2_2 = \frac{4050}{10-1} = 450.$$

Как и следовало ожидать, при одинаковых значениях размаха вариации и лимитов вариабельность значений по показателю дисперсии в первой выборке оказалась выше, чем во второй выборке. Единственное, что не было отмечено в формуле дисперсии, – это находящаяся в знаменателе разность $n - 1$. Эта разность называется в статистике *числом степеней свободы*. Дело в том, что, когда исследователь рассчитывает дисперсию на основе выборки, взятой из генеральной совокупности (а это бывает, как правило, в большинстве случаев), получаемое выборочное значение дисперсии, строго говоря, оказывается заниженным (или, как говорят, смещенным) относительно реально существующей генеральной дисперсии, т. е. того значения дисперсии, которое могло бы быть получено, если бы исследователь использовал все значения признака из генеральной совокупности. Чтобы скорректировать существующее занижение дисперсии, получаемую сумму в числителе делят не на n , а на немного меньшее число – $n - 1$, что приводит к возрастанию величины дисперсии, рассчитанной на основе выборочных данных.

В некоторых случаях использование дисперсии оказывается не очень удобным, поскольку в формуле каждое отклонение варианты от среднего значения возводится в квадрат, в итоге дисперсия измеряется в единицах, равных квадрату единицы измерения. Так, например, если высчитывается дисперсия измеряемого в килограммах веса, то сама дисперсия будет выражаться в квадратных килограммах, что само по себе бессмысленно. Поэтому часто используется другой, очень близкий к дисперсии показатель вариации.

4. Среднее квадратическое (стандартное) отклонение (σ , S) – корень квадратный из дисперсии. Если стандартное отклонение рассчитывается по выборочным данным, то используется обозначение S , если на основе генеральной совокупности, то символ σ (читается как «сигма»). Действительно, для избавления от квадратов отклонений прибегают к действию, противоположному возведению в степень, т. е. извлекают квадратный корень. В итоге стандартное отклонение является в ряде случаев более удоб-

ной характеристикой вариации признаков, поскольку измеряется в тех же единицах, что и исходные данные.

$$S_1 = \sqrt{916.7} = 30.3 \quad S_2 = \sqrt{450} = 21.2$$

Таким образом, дисперсия и стандартное отклонение являются мерой варьирования числовых значений признака вокруг их средней арифметической и одновременно отражают внутреннюю изменчивость значений признака, зависящую от разностей между отдельными значениями признака.

Однако эти показатели затруднительно использовать при решении ряда задач сравнения признаков по степени варьирования. Поэтому в биологии и экологии широкое распространение получила также относительная количественная характеристика вариации.

5. Коэффициент вариации (C_v) – отношение стандартного отклонения к средней арифметической величине, выраженное в процентах:

$$C_v = \frac{S}{\bar{X}} \cdot 100\% .$$

Варьирование считается слабым при $C_v \leq 10\%$, средним при $C_v = 11-25\%$, сильным при $C_v > 25\%$ (Лакин, 1990).

Значение коэффициента вариации

- Дисперсия и стандартное отклонение применимы для сравнительной оценки признаков, выраженных в одних и тех же единицах измерения. Коэффициент вариации позволяет сравнивать вариацию признаков, выраженных разными единицами измерения.

Пример: сравним 2 признака (например, вес и длина тела) по степени вариабельности:

$$\begin{aligned} \bar{X}_1 &= 2.4 \text{ кг}, & S &= 0.6 \text{ кг}, & C_v &= 24\% \\ \bar{X}_2 &= 8.3 \text{ см}, & S &= 1.6 \text{ см}, & C_v &= 19\% \end{aligned}$$

Стандартное отклонение показывает большую изменчивость длины тела по сравнению с весом. Однако стандартное отклонение выражено в тех же единицах измерения, что и исходные данные, поэтому сравнивать эти величины не вполне корректно.

В данном случае уместно использовать безразмерный коэффициент вариации, анализ значений которого даёт обратный результат: вес варьирует сильнее длины тела.

- Коэффициент вариации позволяет сравнивать вариацию признаков, выраженных в одних и тех же единицах измерения, но резко различающихся по величине среднего значения.

Пример:

длина ног кур	$\bar{X}_1 = 10$ см,	$S = 1$ см,	$C_v = 10\%$
длина ног страусов	$\bar{X}_2 = 150$ см,	$S = 6$ см,	$C_v = 4\%$

Стандартное отклонение определяется по отклонениям отдельных значений от среднего, но в данном примере средние значения различны по величине, поэтому только это может давать значительные отличия стандартных отклонений. Расчет коэффициента вариации подтверждает эти рассуждения: длина ног кур, несмотря на меньшее значение стандартного отклонения, оказывается более вариабельным признаком по сравнению с длиной ног страусов.

Примечание 1. Показатели вариации, помимо того что используются как вспомогательные величины при расчетах большого количества статистических характеристик (стандартная ошибка, статистические критерии, коэффициенты корреляций и т. д.) и как показатели вариабельности значений в отдельных выборках, имеют в практике биологических и экологических исследований и самостоятельное значение. К примеру, расчет дисперсии и стандартного отклонения применяется при определении типа пространственной структуры популяций (случайное, равномерное, групповое распределение), эти же показатели и коэффициент вариации широко используются при расчетах индексов агрегированности и установлении неоднородности (или равномерности) пространственного распределения каких-либо показателей на определенной территории или в водоеме. В систематике лимиты относительно широко используются в определительных таблицах с целью упрощения идентификации видов (подвидов). В гидробиологии для характеристики изменчивости биомассы гидробионтов в течение сезонов, года или в межгодовой динамике используется отношение максимальной к минимальной биомассе за эти промежутки времени. Данный показатель получил название «вариабельность динамики

биомассы», ВДБ (Алимов, 2001). Как оказалось, ВДБ связана со многими структурными и функциональными показателями водных экосистем: в частности, в сообществах с высокими значениями ВДБ преобладают эврибионтные виды с r-стратегией, а при загрязнении и эвтрофировании водоемов ВДБ возрастает (Алимов, 2001). Знание степени изменчивости тех или иных признаков имеет большое значение в генетике и селекции. Например, при сравнении двух сходных по продуктивности и качественным показателям сортов предпочтение должно быть отдано тому из них, который при равных условиях обладает меньшей изменчивостью.

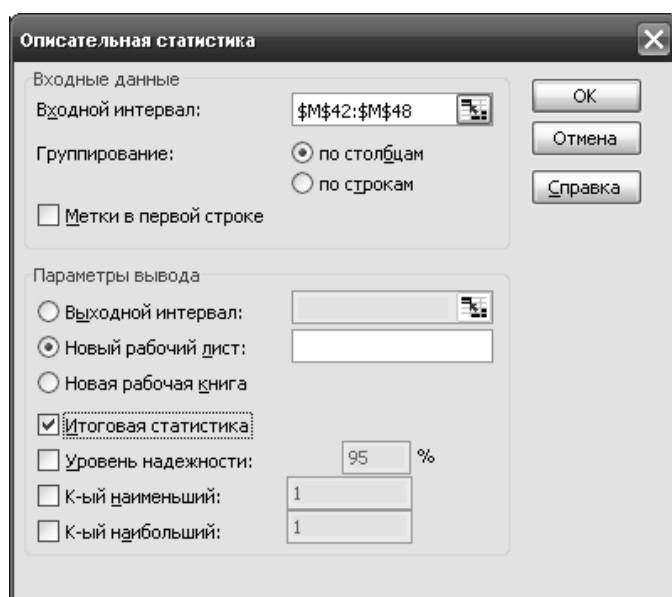
Примечание 2. Важно отметить, что, рассчитывая среднее и показатели вариации для выборочных данных, фактически исследователь оценивает вклад систематических и случайных (неизвестных) факторов в вариационное распределение признака. Средняя арифметическая есть поэтому характеристика действия доминирующего фактора на одну варианту, а показатели вариации есть мера изменчивости признаков, обусловленная влиянием на них случайных факторов (Ивантер, Коросов, 2005). Чем больше случайных факторов, чем они сильнее, тем дальше будут разбросаны отдельные значения признака вокруг средней и тем большими оказываются показатели вариации (Ивантер, Коросов, 2005).

Программное обеспечение. В результате первичной математической обработки исходных данных исследователь получает простой набор обобщающих количественных характеристик, с помощью которых можно сжато описать выборки любого объема. Однако расчет всех этих характеристик по приведенным формулам вручную или на калькуляторе требует значительных затрат времени. Естественно, что подобный подход в современных исследованиях уже давно исключен, поэтому возникает необходимость овладения навыками подобных расчетов на персональном компьютере. Это повышает качество анализа и сокращает затраты времени. В MS EXCEL статистические характеристики выборки можно вычислять с помощью встроенных функций: СРЗНАЧ – средняя арифметическая; СРГЕОМ – средняя геометрическая; ДИСП – дисперсия; СТАНДОТКЛОН – стандартное отклонение; МОДА; МЕДИАНА и др.

Программной реализацией расчета статистических характеристик выборочной совокупности является модуль «*Описательная статистика*» (Descriptive statistics), включенный в большин-

ство пакетов для статистической обработки данных. Работая в данном модуле, достаточно сделать ссылку на массив данных выборки, выбрать необходимые статистические характеристики в специальном диалоговом окне и нажать на кнопку **ОК** или **Выполнить расчет**. В результате исследователь получит таблицу с рассчитанными статистическими характеристиками выборочной совокупности. Примеры подобных модулей приведены ниже (рис. 2.5; 2.6; 2.7).

Краткое описание запуска модуля: в верхнем меню Statistics надо выбрать команду Basic Statistics/Tables (основные статистики/таблицы). В появившемся меню надо выбрать команду Descriptive statistics (описательные статистики). Для выбора переменной, описательные статистики которой нас интересуют, надо нажать кнопку Variables и в открывшемся окне щелкнуть на имени переменной (переменных). Зайти на вкладку Advanced, установив флажки напротив соответствующих показателей. Для просмотра результатов надо нажать кнопку Summary. Откроется таблица с основными статистиками.



Среднее	86.9
Стандартная ошибка	25.3
Медиана	124.5
Мода	120
Стандартное отклонение	62.1
Дисперсия выборки	3857.6
Эксцесс	-1.7
Асимметричность	-0.97
Интервал	128.3
Минимум	1.13
Максимум	129.5
Сумма	521.7
Счет	6
Уровень надежности (95,0%)	65.1

Рис. 2.5. Диалоговое окно процедуры «Описательная статистика» табличного процессора MS Excel и таблица результатов

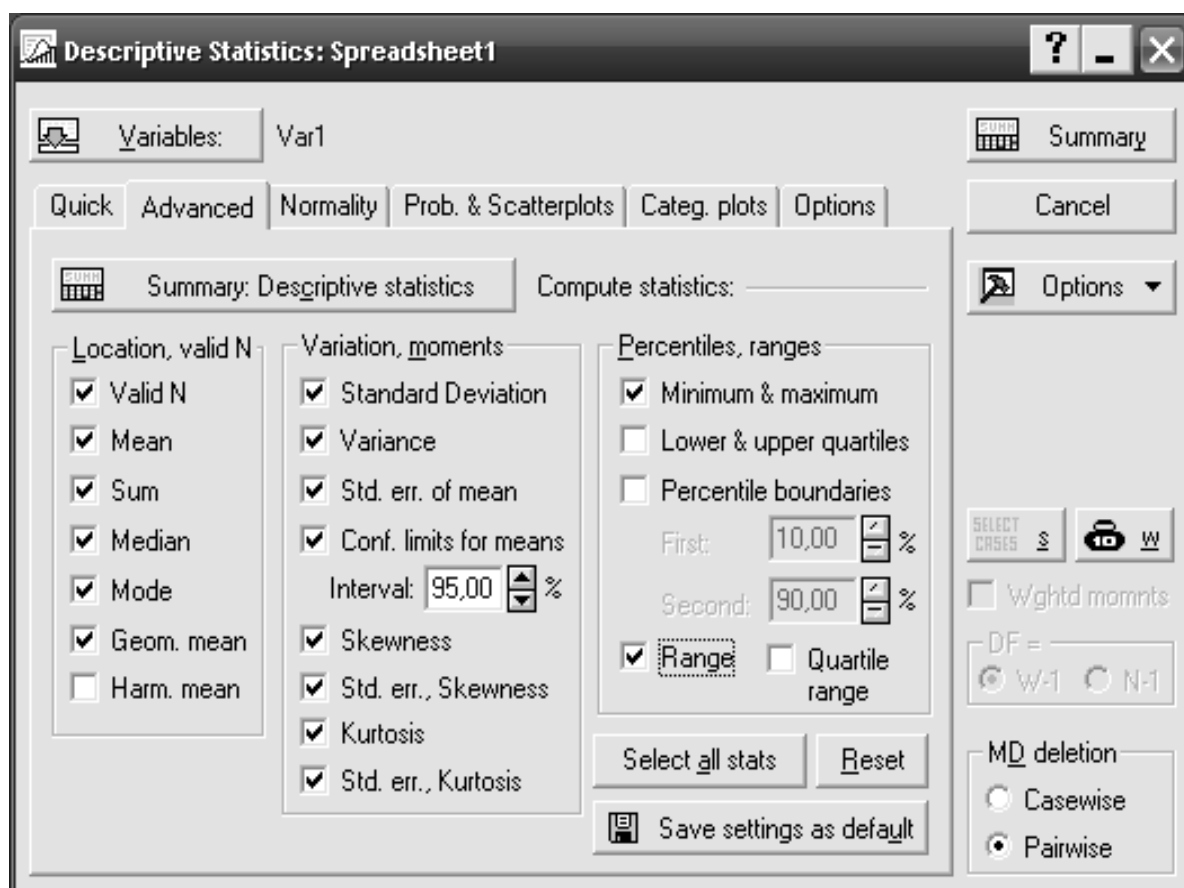


Рис. 2.6. Диалоговое окно модуля «Описательная статистика» (Descriptive statistics) пакета STATISTICA

Краткое описание статистических характеристик:

Valid N – объем выборки; **Mean** – средняя арифметическая; **Median** – медиана; **Mode** – мода; **Geom. mean** – средняя геометрическая; **Standard Deviation** – стандартное отклонение; **Variance** – дисперсия; **Standard error of mean** – стандартная ошибка; **95% confidence limits of mean** – доверительный интервал для среднего; **Minimum & maximum** – минимальное и максимальное значения (лимиты); **Lower & upper quartiles** – нижняя и верхняя квартили; **Range** – размах вариации; **Quartile range** – квартильный размах; **Skewness** – асимметрия; **Kurtosis** – эксцесс; **Standard errors of skewness & kurtosis** – ошибки асимметрии и эксцесса.

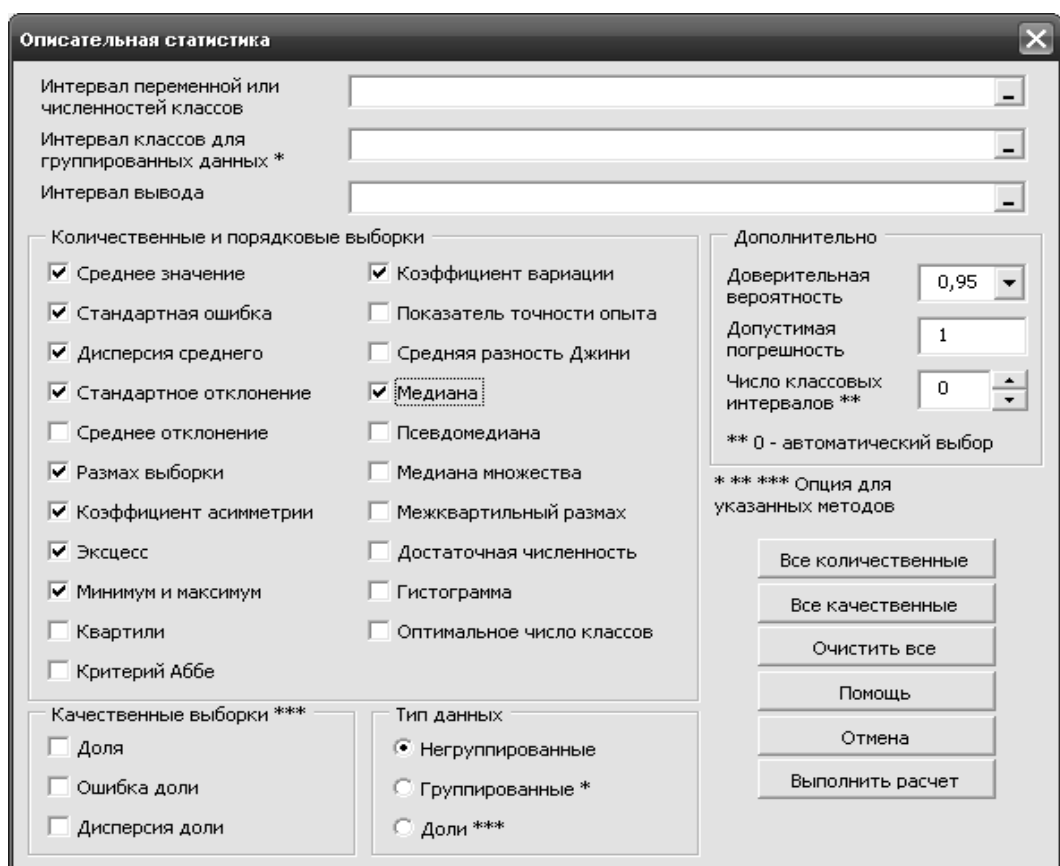


Рис. 2.7. Диалоговое окно модуля «Описательная статистика» программы ATTESTAT: выберите из меню программы пункт AtteStat | Модуль «Описательная статистика», на экране появится диалоговое окно, изображенное на рисунке

Глава 3. Законы распределения биологических и экологических переменных

В данной главе мы познакомимся с фундаментальным понятием вероятности, узнаем, что такое закон распределения признака и для чего исследователю перед применением методов математической статистики необходимо знать, по какому закону распределяются биологические и экологические переменные. И, наконец, кратко рассмотрим наиболее важные из этих законов распределения.

3.1. Вероятность события

Для того чтобы двигаться дальше в понимании более сложных методов математической обработки данных, необходимо остановиться на некоторых ключевых понятиях теории вероятностей.

1. Исход или результат отдельного испытания называется *событием*.

2. Под *испытанием* понимается некоторый комплекс условий, необходимых для того, чтобы мог осуществиться тот или иной исход.

3. У некоторых событий может быть только один исход, заранее предсказуемый. Такие события называются *достоверными* – падение подброшенного камня вниз, по длине стороны квадрата всегда точно можно определить его площадь.

4. Если при осуществлении комплекса условий события заведомо произойти не могут, то они называются *невозможными* – всплытие камня, брошенного в воду, площадь квадрата с длинной стороны 1 см, равная 4 см².

5. Однако существуют события, исход которых заранее непредсказуем, такие события при осуществлении комплекса условий могут произойти, а могут не произойти. Они называются *случайными* – при бросании монеты она может упасть вверх решкой или орлом, в популяции может родиться самец или самка.

6. Случайные события называются *несовместными*, если в серии испытаний всякий раз возможно осуществление только одного из них – при бросании монеты она может упасть вверх орлом или решкой, но одновременно оба события произойти не могут.

Кратко познакомившись с некоторыми понятиями, перейдем к классическому определению вероятности события. *Вероятностью события A* ($P(A)$) называют отношение числа благоприятствующих этому событию исходов (m) к общему числу всех равновозможных и несовместных исходов (n) (Гмурман, 2001).

$$P(A) = \frac{m}{n}.$$

Пример: в урне находится 5 белых и 10 черных шаров. Наугад необходимо вынуть 1 шар. Какова вероятность, что вынутый шар окажется белым? Число всех равновозможных исходов будет равно 15 (5+10). Число благоприятствующих исходов, т. е. случаев вынимания белого шара в однократном испытании, равно 5. Следовательно, вероятность вынуть белый шар равна:

$$P = \frac{5}{15} = \frac{1}{3} \approx 0.33 \approx 33\%$$

Таким образом, вероятность представляет собой число, заключенное между 0 и 1, и может быть выражена либо в долях единицы, либо в процентах от общего числа испытаний. В свете этих представлений *достоверным* называется событие, вероятность которого равна 1. Если вероятность некоторого события равна 0, то это событие рассматривают как *невозможное*; если вероятность заключена где-то между 0 и 1 – как *случайное*. Случайное событие, вероятность которого близка к 0, но нулю не равняется, принято считать *малодостоверным* (*практически невозможным, маловероятным*), и наоборот, если вероятность случайного события приближается к 1, но единице не равняется, то говорят о *практически достоверном* событии.

3.2. Закон распределения

Биологические и экологические явления (события) случайны, точно не предсказуемы. Начиная биологический эксперимент или приступая к наблюдению, невозможно точно сказать, каков будет результат – уровень численности животных в данном районе, выживаемость подопытных особей, артериальное давление через час после введения препарата. Поэтому биологам и экологам часто приходится сталкиваться с вероятностными (стохастическими) суждениями. Так, для гидробиолога, изучающего чистое олиготрофное озеро, ясно, что вероятность обнаружить массовое развитие водоросли *Planktothrix agardhii* (индикатор высокой степени сапробности) крайне мала. Эксперимент по проверке токсичности определенного вещества может показать, что в контрольном варианте выжило на 10% особей больше, чем в опытном (с добавкой вещества). Зависела ли эта разница в выживаемости особей от действия вещества или могла определяться другими факторами (например, изначальной разницей физиологического состояния особей в группах)? Экспериментатор может сказать следующее: «Очень вероятно, что именно тестируемое вещество определило большую смертность особей в опытной группе по сравнению с контрольной». Его более скептически настроенный коллега может заявить: «Небольшая разница, всего лишь в 10%, могла быть следствием действия случайных (неконтролируемых в эксперименте) причин, поэтому маловероятно, что вещество является токсичным».

Однако любому биологу и экологу ясно, что случайность изучаемых ими явлений относительна, несмотря на то, что точный прогноз невозможен, приблизительный результат можно предсказать. Каким образом можно дать такого рода прогноз?

Рассмотрим пример. Зоолог, изучающий популяцию какого-либо вида животного, задался целью дать приблизительный прогноз появления особей в популяции с некой мутацией (например, связанной с окраской). Чтобы рассчитать вероятность, ему потребуются предварительные исследования и данные о том, насколько часто в популяции рождаются особи с данной мутацией. Так, если исследователь обнаружит, что за ряд предшествующих лет из 10 000 родившихся особей 100 имели данную мутацию, то он сможет рассчитать вероятность рождения мутантной особи в данной популяции:

$$P = \frac{100}{10000} = 0.01 .$$

Другими словами, в среднем из 100 родившихся особей одна может быть мутантной. При наличии подобных данных можно решить и обратную задачу – найти вероятность появления в популяции особи без данной мутации:

$$P = \frac{9900}{10000} = 0.99 .$$

Из этого абстрактного примера вытекают 2 важных вывода. Во-первых, сумма вероятностей противоположных событий $(0.01+0.99)$ всегда равна единице. Во-вторых, приблизительный (вероятностный) прогноз можно дать, ориентируясь на повторяемость однотипных событий, на частоту встречаемости значений признака. Зная частоту, с которой данное значение признака встречается в популяции относительно общего количества всех встреченных значений признака (объем выборки), можно установить *статистическую вероятность* появления данного значения признака:

$$P = \frac{f}{n} ,$$

где f – частота встречаемости,
 n – объем выборочной совокупности.

Статистическую вероятность события принято называть *относительной частотой*. Установлено, что относительная частота полностью не совпадает с «классической» вероятностью, однако приближается к ней по мере значительного увеличения числа наблюдений, т. е. объема выборки.

Таким образом, зная ряд распределения частоты встречаемости значений признака, можно легко перейти к построению *распределения вероятностей*. Сделаем это, вновь обратившись к данным о количестве птенцов в гнездах древесной ласточки *Tachycineta bicolor* ($n = 42$) (Рокицкий, 1973):

Количество птенцов	Частота встречаемости	Вероятность
1	1	$1/42 = 0.02$
2	2	$2/42 = 0.05$
4	11	$11/42 = 0.26$
5	18	$18/42 = 0.43$
6	9	$9/42 = 0.22$
7	1	$1/42 = 0.02$
Σ	42	1.0

Кроме того, закономерность, отмеченную в распределении вероятностей, можно выразить не только в табличной форме (ряд распределения), но и графически:

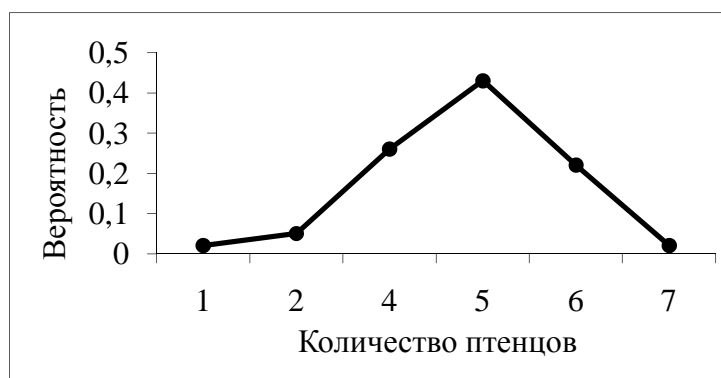


Рис. 3.1. Кривая распределения вероятностей появления в выводке того или иного количества птенцов

Наконец, закономерность распределения вероятностей можно описать с помощью математической формулы. Функция, связывающая значения случайного признака с их вероятностями, называется *законом распределения признака*. Каждый признак (показатель) распределяется по своему закону, имеет специфическую закономерность распределения (повторяемости) отдельных значений. Поэтому закон распределения образно можно сравнить

с «паспортом» признака. В зависимости от типа переменной выделяют *дискретные* и *непрерывные* законы распределения. Описанное распределение относится к дискретным и, вероятнее всего, близко к так называемому *биномиальному распределению*.

К настоящему моменту известны десятки теоретических распределений (их можно построить на основе известных математических формул, рис. 3.2), к которым исследователи могут «подгонять» полученные на основе выборок эмпирические распределения, устанавливая с определенной вероятностью, по какому закону распределяются изучаемые признаки. Из всего многообразия законов распределения кратко остановимся на наиболее значимых в практике биологических и экологических исследований – *нормальном* и *биномиальном*.

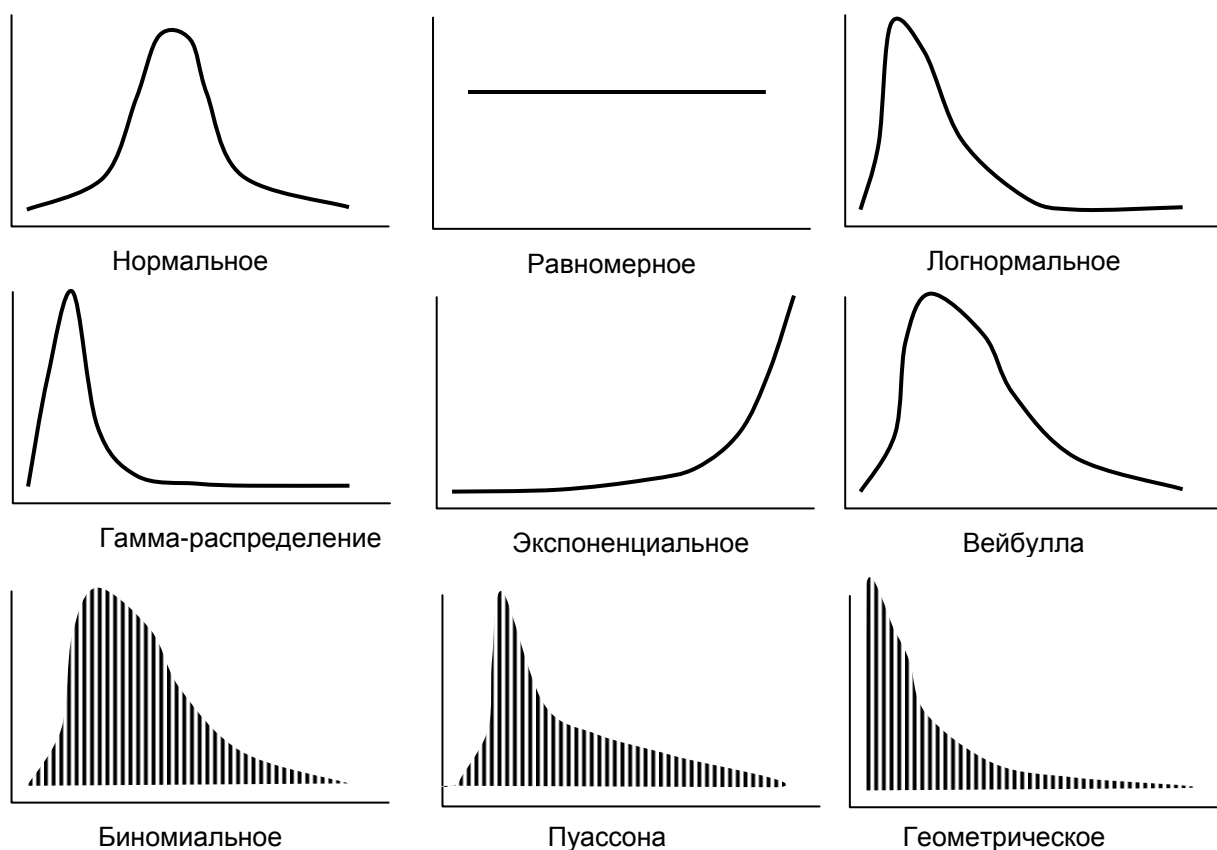


Рис. 3.2. Некоторые типы теоретических распределений случайной величины: *непрерывные* – нормальное, логнормальное, гамма-распределение, экспоненциальное, распределение Вейбулла; *дискретные* – биномиальное, распределение Пуассона, геометрическое, равномерное (по: Шитиков и др. 2003)

3.3. Нормальное распределение

Наиболее характерный тип распределения *непрерывных случайных величин*, из него можно вывести (к нему сводятся) все остальные. Термин «нормальное распределение» введен в биологическую лексику Ф. Гальтоном в 1889 году. Однако ещё задолго до этого оно было хорошо известно математикам, которые это распределение часто называют законом Гаусса – Лапласа. Название распределения, конечно, не означает, что все другие законы распределения «ненормальные», или атипичные. Просто подобное распределение значений признака так часто встречается в самых различных областях науки и практики, что первоначально принималось за «норму» случайного проявления признаков.

Графически нормальное распределение выглядит как симметричная колоколообразная кривая. Основная закономерность при нормальном распределении значений признака заключается в том, что крайние значения (наибольшие и наименьшие) появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем они чаще встречаются (рис. 3.3).

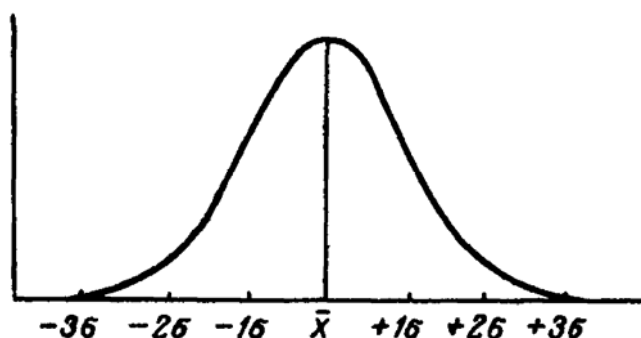


Рис. 3.3. Кривая нормального распределения

Действительно, если взять такой признак, как человеческий рост, и построить распределение, то можно убедиться, что люди с неким средним ростом будут встречаться очень часто (центр распределения), а вероятность обнаружить людей с очень высоким (например, выше 2 м) или очень низким (например, менее 1 м) ростом будет значительно меньше (края распределения). Это означает, что человеческий рост – признак, подчиняющийся нормальному закону распределения.

Важной особенностью нормального распределения является то, что форма и положение его графика определяется только 2 па-

раметрами: средним значением признака и стандартным отклонением. При этом для оценки степени отклонения отдельных значений признака от среднего значения используют не само стандартное отклонение, а величины так называемого *нормированного отклонения* (t) – отношения отклонений отдельных значений признака от среднего значения к стандартному отклонению:

$$t = \frac{(X_i - \bar{X})}{\sigma}.$$

Отсюда

$$(X_i - \bar{X}) = t \cdot \sigma.$$

Таким образом, закон нормального распределения фактически описывает функциональную зависимость между вероятностью P (ось ординат, y) и нормированным отклонением t (ось абсцисс, x). Он утверждает, что вероятность отклонения любого значения признака от центра распределения (т. е. среднего значения) определяется функцией нормированного отклонения. Закон нормального распределения может быть охарактеризован довольно сложной математической формулой:

$$P = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i - \bar{X})^2}{2\sigma^2}},$$

где π – отношение длины окружности к диаметру (равно 3.1416),
 e – основание натуральных логарифмов (равно 2.718).

Остальные параметры уравнения читателю уже известны, отметим лишь, что в показателе степени величины e находится возведенное в квадрат нормированное отклонение $t^2 = \frac{(X_i - \bar{X})^2}{\sigma^2}$.

Данное уравнение определяет ход кривой нормального распределения, т. е. позволяет вычислить ординаты нормальной кривой, или «плотность вероятности» (P).

Нормальная кривая со средним значением, равным 0, и $\sigma = 1$ называется *стандартизованной* и описывается следующей математической формулой:

$$P = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Если данную формулу несколько модифицировать, то на её основе можно вычислять уже не вероятности, а теоретические частоты вариационного ряда и строить, соответственно, *теоретические* (нормальные) *кривые распределения*. Эти кривые будут плавными, и, сопоставляя их с ломаными *эмпирическими кривыми* (т. е. с вариационными кривыми или гистограммами распределения, полученными по выборке), исследователь может в первом приближении определять соответствие изучаемого признака нормальному закону распределения.

Кратко опишем свойства нормального распределения:

1. Нормальная кривая приближается к оси абсцисс асимптотически, т. е. никогда не касаясь её.

2. Все значения признака лежат в интервале плюс – минус бесконечность. Иными словами, с вероятностью $P = 1$ мы вправе ожидать появление нового значения в пределах от $-\infty$ до $+\infty$.

3. Нормальная кривая имеет характерный изгиб по мере удаления от центра распределения: точка перегиба лежит точно на расстоянии в 1σ от \bar{X} .

4. Для нормального распределения характерно совпадение средней арифметической, моды и медианы.

5. Площадь между стандартизованной нормальной кривой и осью абсцисс равна 1. Таким образом, площадь под кривой интерпретируется как вероятность.

Из этих свойств вытекает важное следствие, получившее название *правила 3-х сигм*: отдельные значения любого признака, имеющего нормальное распределение, отклоняются от среднего значения (т. е. от центра распределения) с вероятностью 0.997 не более чем на 3 сигмы влево и вправо ($\pm 3\sigma$). И только с вероятностью 0.003 отдельное значение признака может не попасть в пределы интервала $\pm 3\sigma$ (рис. 3.4).

Откуда появились в этом правиле вероятности 0.997 и 0.003?

Как мы уже знаем, отклонение каждого отдельного значения признака от центра нормального распределения характеризуется определенным значением t , т. е. $(X_i - \bar{X}) = t \cdot \sigma$. Другими словами, зная t , можно установить, в какую сторону и насколько отклоняется варианта от центра распределения.

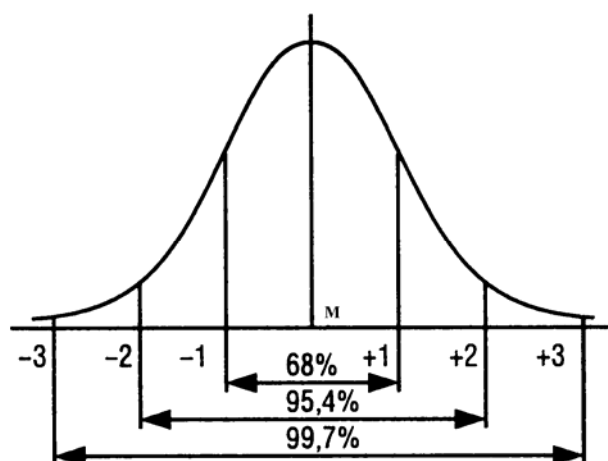


Рис. 3.4. Иллюстрация правила 3-х сигм

Так, при $t = +1$ значение признака будет находиться справа от центра распределения и отклоняться от него (т. е. от \bar{X}) на величину 1σ ; при $t = -1.5$ значение признака будет находиться слева от центра распределения и отклоняться от него на величину 1.5σ . Кроме того, из математической формулы нормального распределения, связывающей вероятность с нормированным отклонением, следует, что в границах от $-t$ до $+t$ всегда будет заключена постоянная вероятность нахождения определенного значения признака. Так вот, в пределах $\pm 1\sigma$ (т. е. при $t = 1$) всегда отсекается 68.3% от общей площади фигуры под кривой нормального распределения (рис. 3.4). Это значит, что с вероятностью 0.683 значение случайной величины попадет в интервал от -1σ до $+1\sigma$, а с вероятностью 0.317 ($1 - 0.683$) это значение может попасть за пределы данного интервала. Если интервал расширить до $\pm 2\sigma$ (т. е. при $t = 2$), то отсекается уже 95.4% от общей площади фигуры под кривой нормального распределения (рис. 3.4). Другими словами, в интервал от -2σ до $+2\sigma$ наугад отобранная варианта попадет уже с вероятностью 0.954, и лишь с вероятностью 0.046 ($1 - 0.954$) она может не попасть в этот интервал. И, наконец, в пределах $\pm 3\sigma$ (т. е. при $t = 3$) заключено 99.7% от общей площади фигуры под кривой нормального распределения (рис. 3.4). Фактически можно утверждать (предсказать), что с вероятностью 0.997 все значения признака будут отклоняться от центра распределения (средней арифметической) на величину, не превышающую $\pm 3\sigma$. И лишь с вероятностью 0.003 ($1 - 0.997$) наугад отобранная варианта может не попасть в заданные границы.

Примечание. Нормальные распределения встречаются очень часто, когда некая величина отклоняется от средней под действием множества слабых, случайных, независимых друг от друга факторов, которые приводят к формированию симметричного распределения. Таким образом, нормальное распределение является моделью идеального равновесного состояния, не подверженного действию какого-либо одного специфического фактора. Однако в естественных условиях нормальные распределения признаков часто нарушаются.

3.4. Понятие асимметрии и эксцесса распределения

Существует 2 типа отклонений распределения признака от нормальной кривой:

1. Асимметрия (As) – при нормальном распределении мода и медиана совпадают. При асимметрии мода отклоняется от медианы вправо либо влево. Если мода отклоняется влево от медианы, а правая ветвь кривой длиннее левой (т. е. является более полой), то говорят о *положительной (правосторонней) асимметрии*, при этом коэффициент асимметрии $As > 0$ (рис. 3.5). Если мода отклоняется вправо от медианы, а левая ветвь кривой длиннее правой, то говорят об *отрицательной (левосторонней) асимметрии*, при этом коэффициент асимметрии $As < 0$ (рис. 3.5).

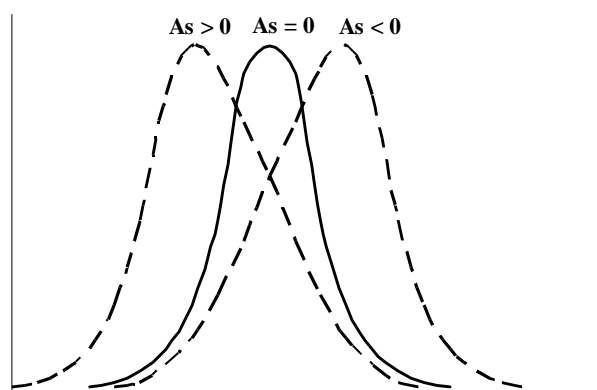


Рис. 3.5. Асимметрия распределения

2. Эксцесс (Ex) – изменение высоты модальных классов эмпирической кривой относительно высоты модальных классов

нормальной теоретической кривой. Если вершина эмпирической кривой оказывается сильно поднятой относительно вершины нормальной кривой, то говорят о *положительном эксцессе распределения*, при этом коэффициент эксцесса $E_x > 0$ (рис. 3.6). Если вершина эмпирической кривой оказывается ниже вершины нормальной кривой, то говорят об *отрицательном эксцессе распределения*, при этом коэффициент эксцесса $E_x < 0$ (рис. 3.6).

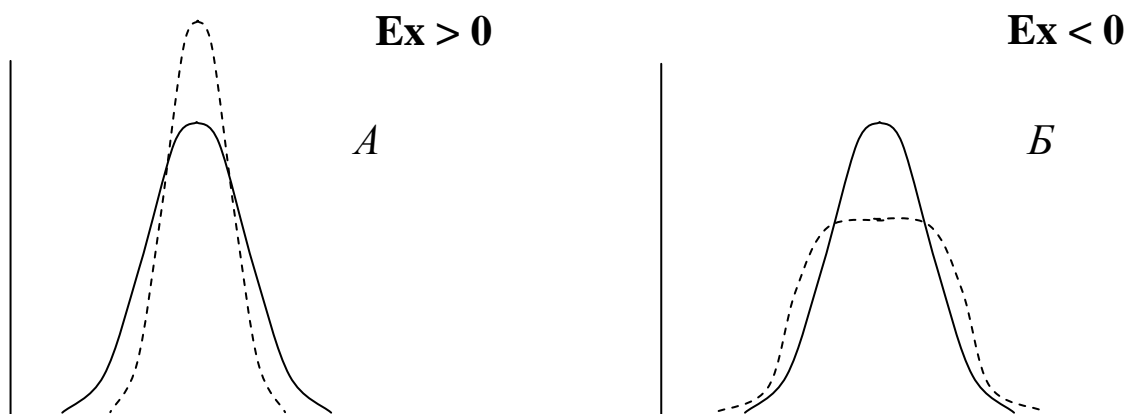


Рис. 3.6. Эксцесс распределения

Таким образом, теоретическое нормальное распределение имеет $A_s = 0$ и $E_x = 0$.

Примечание. Рассчитать коэффициенты асимметрии и эксцесса можно в модуле «*Описательная статистика*» (см. главу 2).

Причины возникновения асимметричных эмпирических распределений могут быть разными:

1. «*Механическая*» причина – асимметричность распределения связана с неправильной группировкой значений признака по классовым интервалам или с неправильным расчетом ширины классového интервала. В результате с одной стороны кривой частота встречаемости значений признака может оказаться больше, чем с другой. Подобные асимметричные распределения называются ложными.

2. *Модифицирующие условия внешней среды* – действие экстремальных или специфических факторов приводит к отклонению значений биологических признаков организма от «нормы», и при изучении данных признаков распределение их оказывается асимметричным.

3. *Неоднородность выборки* – объединение 2 разнородных совокупностей, каждая из которых имеет нормальное распределение, но в сумме они образуют асимметричное распределение (как правило, бимодальное или двухвершинное, в общем случае – полимодальное).

3.5. Биномиальное распределение

Во многом близко к нормальному. Отличие состоит лишь в том, что оно характеризует поведение дискретных признаков, выраженных целыми числами. Таким образом, при биномиальном распределении проявляется та же самая закономерность, что и при нормальном распределении: чем ближе значения дискретного признака к центру распределения, тем выше вероятность их появления.

Математически распределение называется *биномиальным*, если вероятности появления отдельных значений признака выражаются величинами, соответствующими коэффициентам разложения бинома Ньютона:

$$(p + q)^k,$$

где p – вероятность появления признака,

q – вероятность неоявления признака,

k – число классов, отличающихся по появлению признака.

Коэффициенты при отдельных членах разложения бинома Ньютона при возведении его в разные степени будут следующими:

$$\begin{aligned}(p + q)^1 &= p + q \\(p + q)^2 &= p^2 + 2pq + q^2 \\(p + q)^3 &= p^3 + 3p^2q + 3pq^2 + q^3 \\(p + q)^4 &= p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4\end{aligned}$$

Эти коэффициенты можно легко получить с помощью треугольника Паскаля, в котором цифры каждого последующего ряда получаются путем сложения двух цифр ряда, расположенного над ним (рис. 3.7).

В основе биномиального распределения лежит альтернативное проявление качественного признака: он может быть

у единичного объекта или отсутствовать, проявиться или нет. В гнездах древесной ласточки *Tachycineta bicolor* можно обнаружить 1 птенца или не одного, 2-х птенцов или не двух птенцов, 3-х птенцов или другое их количество и т. д.

<i>n</i>	Биномиальные коэффициенты																				
0						1															
1					1		1														
2				1		2		1													
3			1		3		3		1												
4			1		4		6		4		1										
5		1		5		10		10		5		1									
6		1		6		15		20		15		6		1							
7		1		7		21		35		35		21		7		1					
8		1		8		28		56		70		56		28		8		1			
9		1		9		36		84		126		126		84		36		9		1	
10	1		10		45		120		210		252		210		120		45		10		1

Рис. 3.7. Арифметический треугольник Паскаля

Отдельный корнеплод может быть больным или здоровым (признак качественный), тогда *проба* из нескольких корнеплодов будет содержать некоторое *число* здоровых корнеплодов (признак количественный), а множество равнообъемных проб образует уже выборку чисел, для которой можно построить гистограмму распределения. Вероятность отдельного события (корнеплод больной) составляет p , а вероятность альтернативного события (корнеплод здоровый) равна $q = 1 - p$. При равенстве вероятностей событий $p = q = 0.5$ большинство проб (вариант) будет иметь около половины возможных событий (поровну больных и здоровых корнеплодов); распределение примет симметричную форму (рис. 3.8). В случае неравенства вероятностей наблюдается та или иная степень асимметрии распределения (Ивантер, Коросов, 2003).

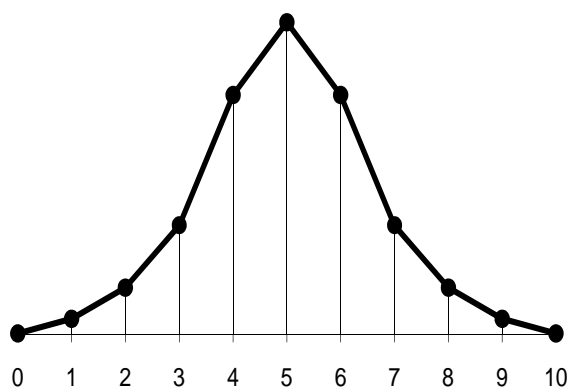


Рис. 3.8. График биномиального распределения (при $p = q = 0.5$)

Примерами описания признаков с помощью биномиального распределения могут служить поражения глистными инвазиями рыб, пелорическая форма цветка в популяциях львиного зева (Плохинский, 1970); число поврежденных участков на листьях, число волосков на единице площади шкурки, количество лучей в плавниках рыб, число хвостовых щитков у рептилий, плодовитость (размер выводка) самок (Ивантер, Коросов, 2003); число листочков околоцветника у *Anemone nemorosa* L. (Шмидт, 1984); типичная область применения в экологии – описание однородности сообщества по встречаемости видов (Пузаченко, 2004).

3.6. Другие типы теоретических распределений

Закон Пуассона описывает редкие события (признаки), происходящие 1, 2, 3 и т. д. раз на сотни и тысячи обычных событий. Другими словами, события, вероятность p которых крайне мала, а q , соответственно, приближается к единице. Таким образом, распределение Пуассона можно рассматривать как предельный случай биномиального распределения. При $p = q$ биномиальная кривая строго симметрична, при значительном уменьшении p биномиальная кривая становится асимметричной (рис 3.2).

В биологии и экологии закону Пуассона удовлетворяют редко наблюдаемые явления: частота нарушений хромосомного аппарата на каждую тысячу митозов, встречаемость семян сорняка в большой серии навесок семян культурного растения, число повторных попаданий животных в ловушки, встречаемость животных на отрезках длинных маршрутов (или на пробных площадках обширной территории), отловы животных в отдельные

промежутки времени при длительных наблюдениях (Ивантер, Коросов, 2003); явление полиэмбрионии в семенных растениях, частота рождения троен и четверен у человека, частота островков Лангерганса в тканях поджелудочной железы (Рокицкий, 1973); численность перезимовавших клопов вредной черепашки на пробных площадках, частоты спонтанных мутаций у кишечной палочки (Лакин, 1990); число пар ветвей у береговой экологической формы *Odontites serotina* Dum. (Шмидт, 1984).

Распределение *логнормальное*, или *логарифмически нормальное*, характеризуется тем, что логарифмы исходных значений выборки образуют правильное нормальное распределение; распределение же исходных значений, как правило, умеренно смещено в правую сторону вариационной кривой (рис. 3.2). Эта модель подходит для описания таких показателей, как концентрации веществ в различных средах, гидрохимические, физиологические и биохимические характеристики (Ивантер, Коросов, 2003); данной моделью удачно может быть описано распределение атмосферного и почвенного загрязнения (Пузаченко, 2004), а также распределение численности и биомассы бентосных организмов (Шитиков и др., 2003).

Равномерное распределение характеризуется одинаковой частотой встречаемости всех значений дискретного признака ($p = q$ для двух классов или $p_1 = p_2 = \dots = p_j \dots = p_k$ для нескольких классов) (рис. 3.2). Такой тип распределения можно использовать при анализе частот генов и фенотипов в популяциях, при подсчете тест-организмов, выживших в токсикологическом эксперименте (Ивантер, Коросов, 2003).

Гамма-распределение используется для описания распределения атмосферных осадков, аэрозолей, химических веществ в почве, стоке, численности некоторых видов норных животных (Пузаченко, 2004).

После описания некоторых законов распределения осталось понять, зачем биологу и экологу необходимо знать, какому закону соответствует распределение изучаемых или контролируемых признаков и показателей в процессе исследования. Эта задача внешне кажется вспомогательной, поскольку само по себе оценивание закона распределения не имеет большого практического смысла. С другой стороны, эта операция может дать исследователю некоторую важную информацию о состоянии экосистемы

или популяции, об экологии вида, о действии в экосистеме некоторых экстремальных факторов, указать на определенную тенденцию в направлении естественного отбора. Например, сильный положительный эксцесс в распределении жизненно важных признаков популяции может указывать на ужесточение стабилизирующего отбора, а асимметричное отклонение от нормального распределения – на смену стабилизирующего отбора на движущий (Ивантер, Коросов, 2003). Кроме того, резко асимметричное распределение какого-либо признака может свидетельствовать о влиянии на признак неизвестного лимитирующего фактора, приводящего к подобному смещению. При этом сделать вывод о том, что это за фактор из анализа закона распределения будет, конечно, проблематично. Поэтому подобные упражнения могут сильно напоминать «гадания на кофейной гуще». Более важным приложением практических навыков исследователя определять тип распределения является корректное применение большинства методов математической статистики. Зная тип распределения, можно воспользоваться разработанными специально для него приемами математической обработки и получить максимальную, а главное, достоверную информацию о явлении, сделать более точный прогноз, правильно оценить различия между параметрами разных выборок. В большинстве случаев исследователю перед применением конкретного метода математической обработки данных достаточно ответить на вопрос: отличается ли распределение изучаемого показателя от нормального теоретического (в случае дискретных признаков от биномиального) или нет? Если распределение нормальное или близкое к нормальному, то необходимо применять точные и высокоэффективные параметрические методы. Если распределение сильно отклоняется от нормального, то пользоваться параметрическими методами неправомерно. В этом случае корректным будет использование непараметрических методов анализа.

Примечание. Используя параметрические статистические методы для описания непрерывных признаков, нужно быть уверенным, что они действительно подчиняются нормальному закону, а в случае дискретных признаков – биномиальному. В дальнейшем для простоты изложения, подразумевая сказанное, мы будем говорить лишь о проверке нормальности распределения изучаемых показателей.

Глава 4. Статистические оценки генеральных параметров, или насколько точно данные выборки соответствуют реальности

В этой главе читатель познакомится с тем, каким образом можно оценивать статистические параметры генеральной совокупности, если в распоряжении исследователя имеются рассчитанные статистические характеристики выборки, извлеченной из этой генеральной совокупности.

Напомним, что основной целью выборочного метода исследования является оценка генеральной совокупности на основе известных характеристик выборки, представляющей лишь часть этой совокупности. Любого исследователя в первую очередь будет интересовать, к примеру, средняя численность (биомасса) вида или возрастная структура популяции не в изученных на пробных площадках выборках или в отобранной из озера пробе, а во всей популяции вида или в целом для всего озера. Числовые показатели, характеризующие генеральную совокупность, называют *генеральными параметрами*, а статистические показатели, характеризующие выборку, – *выборочными характеристиками*. Из этого разделения становится понятным, почему для одних и тех же статистических показателей применяются разные буквенные обозначения. Так, если средняя арифметическая рассчитана на основе данных генеральной совокупности, то её обозначают буквой μ , если на основе выборки – \bar{x} или M ; генеральная дисперсия – это σ^2 , выборочная дисперсия – S^2 ; генеральное стандартное отклонение – это σ , выборочное стандартное отклонение обозначается как S . Различия в обозначениях подчеркивают тот факт, что выборочные характеристики, как величины случайные, не совпадают полностью с генеральными параметрами, а с некоторой степенью отклонения в ту или иную сторону варьируют вокруг них. Исследователя же всегда интересуют истинные параметры генеральной совокупности как основного объекта исследования, но в непосредственном распоряжении имеются, как правило, лишь выборочные характеристики. Поэтому часто и возникает задача каким-то образом оценить точность,

с какой случайная выборочная характеристика соответствует истинному генеральному параметру.

Для понимания того, почему выборочные характеристики являются лишь приближенными оценками генеральных параметров, обратимся к понятию «*стандартная ошибка*».

4.1. Стандартная ошибка среднего значения

Стандартная ошибка среднего является интервальной оценкой («от – до» или \pm) генерального среднего значения и рассчитывается на основе известного выборочного среднего. Вначале приведем определение: *стандартная ошибка (средняя, среднеквадратическая, статистическая ошибка; ошибка репрезентативности, ошибка выборочности)* – это средняя величина отклонения выборочной характеристики от её генерального параметра. Наиболее часто в биологии и экологии используется понятие «стандартная ошибка среднего значения».

Следует сразу оговориться, что стандартная ошибка не является погрешностью (ошибкой измерения), а имеет совершенно другую природу. Чтобы понять, каким образом возникает стандартная ошибка, рассмотрим простой пример выборочного исследования, который, однако, в реальности неосуществим, поскольку в действительных условиях применения выборочного метода объем генеральной совокупности бывает подавляюще велик и теоретически стремится к бесконечности. Допустим, что интересующая исследователя генеральная совокупность состоит всего из 5 вариантов, представляющих собой, к примеру, длину тела какого-либо вида животного (см):

8 16 20 24 32

Таким образом, объем генеральной совокупности (N) будет равен 5, значит генеральная средняя будет равна:

$$\mu = \frac{8+16+20+24+32}{5} = 20 \text{ см.}$$

Очевидно, что в реальности исследователь никогда не знает ни объема генеральной совокупности, ни генеральных пара-

метров. Предположим, что мы захотели оценить среднюю длину тела животного, используя обычный для этих целей выборочный метод. Запланированный нами объем выборки должен составить $n = 4$. Поскольку выборка должна быть взята случайным образом, это означает, что есть равная вероятность всех вариантов генеральной совокупности попасть в состав этой выборки. Другими словами, каждая из этих 5 вариантов генеральной совокупности может попасть в выборку в составе ещё 3-х равновозможных вариантов, в итоге возможны следующие равновероятные сочетания:

8 16 20 24 8 16 20 32 8 16 24 32 8 20 24 32 16 20 24 32

Если выборка составляется случайным способом, значит каждое из пяти сочетаний вариантов с равной вероятностью может быть отобрано исследователем для установления выборочного среднего значения длины тела. Это означает, что в наших расчетах необходимо учитывать все равновероятные сочетания, не пропуская ни одного из них.

Теперь рассчитаем все выборочные средние значения, которые мог бы получить исследователь при случайном отборе:

$$\bar{X}_1 = \frac{8+16+20+24}{4} = 17 \text{ см}$$

$$\bar{X}_2 = \frac{8+16+20+32}{4} = 19 \text{ см}$$

$$\bar{X}_3 = \frac{8+16+24+32}{4} = 20 \text{ см}$$

$$\bar{X}_4 = \frac{8+20+24+32}{4} = 21 \text{ см}$$

$$\bar{X}_5 = \frac{16+20+24+32}{4} = 23 \text{ см}$$

Каждое из рассчитанных выборочных средних с некоторым приближением и характеризовало бы неизвестное исследователю генеральное среднее значение признака, причем, принимая одно из этих выборочных средних за генеральную среднюю, исследователь в каждом случае допускал бы некоторую неизбежную ошибку.

Воспользуемся тем, что в нашем числовом примере это генеральное среднее уже известно ($\mu = 20$), и выясним величину такой ошибки в применении ко всем пяти выборочным средним. Так, если бы исследователь имел дело с первым выборочным средним ($\bar{X}_1 = 17$), принимая его за генеральное среднее ($\mu = 20$), он сделал бы ошибку, равную разности:

$$\bar{X}_1 - \mu = 17 - 20 = -3, \text{ т. е.}$$

ошибся бы на 3 см в сторону приуменьшения длины тела. Расчет остальных отклонений дает следующие ошибки: -1; 0; +1; +3.

Теперь нужно определить, какова средняя величина этих пяти равновероятных ошибок. Другими словами, насколько в среднем мог бы ошибиться исследователь, применяя выборочный метод. Как и в случае расчета дисперсии и стандартного отклонения (см. главу 2), чтобы избавиться от отрицательных значений, возведем вначале каждую ошибку в квадрат, затем, суммировав получившиеся значения и поделив их на число слагаемых (5), произведем обратную арифметическую операцию – извлечение квадратного корня:

-3	-1	0	+1	+3
9	1	0	1	9

$$9 + 1 + 0 + 1 + 9 = 20$$

$$\sqrt{\frac{20}{5}} = \sqrt{4} = \pm 2 \text{ см}$$

Эта средняя величина пяти равновероятных ошибок в нашем примере и есть стандартная ошибка среднего значения длины тела животного. Итак, стандартная ошибка показывает, насколько в среднем (но не максимально!!!) рискует ошибиться исследователь, принимая одно из случайных выборочных средних за неизвестное ему генеральное среднее. В нашем примере, следовательно, принимая за среднюю длину тела всех 5 животных (генеральная совокупность) средний размер тела только 4 животных (выборка), исследователь в среднем рискует ошибиться на величину ± 2 см.

Вычисленные значения ошибок подставляют к соответствующим выборочным характеристикам со знаками «плюс – минус»

(характеристика \pm ошибка) и в такой форме представляют в научных отчетах и публикациях. Если бы исследователю случайным образом попала первая выборка, то стандартная запись полученной ошибки выборочности имела бы вид:

$$\bar{X}_1 = 17 \pm 2 \text{ см.}$$

Из рассмотренного гипотетического примера вытекают два важных вывода:

1. Стандартная ошибка по своей природе является не ошибкой измерения, а статистической ошибкой, неизбежно возникающей при отборе выборок из генеральной совокупности и, соответственно, связанной с перенесением результатов, полученных при изучении выборки, на всю генеральную совокупность. При этом очевидно, что ошибки измерения могут увеличивать стандартную ошибку. Также следует понимать, что определять величину ошибок репрезентативности требуется только для выборочных характеристик, генеральные параметры не имеют стандартных ошибок.

2. Расчет стандартной ошибки фактически совпадает с вычислением стандартного отклонения, произведенного для выборки в главе 2. Поэтому стандартная ошибка не что иное, как стандартное отклонение множества случайных выборочных средних от истинной генеральной средней.

На практике обычно нет возможности делать несколько выборок и вычислять несколько выборочных средних, чтобы по ним проводить расчеты. Статистическая теория показывает, что стандартная ошибка среднего значения в \sqrt{n} раз меньше, чем стандартное отклонение. Поэтому ошибку можно рассчитать для единичной отдельной выборки (на основе всего одного выборочного среднего значения) по формуле:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}},$$

где S_x – стандартное отклонение,

n – объем выборки.

Другим часто используемым обозначением стандартной ошибки, кроме $S_{\bar{x}}$, является буква m .

Примечание 1. Полезно знать, что приведенная формула всегда даёт значения стандартной ошибки, несколько завышенные по сравнению с действительными, поскольку в расчетах используется выборочное стандартное отклонение, а не истинное генеральное. Данная неточность при расчетах стандартной ошибки считается допустимой, поскольку в статистике из альтернативы – преувеличение или преуменьшение ошибки – именно первое является менее опасным.

Формула стандартной ошибки показывает, что величина ошибки тем больше, чем больше варьирование признака (S_x) и чем меньше выборка (n). При увеличении объема выборки ошибки репрезентативности стремятся к нулю, поэтому при планировании научных работ предполагаемый объем собираемого материала всегда является важным критерием адекватности тех данных, которые будут получены в процессе исследования. При всей неизбежности статистической ошибки она может быть сведена к минимуму отбором достаточного числа особей (вариант).

Таким образом, стандартная ошибка указывает на точность, с какой выборочный показатель характеризует генеральный параметр. Чем меньше ошибка, тем ближе выборочная характеристика к величине генерального параметра, и, наоборот, чем больше ошибка, тем менее точно выборочная характеристика определяет генеральный параметр, и значит пользоваться подобными данными необходимо с особой осторожностью.

4.2. Доверительный интервал для среднего значения

Стандартная ошибка характеризует лишь *средние* пределы варьирования выборочных средних около истинного генерального среднего значения. Если средняя ошибка оказывается равной 2 см, как в нашем предыдущем примере, это свидетельствует лишь о том, что некоторые из выборочных средних будут отклоняться от генерального среднего меньше чем на 2 см, а другие, наоборот, могут отстоять от него больше чем на 2 см, вследствие чего разница в 2 см и является *средней* характеристикой всех возможных отклонений этих выборочных средних от их общего генерального среднего. Однако часто исследователя может интересовать не столько средняя величина этих разностей,

сколько предельно возможное отклонение выборочных средних от генеральной средней, другими словами, не средняя, а максимальная ошибка.

Определить максимальное отклонение выборочной средней от истинной генеральной средней можно лишь с определенной вероятностью. В случае нормального распределения изучаемого признака это можно сделать на основе правила 3 сигм. Рассмотрим следующий пример: пусть имеется бесконечно большая генеральная совокупность, например популяция какого-либо вида животного. Отберем из этой популяции 10 000 выборок, в каждой из которых будет по 100 особей, и измерим у всех особей длину тела. В итоге мы можем рассчитать 10 000 выборочных средних значений длины тела. Если признак соответствует нормальному закону, то распределение этих выборочных средних на графике примет уже известный читателю «колоколообразный» вид. В центре данного распределения будет находиться истинная генеральная средняя (μ), от неё влево и вправо будут отклоняться выборочные средние значения ($\bar{X}_1 \dots \bar{X}_{10000}$), как характеристики, случайно варьирующие около μ (рис. 4.1).

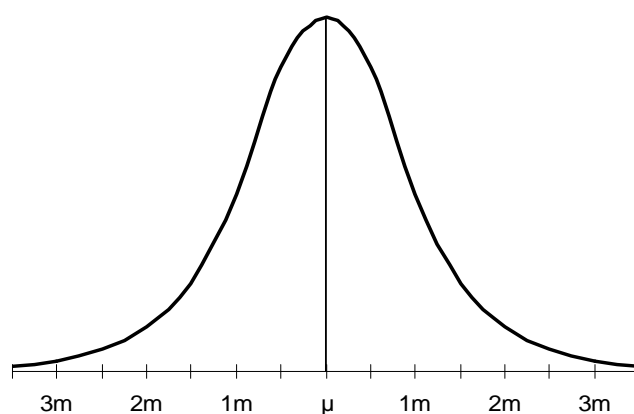


Рис. 4.1. График, иллюстрирующий отклонения выборочных средних значений от истинного генерального среднего

Причем можно установить с определенной вероятностью границы, в пределах которых будут происходить подобные отклонения. Читатель уже знаком с подобными вероятностями и границами. Так, например, с вероятностью 0.683 выборочные средние будут отклоняться от истинного генерального значения в пределах $\mu \pm 1m$ (одна стандартная ошибка), а с вероятностью

0.997 в пределах $\mu \pm 3m$ (три стандартные ошибки). В данном случае используется не сигма (стандартное отклонение), а стандартная ошибка, поскольку именно она, как было показано выше, является стандартным отклонением, характеризующим разброс выборочных средних относительно генерального среднего. *Верно и обратное заключение*: можно утверждать, что с вероятностью 0.997 истинная генеральная средняя окажется внутри интервала $\bar{x} \pm 3m$. Таким образом, мы подошли к понятию доверительного интервала. Приведем определение.

Доверительный интервал – границы, в которых с заданной вероятностью (степенью достоверности) находится изучаемый генеральный параметр. В экологии и биологии наиболее часто используется доверительный интервал для среднего значения.

При расчетах (построениях) доверительных интервалов, однако, не применяются вероятности, которыми мы оперировали в правиле трех сигм ($P = 0.683; 0.954; 0.997$). Напомню, что разным значениям $\pm t$ соответствуют строго определенные вероятности, так:

$$t = 1 \Rightarrow P = 0.683; t = 2 \Rightarrow P = 0.954; t = 3 \Rightarrow P = 0.997$$

В случае с доверительными интервалами важнейшее значение имеет так называемое «соглашение о 95%-й вероятности». В соответствии с ним совокупности, состоящей из 95% особей (объектов), мы доверяем так же, как и 100%-й. Другими словами, данная вероятность принята как наименьшая, которой можно доверять как 100%-й при принятии того или иного решения, связанного с математической обработкой данных. Поэтому подобная вероятность получила обозначение «*доверительная вероятность*» – вероятность, признанная достаточной для уверенного суждения о генеральных параметрах на основании известных выборочных характеристик. Применительно к доверительному интервалу это *вероятность того, что генеральный параметр (среднее значение) действительно окажется внутри доверительного интервала*. Если вероятность 0.95 является наименьшей в рейтинге доверия, то, значит, существуют и другие доверительные вероятности. Действительно, если решение, которое нужно принять при математической обработке данных, является очень ответственным, то его

стараясь принимать с ещё большей вероятностью, к примеру 0.99 или 0.999, чтобы свести возможные ошибки практически к нулю. Все эти три вероятности относятся к доверительным, и именно они используются при построении доверительных интервалов. Какие же значения t соответствуют этим вероятностям? Ответ читатель найдет ниже:

$P = 0.683$	$t = 1 \Rightarrow$	$P = 0.95$	$t = 1.96$
$P = 0.954$	$t = 2 \Rightarrow$	$P = 0.99$	$t = 2.58$
$P = 0.997$	$t = 3 \Rightarrow$	$P = 0.999$	$t = 3.29$

Теперь настало время ввести ещё одно важное понятие в математической статистике, его обозначают как *уровень значимости* (p или α). Данное понятие имеет много определений, в общем виде это вероятность допустить ошибку, принимая то или иное решение, связанное с математической обработкой данных, или вероятность, противоположная доверительной. Более конкретное определение применительно к доверительному интервалу – *это вероятность того, что генеральный параметр (среднее значение) при заданной доверительной вероятности ($P=0.95$, $P=0.99$, $P=0.999$) окажется за границами доверительного интервала.*

В статистике приняты 3 уровня значимости, соответствующие доверительным вероятностям:

$P = 0.95$	\Rightarrow	$p = 0.05 (1-0.95)$
$P = 0.99$	\Rightarrow	$p = 0.01 (1-0.99)$
$P = 0.999$	\Rightarrow	$p = 0.001 (1-0.999)$

Отсюда следует, что при $p = 0.05$ риск ошибиться составляет 1 раз на 20 случаев (5%), при $p = 0.01$ – 1 раз на 100 случаев (1%), при $p = 0.001$ – 1 раз на 1000 случаев (0.1%). Таким образом, чем меньше уровень значимости и, соответственно, выше доверительная вероятность, тем меньше риск ошибки.

Доверительный интервал для оценки генерального среднего значения можно рассчитать, исходя из 3-х параметров:

1. Выборочное среднее значение – \bar{X} .

2. Стандартная ошибка выборочного среднего (m) в данном случае является стандартным отклонением выборочных средних от генеральной средней.

3. Нормированное отклонение (t) необходимо для установления доверительной вероятности, с которой будет рассчитан доверительный интервал.

Таким образом, генеральное среднее значение находится в интервале:

$$\bar{X} \pm t \cdot m$$

В зависимости от заданной доверительной вероятности можно рассчитать следующие виды доверительных интервалов:

95%-й доверительный интервал	$\bar{X} \pm 1.96 \cdot m$
99%-й доверительный интервал	$\bar{X} \pm 2.58 \cdot m$
99.9%-й доверительный интервал	$\bar{X} \pm 3.29 \cdot m$

Примечание 2. Если выборка невелика ($n < 20$), то необходимо вводить поправки на объем выборки, расширяя область возможного пребывания генерального параметра. Это понятно, поскольку при дефиците информации любые заключения не могут быть очень точными. Фактически в этих случаях поправки представляют собой более высокие значения t , чем те, которые были представлены. Так, если объем выборки равен $n = 4$, то для расчета 95%-го доверительного интервала необходимо взять значение t , равное 3.18, а не 1.96. Данные поправки читатель при желании может найти в любом пособии по статистике в виде так называемой таблицы критических точек t -критерия Стьюдента. Однако если расчеты производятся с использованием статистических программ, то в поисках данной таблицы нет необходимости.

Биологическая (экологическая) интерпретация доверительного интервала довольно проста: в примере с расчетом стандартной ошибки для длины тела «неизвестного» вида животного первая выборочная средняя оказалась равна 17 см, рассчитанный для неё 95%-й доверительный интервал составляет ± 10.9 см. Это означает, что с вероятностью 0.95 истинная генеральная средняя длины тела находится внутри границ от 6.1 см до 27.9 см (т. е. 17 ± 10.9 см), но с вероятностью 0.05 средняя генеральной сово-

купности может находиться вне границ данного интервала. Генеральная средняя в данном примере была равна 20 см и действительно попадает в рассчитанный интервал, а большие размеры интервала, как было сказано, объясняются малочисленностью выборки ($n = 4$).

Программное обеспечение. Расчет стандартной ошибки и доверительного интервала для среднего значения проводится в известном читателю модуле «*Описательная статистика*», который имеется в большинстве статистических программ (см. выше). Кроме того, в MS EXCEL можно воспользоваться функцией **ДОВЕРИТ**. Часто удобно представлять интервальные оценки генеральных параметров графически в виде отрезков, откладываемых от среднего значения в обе стороны (\pm). Подобные графики можно назвать *диаграммами размаха*, в программе STATISTICA они обозначаются как «ящики с усами» (Box & Whisker plot) (рис. 4.2).

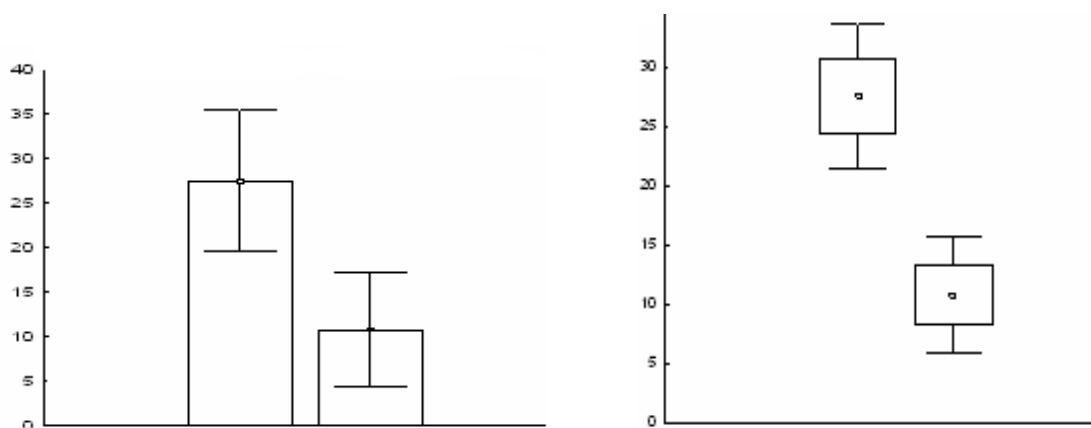


Рис. 4.2. Диаграммы размаха, визуализирующие интервальные оценки генеральных параметров.

На левом графике: столбец – выборочное среднее, отрезок – стандартная ошибка либо доверительный интервал. На правом графике: точка – выборочное среднее, прямоугольник – стандартная ошибка, отрезок – доверительный интервал

Примечание 3. Выбор доверительной вероятности осуществляется исследователем в зависимости от той ответственности, с какой должны быть сделаны выводы о генеральном параметре. Если, к примеру, исследование связано с токсичностью вещества или с дозами лекарственных препаратов, от которых зависит

жизнь пациентов, то для уверенного суждения о генеральных параметрах необходимо оперировать более высокими вероятностями – 0.99 или даже 0.999. С другой стороны, важно понимать, что чем выше доверительная вероятность, тем шире будет доверительный интервал и тем менее чёткой становится оценка генерального параметра. В большинстве экологических и биологических исследований достаточно надежной считается 95%-я доверительная вероятность (или 5% уровень значимости), которые и используются наиболее часто.

Примечание 4. Рассмотренный в данной главе способ расчета доверительного интервала применим лишь в случае нормальности распределения изучаемых признаков. Более широкие возможности использования доверительного интервала читатель может найти в модуле «*Описательная статистика*» программы ATTESTAT.

Глава 5. Проверка статистических гипотез

В данной главе будут обсуждаться *сравнительные* оценки генеральных параметров, т. е. методы, позволяющие установить, насколько различия между выборками «правильно» отражают различия между генеральными совокупностями, из которых они взяты.

В биологических и экологических исследованиях анализ отдельных выборок редко является конечной целью. Очень часто приходится сравнивать эти выборки между собой. Метод сравнения является наиболее общим способом эмпирического познания и используется для получения информации об объекте исследования как в естественных, так и в гуманитарных науках. Ни одно исследование биологов и экологов не обходится без сравнения, сравнивать приходится данные опытной и контрольной групп (выборок) в эксперименте, показатели качества воды за разные промежутки времени, степень загрязнения тех или иных участков, популяции по численности и структуре, продуктивность разных озер, морфофизиологические особенности разных групп людей и животных и т. д. При этом сравнение двух выборок не может быть самоцелью ни биологического, ни экологического исследования, поскольку современную науку интересуют не просто факты, но причина их возникновения. В этом ключе сравнение двух выборок выступает в роли метода

поиска отличий в причинах, обеспечивших существование двух групп объектов (выборок) разного качества; в конце концов, это поиск влияния фактора, поиск закономерности, отделение её от случайности (Ивантер, Коросов, 2005).

При сравнении двух выборок всегда возникает вопрос, достоверны ли наблюдаемые отличия между выборками или они обусловлены лишь какими-то случайными причинами? Другими словами, можно ли данное различие считать закономерным, *характерным для всей генеральной совокупности* и рассматривать его как результат реально действующих в системе факторов или же оно случайно и является следствием недостаточного количества данных и в следующих опытах (наблюдениях) может не проявиться? Поясним сказанное на примере. Вернемся к росту детей дошкольного возраста, в главе 2 мы установили, что в полученных выборках девочки в среднем выше мальчиков, причем выборочная разность роста составила **6 см** (средний рост девочек **127 см**, а мальчиков – **121 см**). Можно ли сразу утверждать, что и в генеральных совокупностях, из которых извлечены 2 выборки детей, наблюдается такое же различие между генеральными средними значениями (т. е. теми, которые можно было бы получить, изучив всех детей данного возраста)? Ответ очевиден – нет! В предыдущем разделе было показано, что любые выборочные средние являются величинами случайными, отклоняющимися от истинных генеральных средних и поэтому с ними не совпадающими. Эту среднюю величину отклонения оценивает стандартная ошибка. Фактически это означает, что в «реальности» средний рост мальчиков и девочек может быть равным (например, и мальчики и девочки могут иметь средний рост $\mu_1 = \mu_2 = 125$ см) и только в силу случайности, возникающей при отборе выборок из генеральных совокупностей, полученные исследователем выборочные средние ($\bar{X}_1 = 127$ см и $\bar{X}_2 = 121$ см) могут отличаться друг от друга, т. е. их разность может быть недостоверна. Но, с другой стороны, поскольку исследователь никогда не знает истинных генеральных средних, полученная разность между выборочными средними вполне может оказаться не случайной, а закономерной, т. е. реально существующей в генеральной совокупности. Приведенный пример показывает, что при сравнении любых выборок исследователю всегда необходимо каким-то образом

устанавливать достоверность наблюдаемых различий, для того чтобы подтвердить реальность их существования.

Прежде чем приступить к практическому освоению методов, позволяющих устанавливать достоверность выборочной разности, остановимся на основополагающих понятиях.

5.1. Достоверность выборочной разности. Нулевая и альтернативная гипотезы. Понятие критерия достоверности

Достоверность – это свойство выборочной разности (различие средних, дисперсий 2-х выборок) правильно с заданной вероятностью отражать генеральную разность (различие генеральных средних и дисперсий). Выборочная разность может быть *достоверна (статистически значима)* или *недостоверна (случайна, статистически незначима)*.

Выборочная разность достоверна – это означает, что если в выборочном исследовании зафиксировано различие выборочных характеристик (средних значений, дисперсий), то точно такое же различие наблюдается между соответствующими генеральными параметрами в генеральных совокупностях, из которых извлечены выборки.

Если получена *недостоверная выборочная разность*, это значит не получено **никакого определенного ответа о разности** между соответствующими генеральными параметрами в генеральных совокупностях, из которых извлечены выборки. Другими словами, ничего нельзя заключить с заданной вероятностью о генеральной разности – ни что она есть, ни что её нет, т. е. разница остаётся статистически недоказанной.

Примечание 1. Распространенная ошибка среди исследователей – это неправильная интерпретация *недостоверности* различий выборочных характеристик: наличие между выборками недостоверной разности не свидетельствует об отсутствии разности между соответствующими генеральными параметрами, фактически отсутствие различий в генеральных совокупностях доказать невозможно.

Примечание 2. В современной литературе по математической статистике вместо термина «достоверный» при проверке статистических гипотез рекомендуется использовать слово-

сочетание «статистически значимый», особенно в научных публикациях. В данном пособии применяются оба термина.

Для установления того, достоверна или недостоверна выборочная разность, исследователь вначале должен сформулировать 2 противоположные статистические гипотезы:

1. *Нулевая гипотеза* (H_0) – различия между выборочными характеристиками случайны, недостоверны.

2. *Альтернативная гипотеза* (H_a) – различия между выборочными характеристиками достоверны, т. е. реально наблюдаются между генеральными параметрами в генеральных совокупностях, из которых извлечены выборки.

Для отклонения или принятия той или иной гипотезы применяются так называемые *критерии достоверности* – специально разработанные статистические показатели с известными функциями распределения, позволяющие с заданной доверительной вероятностью проверять истинность нулевой или альтернативной гипотез.

Рассмотрим упрощенную схему проверки истинности статистических гипотез критериями достоверности (рис. 5.1).

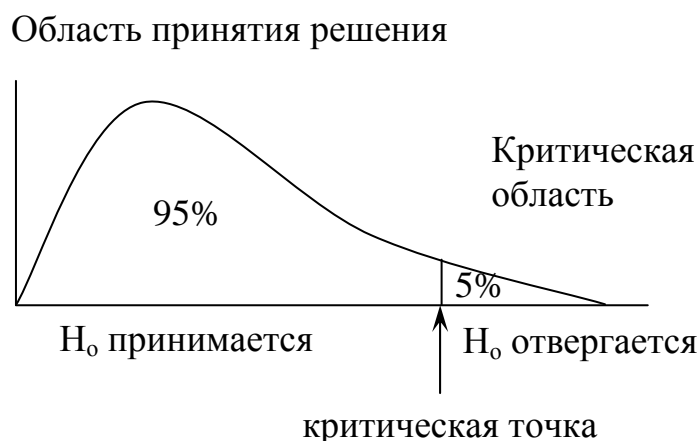


Рис. 5.1. Схема использования критериев достоверности

Разместив в представленной схеме значения критерия достоверности, можно заметить, что при проверке нулевой гипотезы они разбиваются некоей критической точкой на 2 подмножества: одно из подмножеств содержит значения критерия достоверности, при которых H_0 отвергается (критическая область), а другое – при которых H_0 принимается (область принятия решения) (рис. 5.1).

Критическую точку, таким образом, можно определить как числовое значение критерия достоверности, при достижении или превышении которого нулевая гипотеза отвергается, т. е. доказываемая достоверность выборочной разности. Критические точки (значения) для каждого критерия табулированы, т. е. сведены в специальные таблицы, которыми до сих пор пользуются многие исследователи. На основе сравниваемых выборок можно рассчитать так называемое фактическое (достигнутое, эмпирическое) значение критерия (Φ_3), которое и сравнивают с критической точкой (значением) (K_3) по определенному правилу, отраженному на схеме:

Если $\Phi_3 \geq K_3$, то H_0 отвергается, доказываемая достоверность разности.
Если $\Phi_3 < K_3$, то H_0 принимается, доказываемая недостоверность разности.

Мы будем пользоваться несколько другим подходом, более удобным при применении статистических компьютерных программ. Дело в том, что принимать или отвергать нулевую гипотезу можно лишь с заданной доверительной вероятностью: 0.95, 0.99 или 0.999. Для биологических и экологических исследований вполне достаточна 95%-я доверительная вероятность. Каждой доверительной вероятности соответствует обратная ей величина – *уровень значимости* (0.05, 0.01 и 0.001). Применительно к критериям достоверности – это вероятность ошибочно отвергнуть нулевую гипотезу, когда она на самом деле верна (т. е. когда гипотеза верна, но отклоняется исследователем). Более простое определение *p-уровня значимости* – это вероятность справедливости нулевой гипотезы. Смысл 2-го подхода к оценке истинности статистических гипотез заключается в том, что для каждого фактического значения критерия достоверности с учетом объема выборки можно рассчитать уровень значимости. Это означает, что можно перейти от сравнения фактических значений критерия с критической точкой к сравнению фактического уровня значимости с критическим уровнем, что является более удобной процедурой при обработке данных на компьютере. В этом случае вероятность 5% (или 0.05) является принятым научным сообществом верхним «порогом ошибки» при проверке статистических гипотез (критический уровень). Если факти-

ческий p -уровень значимости (он рассчитывается статистической программой на основе сравниваемых выборок) превышает критический p -уровень (обычно 0.05), то вероятность ошибиться, отвергая нулевую гипотезу, считается высокой (больше 0.05), что является основанием не отклонять нулевую гипотезу. И наоборот, если получается низкая вероятность ошибиться, отвергая нулевую гипотезу (фактический p -уровень ≤ 0.05), то, значит, есть все основания отвергнуть нулевую гипотезу. Другими словами, выборочная разность может считаться достоверной только с вероятностью ≥ 0.95 (при $p \leq 0.05$), если вероятность этого очень высокая, но равна, допустим, 0.94 ($p = 0.06$), то различие между выборочными характеристиками следует признать недостоверным. Подобное жесткое соглашение о 95% является необходимым условием при математической обработке данных, поскольку исследователь имеет дело со случайными явлениями. Исходя из этих рассуждений все статистические сравнения сводятся к вычислению фактического уровня значимости, что легко можно делать, используя статистические программы и следующее условие:

Если $p \leq 0.05$, то H_0 отвергается, доказывается достоверность разности.
Если $p > 0.05$, то H_0 принимается, доказывается недостоверность разности.

Примечание 3. В качестве критического уровня значимости исследователь в зависимости от той ответственности, с которой необходимо принять решение, может выбирать другие, более низкие вероятности – 0.01 или 0.001. Таким образом, процесс проверки истинности статистических гипотез всегда относителен. Допустим, если достигнутый уровень значимости равен 0.04, то при критическом p -уровне, равном 0.05, исследователь должен отвергнуть H_0 , а если выбран критический p -уровень 0.01, то необходимо принять H_0 .

От общетеоретических рассуждений перейдем непосредственно к практическим задачам, решаемым при проверке статистических гипотез. Использование критериев достоверности позволяет решать 3 основные задачи:

1. Осуществлять оценку соответствия между эмпирическим распределением изучаемого показателя и известным теоретическим распределением – *подгонка распределения*.

2. Производить сравнение двух выборок и определение достоверности различий средних величин и дисперсий этих выборок.

3. Устанавливать принадлежность отдельного значения изучаемой переменной к выборке (генеральной совокупности) – *браковка выбросов или выскакивающих значений*.

5.2. Классификация критериев достоверности

В зависимости от типа сравнения:

1. Критерии, определяющие степень соответствия эмпирического распределения известному теоретическому, называются *критериями согласия*.

2. Критерии, определяющие достоверность различий 2-х выборок по определенным характеристикам (среднее или дисперсия), называются *критериями различия*.

3. Критерии, определяющие принадлежность выбросов к выборке (генеральной совокупности), называются *критериями исключения*.

В зависимости от способа расчета критерия:

1. *Параметрические критерии* – рассчитываются на основе параметров выборочной совокупности (на основе среднего значения, дисперсии, стандартной ошибки и т. д.).

• *T-критерий Стьюдента (t-test)* – основан на выборочном среднем значении. Если возникает задача сравнить две выборки по средним значениям, то фактическое значение критерия Стьюдента рассчитывается в классическом варианте как отношение разности 2-х выборочных средних к стандартной ошибке этой разности:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}.$$

При этом совершенно очевидно, что чем больше разность между средними значениями 2-х выборок и чем меньше стандартная ошибка этой разности, тем больше вероятность того, что 2 выборки достоверно различаются между собой по средней тенденции. Как отмечалось выше, для каждого подобного фактического значения критерия можно рассчитать фактический р-уровень значимости, сравнить его с критическим уровнем (0.05) и определить достоверность различий. Подобная схема принятия решения сохраняется для всех других критериев достоверности.

• *F-критерий Фишера (F-test)* – основан на выборочной дисперсии. Фактическое значение F-критерия рассчитывается как отношение большей выборочной дисперсии к меньшей:

$$F = \frac{S_1^2}{S_2^2}.$$

2. *Непараметрические критерии* – рассчитываются на основе частоты встречаемости или рангов.

• *Частотные критерии:*

Критерий χ^2 («хи квадрат»), или критерий согласия Пирсона, (Chi-Square test) – представляет собой сумму квадратов отклонений эмпирических частот (f) от вычисленных теоретических частот распределения (f'), отнесенную к теоретическим частотам:

$$\chi^2 = \sum_{i=1}^n \frac{(f - f')^2}{f'}.$$

• *Ранговые критерии:*

Критерий Манна – Уитни (U) (Mann – Whitney test)

Критерий Вилкоксона (T) (Wilcoxon test)

Критерий знаков (Z) (Sign test)

Критерий серий Вальда – Вольфовица (S) (Wald-Wolfowitz test)

Применение ранговых непараметрических критериев основано на ранжировании отдельных значений двух сравниваемых выборок. При этом сопоставляются не сами по себе отдельные

значения ранжированного ряда, а его порядковые номера или ранги. Алгоритмы расчета ранговых критериев отличаются простотой. Если значения разных выборок более или менее регулярно чередуются в общем ранжированном ряду, значит они распределены сходным образом и отличий между выборками нет. Если же значения выборок пересекаются не полно (перекрываются только краями распределений) или вообще не пересекаются, то очевидно, что эти выборки достоверно отличаются друг от друга (выборки со смещенными центрами или разными дисперсиями). Конструкции ранговых критериев подробно разбираются в ряде пособий по биометрии (Урбах, 1975; Малета, Тарасов, 1982; Лакин, 1990), к которым автор и направляет читателя при необходимости. Нулевая гипотеза, как правило, состоит в том, что сравниваемые выборки различаются случайно, недостоверно. В некоторых работах подход оказывается более дифференцированным: выделяют ранговые критерии, определяющие достоверность отличий выборок по центральной тенденции (среднему значению) и по разбросу значений (дисперсии) или одновременно по обоим параметрам.

Разделение критериев достоверности на параметрические и непараметрические является наиболее важным в практике биологических и экологических исследований, поэтому рассмотрим более подробно преимущества и недостатки двух групп методов.

Параметрические критерии

Преимущество – большая «мощность» по сравнению с непараметрическими критериями. Мощностью критерия – способность более безошибочно отвергнуть нулевую гипотезу, если она неверна.

Недостатки:

1. Могут применяться только при нормальном распределении сравниваемых переменных.
2. Более чувствительны к малому объему выборки, чем непараметрические критерии.
3. Используются только для количественных переменных.

Непараметрические критерии

Преимущества:

1. Могут применяться независимо от закона распределения сравниваемых переменных, т. е. в случаях, когда распределение переменной отличается от нормального.

2. Используются для количественных, порядковых и качественных данных.

3. Хорошо работают при малых объемах выборки.

Недостаток – дают более грубую оценку различий выборок по сравнению с параметрическими критериями.

В зависимости от типа выборки:

1. Критерии для независимых выборок (*t-критерий Стьюдента, критерий Манна – Уитни, критерий серий Вальда – Вольфовица*).

2. Критерии для зависимых выборок (*парный t-критерий Стьюдента, критерий Вилкоксона, критерий знаков*).

Независимые выборки – сравниваемые выборки, отдельные значения в которых никак не связаны между собой. Объем обеих выборок может как различаться, так и быть равным (т. е. $n_1 \neq n_2$ или $n_1 = n_2$).

Пример: сравнение роста или веса животных из 2-х популяций, сравнение концентраций вещества за разные годы.

Зависимые выборки – сравниваемые выборки, отдельные значения которых попарно связаны между собой. Объем обеих выборок всегда должен быть равным (т. е. $n_1 = n_2$). Преимущество подобных выборок в том, что при сравнении различия внутри выборок становятся меньшими, чем между выборками, это повышает вероятность установления достоверности выборочной разности.

Пример: сравнение силы левой и правой руки у группы испытуемых, сравнение физиологических показателей у одних и тех же животных до и после проведения опыта.

При использовании метода сравнений важно знать не только к какой группе относится и какую задачу способен решать тот или иной критерий достоверности, но и условия его применимости, соблюдение которых позволяет статистически корректно производить количественную обработку данных. Информация подобного рода кратко сведена в таблицу и будет полезна читателю (табл. 5.1).

Подробный разбор практического применения каждого из критериев не входит в задачи данного пособия, для этого можно

обратиться к разнообразной статистической литературе, указанной в конце пособия. Единственное, что осталось продемонстрировать читателю, – это некоторые примеры проверки статистических гипотез с применением специализированных программ.

Таблица 5.1

***Классификация и особенности применения
некоторых критериев достоверности***

<i>Задача</i>	<i>Критерий</i>	<i>Тип критерия</i>	<i>Условия и особенности применения критерия</i>
Доказать различие между эмпирическим и теоретическим частотными распределениями.	Критерий Пирсона хи-квадрат	Критерии согласия (нормальности)	1. Объем выборки должен быть большим ($n \geq 50$). 2. Неприменим для малых выборок ($n < 10 - 20$). 3. Необходимо, чтобы частоты значений признака в крайних классах были ≥ 5 . 4. Применяется при проверке как непрерывных, так и дискретных законов распределения.
	Критерий Колмогорова – Смирнова (d) (Kolmogorov – Smirnov test)		Обладает большей мощностью для непрерывных переменных.
	Критерий Шапиро – Уилка (W) (Shapiro-Wilk's test)		1. Считается наиболее мощным критерием, если объем выборки небольшой ($8 \leq n \leq 50$). 2. Более мощный при определении различий в асимметрии распределения, чем в эксцессе.
Доказать различие двух средних арифметических для одного признака.	t-критерий Стьюдента	Параметрический, критерий различия	1. Нормальность распределения сравниваемых переменных.

			2. Недостоверное отличие дисперсий сравниваемых переменных. Если данное условие не соблюдается, то в MS Excel и Statistica необходимо рассчитывать t-критерий Стьюдента с различающимися дисперсиями.
Доказать различие двух дисперсий для одного признака.	F-критерий Фишера	Параметрический, критерий различия	Нормальность распределения сравниваемых переменных.
Доказать различие двух выборок в целом.	Критерий Манна – Уитни	Непараметрические ранговые, для независимых выборок, критерии различия	1. Наиболее мощная непараметрическая альтернатива t-критерия Стьюдента для независимых выборок. 2. Наиболее целесообразно применять для малых выборок ($3 \leq n \leq 60$). 3. В 2-х сравниваемых выборках не должно быть совпадающих значений или таких совпадений должно быть очень мало.
	Критерий Вальда – Вольфовица		1. Применим для малых выборок. 2. В достаточно больших выборках улавливает различия практически любого типа: по центральной тенденции, по дисперсии, по характеру распределения. В выборках малого объема реагирует в основном на сдвиги по центральной тенденции.
	Критерий знаков		1. Обладает большей чувствительностью к выборкам среднего и большого объема. 2. Учитывает лишь направленность изменений в каждой паре значений признака.

		Непараметрические ранговые, для зависимых выборок, критерии различия	3. Применим к порядковым и качественным данным. 4. Если сравниваются количественные данные, то применение критерия знаков возможно только для предварительной оценки различий 2-х выборок.
	Критерий Вилкоксона		Является более мощным критерием по сравнению с критерием знаков, т. к. учитывает не только знак разности между связанными значениями 2-х выборок, но и величину этой разности.

5.3. Проверка нормальности распределения в пакете STATISTICA

При рутинных биологических и экологических исследованиях, как правило, достаточно бывает проверить предположение о соответствии эмпирического распределения изучаемых переменных теоретическому нормальному закону. Как знает читатель, от этого зависит корректный выбор либо параметрических (основанных на нормальном распределении), либо непараметрических методов анализа данных. Если интерес представляет анализ самого эмпирического распределения, то, используя *критерий согласия Пирсона*, можно с определенной вероятностью установить теоретический закон, которому соответствует распределение (подгонка распределения). Мы остановимся на первой задаче.

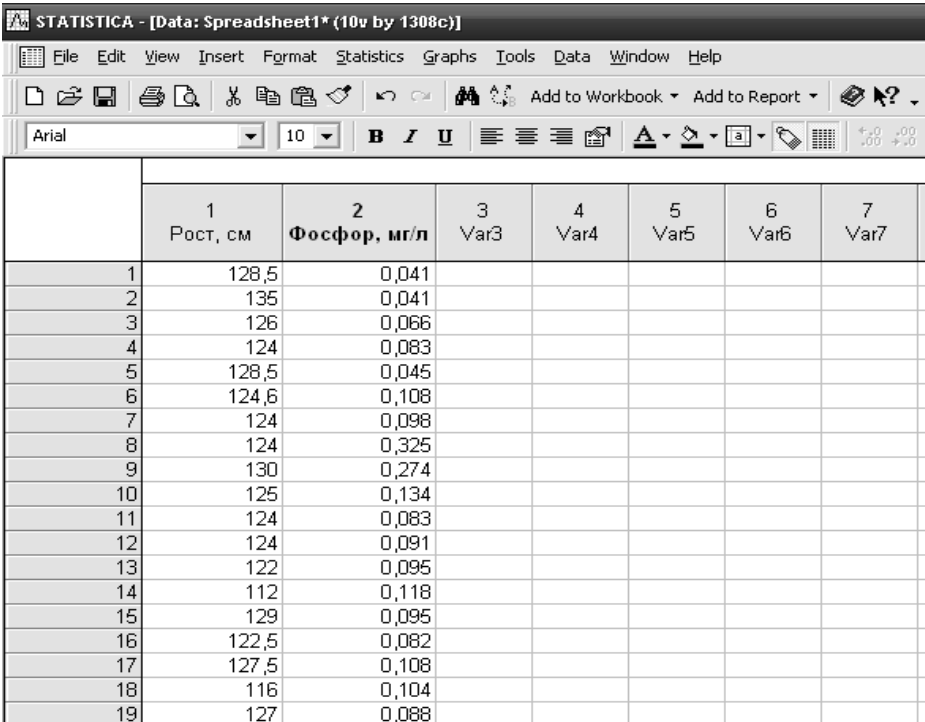
При проверке нормальности распределения необходимо сформулировать статистические гипотезы:

H_0 – эмпирическое распределение недостоверно отличается от нормального теоретического, т. е. отличия частот случайны.

H_a – эмпирическое распределение достоверно отличается от нормального теоретического.

В программе STATISTICA проверку нормальности распределения переменных можно осуществить в модуле Basic Statistics / Tables (Основные статистики / таблицы). Перед

анализом в отдельные столбцы электронной таблицы вводятся числовые значения переменных. В качестве примера проверим нормальность распределения следующих показателей: рост детей дошкольного возраста и концентрации общего фосфора в воде озера Неро (Ярославская область) (рис. 5.2).



STATISTICA - [Data: Spreadsheet1* (10x by 1308c)]

	1 Рост, см	2 Фосфор, мг/л	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7
1	128,5	0,041					
2	135	0,041					
3	126	0,066					
4	124	0,083					
5	128,5	0,045					
6	124,6	0,108					
7	124	0,098					
8	124	0,325					
9	130	0,274					
10	125	0,134					
11	124	0,083					
12	124	0,091					
13	122	0,095					
14	112	0,118					
15	129	0,095					
16	122,5	0,082					
17	127,5	0,108					
18	116	0,104					
19	127	0,088					

Рис. 5.2. Электронная таблица программы STATISTICA

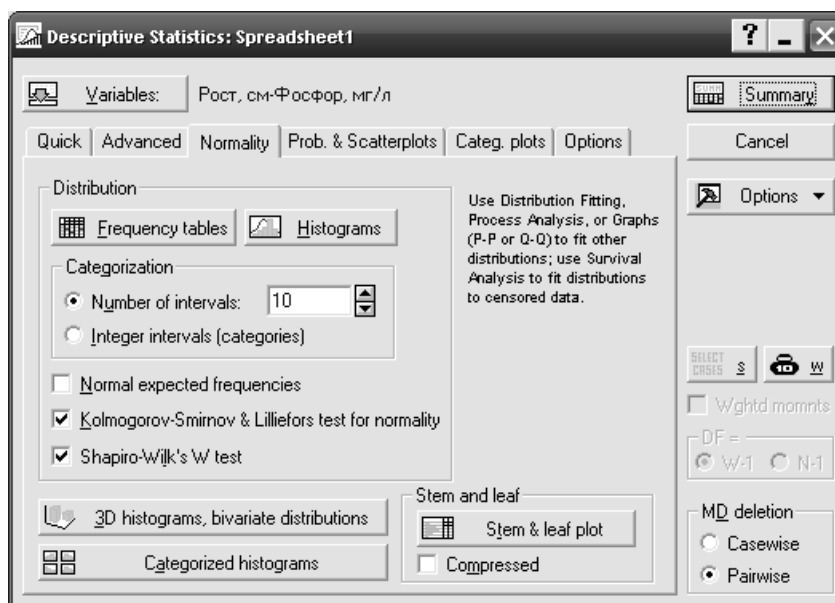


Рис. 5.3. Диалоговое окно модуля «Описательная статистика» (Descriptive statistics), вкладка «Нормальность» (Normality) пакета STATISTICA

Для запуска программы в верхнем меню **Statistics** надо выбрать команду **Basic Statistics / Tables**. В появившемся меню надо выбрать команду **Descriptive statistics** (Описательные статистики). Для выбора переменной, распределение которой необходимо проверить, надо нажать кнопку **Variables** и в открывшемся окне щелкнуть на имени переменной (переменных). Зайти во вкладку **Normality** (Нормальность) и поставить флажки напротив критериев Колмогорова – Смирнова и Шапиро – Уилка, как показано на рисунке (рис. 5.3). Остается нажать на кнопку **Histograms** (Гистограммы) и провести корректную интерпретацию результатов (рис. 5.3).

Рассмотрим результаты по росту детей дошкольного возраста (рис. 5.4).

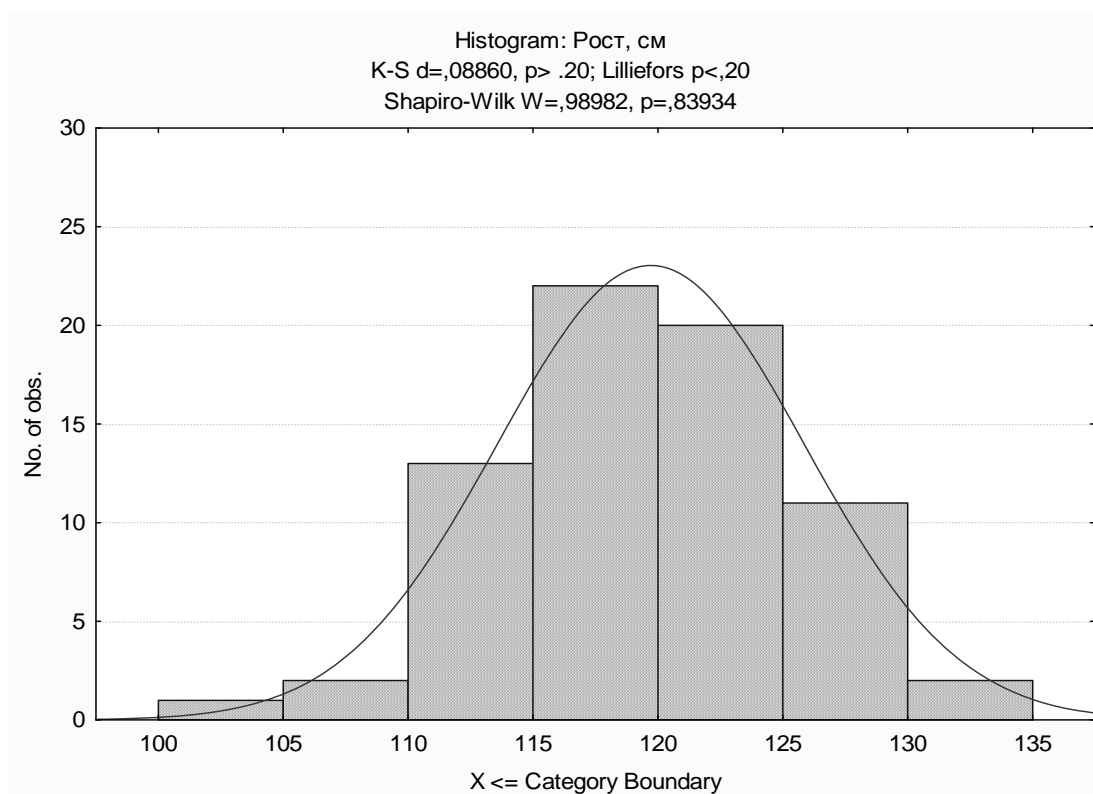


Рис. 5.4. Результаты проверки нормальности распределения в пакете STATISTICA: рост детей дошкольного возраста

На представленной гистограмме сравнения эмпирического распределения признака (столбчатая диаграмма) с теоретической нормальной кривой видно хорошее совпадение, что указывает на нормальность распределения роста. Более точный вывод можно сделать, обратившись к рассчитанному фактическому р-уровню значимости, представленному для обоих критериев вверху

гистограммы (рис. 5.4). По критерию Колмогорова – Смирнова (K-S) $p > 0.2$, а согласно критерию Шапиро – Уилка $p = 0.839$, таким образом, оба значения уровней значимости намного превышают критический уровень (0.05), значит нулевая гипотеза принимается, что доказывает недостоверное отличие распределения роста детей дошкольного возраста от нормального закона.

Обратимся теперь к результатам по концентрациям общего фосфора (рис. 5.5).

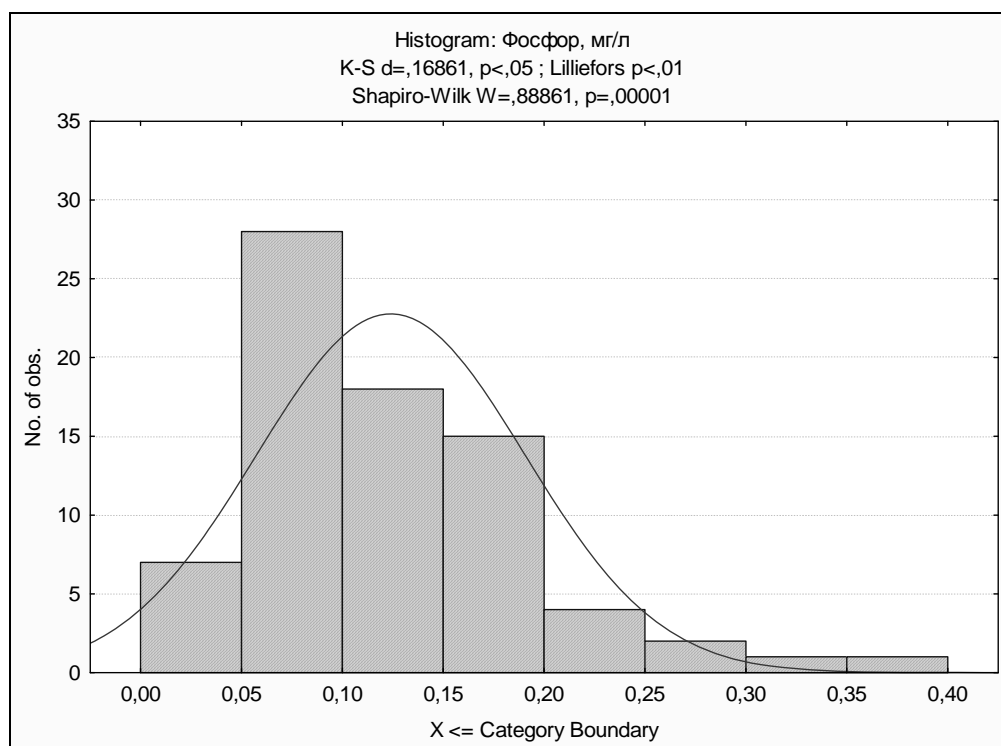


Рис. 5.5. Результаты проверки нормальности распределения в пакете STATISTICA: общий фосфор

График указывает на явную положительную асимметрию в распределении данного показателя (рис. 5.5). Фактический уровень значимости по критерию Колмогорова – Смирнова $p < 0.05$, по Шапиро – Уилка $p = 0.00001$ (рис. 5.5). Таким образом, с вероятностью большей 95% доказана достоверность отличия распределения концентраций общего фосфора от нормального закона.

5.4. Использование параметрических критериев в MS EXCEL

Рассмотрим алгоритм применения t-критерия Стьюдента и F-критерия Фишера в табличном процессоре MS EXCEL. Парамет-

рические критерии вычисляются в условиях строгих предпосылок (перечислены выше). Если указанные предпосылки, особенно нормальность распределения, не соблюдены при обработке данных, пользоваться параметрическими критериями неправомерно. Применение этих критериев в таком случае может искусственно завышать достоверность обсуждаемых различий – это фальсификация данных. Таким образом, первый шаг при использовании критериев различия – это проверка нормальности распределения по алгоритму, показанному выше. Если условия применимости t-критерия не выполнены, следует использовать непараметрические альтернативы t-критерия (например, U-тест Манна – Уитни).

В качестве примера рассмотрим применение t-критерия Стьюдента для выяснения достоверности различий средних величин двух независимых выборок в электронных таблицах. Вернемся к известному читателю примеру и сравним средний рост детей двух групп, предварительно сформировав 2 столбца с данными в таблице MS EXCEL:

Рост 6-летних девочек, см	Рост 6-летних мальчиков, см
128.5	111
135	112
126	125
124	116
128.5	120
124.6	127
124	119
124	127
130	135
125	116

Статистические гипотезы для решения данной задачи будут звучать следующим образом:

H_0 – средний рост 6-летних девочек недостоверно отличается от среднего роста 6-летних мальчиков, т. е. различия в выборках случайны.

H_a – средний рост 6-летних девочек достоверно отличается от среднего роста 6-летних мальчиков.

Первое условие применимости t-критерия Стьюдента – нормальность распределения – было проверено на данных о росте

детей дошкольного возраста. Равенство дисперсий проверим с помощью F-критерия Фишера. Предварительно следует сформулировать нулевую гипотезу: различия между дисперсиями роста двух групп детей недостоверны, а если и наблюдаются между выборочными дисперсиями, то это случайное явление. В меню Сервис надо выделить команду Анализ данных, в открывшемся диалоговом окне выбрать модуль Двухвыборочный F-тест для дисперсии и нажать кнопку Ок (рис. 5.6).

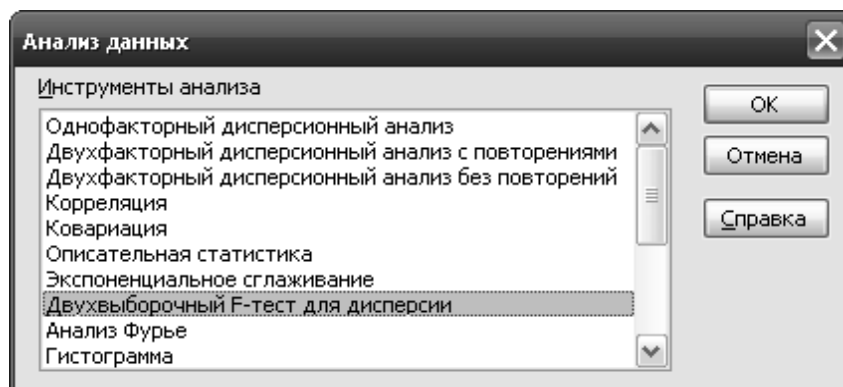


Рис. 5.6. Общий вид меню пакета «Анализ данных» с выбранным модулем «Двухвыборочный F-тест для дисперсии»

Внутри данного модуля указателем мыши необходимо автоматически ввести данные первой и второй выборки в поля Интервал переменной 1 и Интервал переменной 2 и нажать кнопку Ок (рис. 5.7).

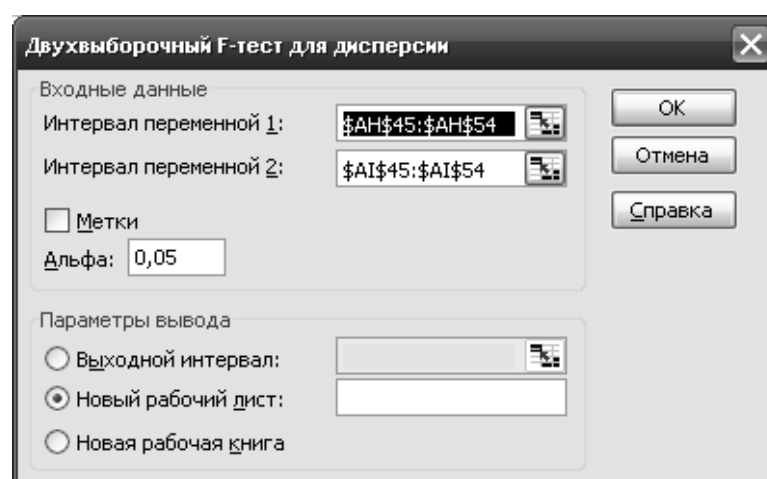


Рис. 5.7. Диалоговое окно процедуры «Двухвыборочный F-тест для дисперсии» табличного процессора MS EXCEL

В результате появится таблица с рассчитанными значениями выборочных дисперсий для 2-х выборок и фактическим p -уровнем значимости (табл. 5.2).

Таблица 5.2

Результаты проверки гипотезы о равенстве генеральных дисперсий с помощью F-критерия Фишера в табличном процессоре MS EXCEL

<i>Двухвыборочный F-тест для дисперсии</i>		
	<i>Переменная 1</i>	<i>Переменная 2</i>
Среднее	126.96	120.8
Дисперсия	12.8048	57.7333
Наблюдения	10	10
Df	9	9
F	0.2218	
P(F<=f) одностороннее	0.0175	
F критическое одностороннее	0.3146	

Фактический уровень значимости (0.0175) меньше критического (0.05) и нулевую гипотезу о равенстве дисперсий следует отклонить, о чем свидетельствуют и сильно различающиеся значения выборочных дисперсий для 2-х выборок (12.8 и 57.7). Таким образом, второе условие применимости t -критерия Стьюдента не выполнено, поэтому в MS EXCEL необходимо воспользоваться модулем **Двухвыборочный t -тест с различными дисперсиями**, в котором используется специальный метод расчета t -критерия. Сделаем это: в меню **Сервис** выделим команду **Анализ данных**, в открывшемся диалоговом окне выберем модуль **Двухвыборочный t -тест с различными дисперсиями** и нажмем кнопку **Ок** (рис. 5.8).

Примечание 4. В случае принятия нулевой гипотезы о равенстве дисперсий при использовании F-критерия Фишера следует применять модуль **Двухвыборочный t -тест с одинаковыми дисперсиями** (рис. 5.8).

Внутри модуля **Двухвыборочный t -тест с различными дисперсиями** проделываем аналогичные описанным выше процедуры (рис. 5.9).

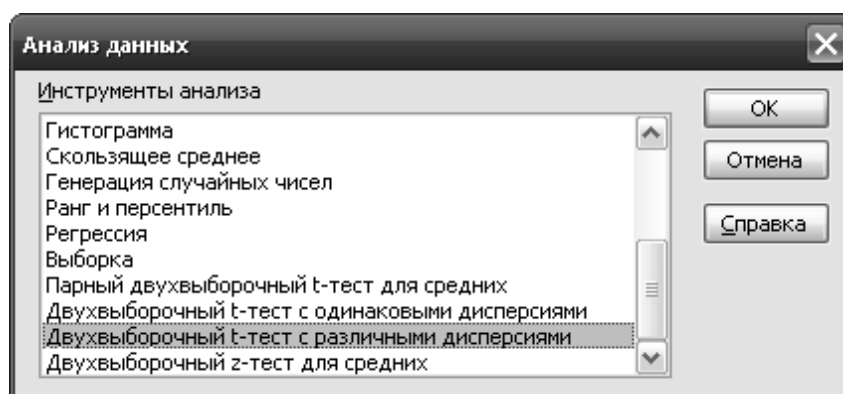


Рис. 5.8. Общий вид меню пакета «Анализ данных» с выбранным модулем «Двухвыборочный t-тест с различными дисперсиями»

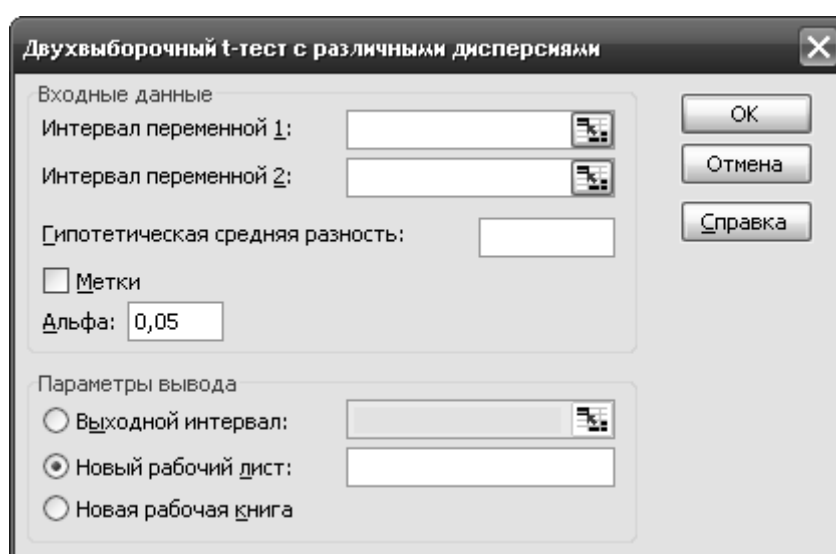


Рис. 5.9. Диалоговое окно процедуры «Двухвыборочный t-тест с различными дисперсиями» табличного процессора MS EXCEL

Окончательные результаты анализа представлены в таблице 5.3.

Таблица 5.3

***Результаты проверки гипотезы о равенстве
генеральных средних с помощью t-критерия Стьюдента
в табличном процессоре MS EXCEL***

<i>Двухвыборочный t-тест с различными дисперсиями</i>		
	<i>Переменная 1</i>	<i>Переменная 2</i>
Среднее	126.96	120.8
Дисперсия	12.80489	57.73333
Наблюдения	10	10

Гипотетическая разность средних	0	
Df	13	
t-статистика	2.319362	
P(T<=t) одностороннее	0.018647	
t критическое одностороннее	1.770933	
P(T<=t) двухстороннее	0.037293	
t критическое двухстороннее	2.160369	

Поскольку величина фактического уровня значимости (0.037) меньше критического уровня (0.05), нулевая гипотеза отклоняется и можно сделать заключение о достоверном различии роста девочек (127 см) и роста мальчиков (121 см).

5.5. Использование непараметрических критериев в пакете STATISTICA

Ранее было установлено, что распределение концентраций общего фосфора в воде озера Неро не подчиняется нормальному закону, поэтому применение параметрических методов неправомерно. Обратимся к следующему примеру: в результате ряда измерений были получены концентрации общего фосфора для озера Неро в 2010 и в 2011 годах. Необходимо установить достоверность снижения среднего содержания общего фосфора в воде озера за данный промежуток времени. Наиболее мощной непараметрической альтернативой t-критерия Стьюдента считается критерий Манна – Уитни, поэтому попытаемся решить предложенную задачу на его основе. Данные по концентрациям общего фосфора представлены ниже:

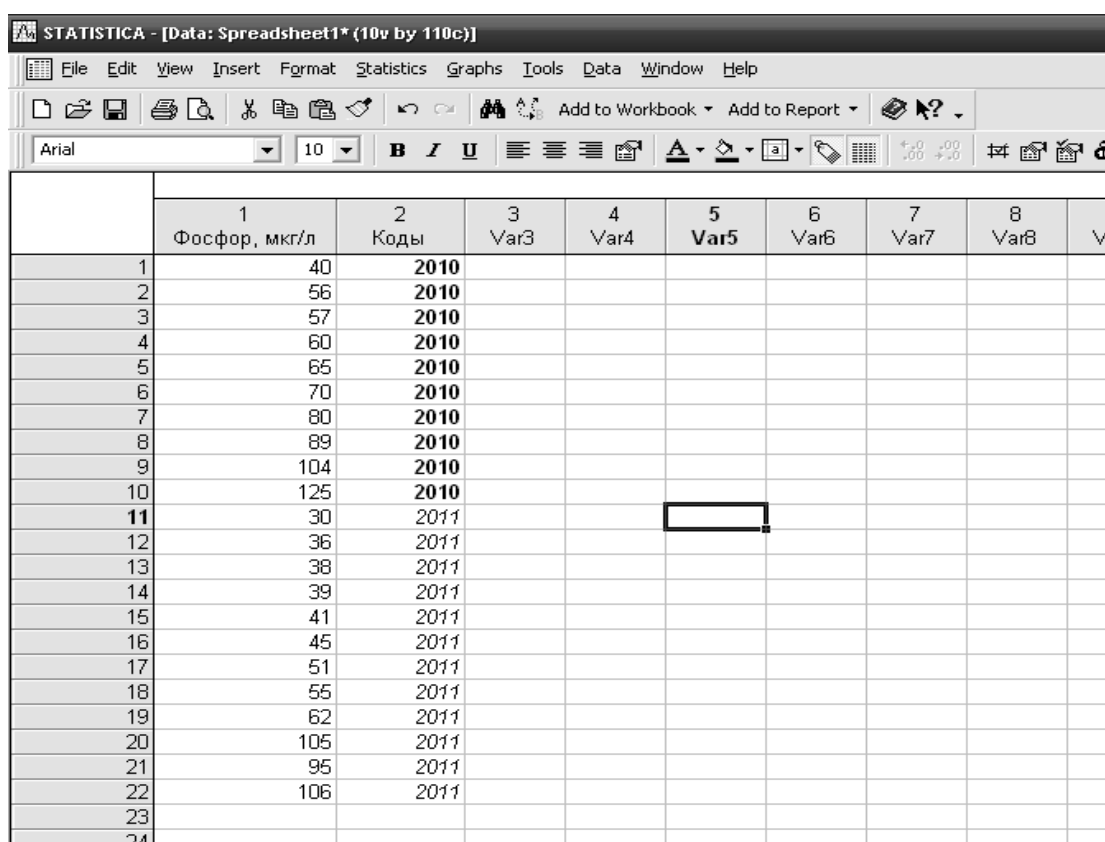
Общий фосфор, мкг/л, 2010 г.	40	56	57	60	65	70	80	89	104	125			Среднее 74.6
Общий фосфор, мкг/л, 2011 г.	30	36	38	39	41	45	51	55	62	105	95	106	Среднее 58.6

Выборочные средние значения действительно различаются, сформулируем гипотезы:

H_0 – концентрации общего фосфора в озере недостоверно отличаются в разные годы.

H_a – концентрации общего фосфора в озере достоверно отличаются в разные годы.

Для правильной организации исходных данных в электронной таблице STATISTICA необходимо обе выборки поместить в один столбец (Dependent variable), а в соседнем столбце разбить исходные данные на 2 группы, введя соответствующие коды, например 2010 и 2011. Этот столбец обозначается как группирующая переменная (Grouping variable) (рис. 5.10).



	1 Фосфор, мкг/л	2 Коды	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8	9 Var9
1	40	2010							
2	56	2010							
3	57	2010							
4	60	2010							
5	65	2010							
6	70	2010							
7	80	2010							
8	89	2010							
9	104	2010							
10	125	2010							
11	30	2011							
12	36	2011							
13	38	2011							
14	39	2011							
15	41	2011							
16	45	2011							
17	51	2011							
18	55	2011							
19	62	2011							
20	105	2011							
21	95	2011							
22	106	2011							
23									
24									

Рис. 5.10. Организация исходных данных в электронной таблице программы STATISTICA при использовании критериев достоверности для независимых выборок

Для проведения анализа в верхнем меню Statistics надо выбрать команду Nonparametrics (Непараметрическая статистика). В появившемся меню выбираем модуль Comparing two independent samples (groups) (Сравнение двух независимых выборок). Для выбора переменных надо нажать кнопку Variables и в открывшемся окне щелкнуть на имени переменных, как показано на рисунке (рис. 5.11).

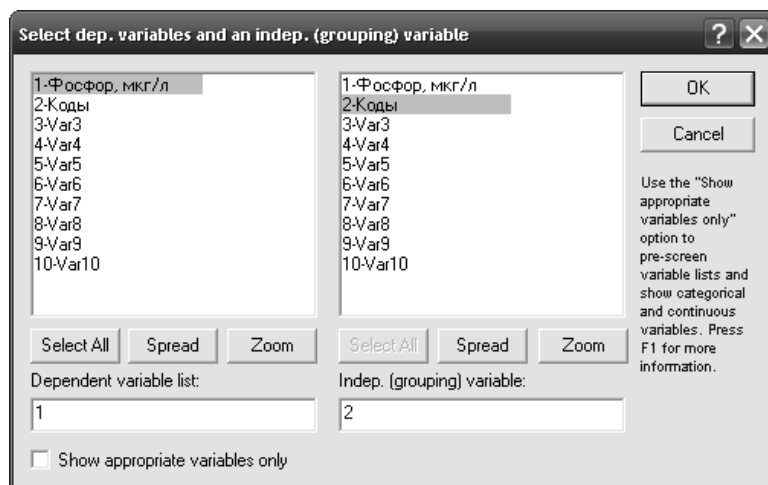


Рис. 5.11. Диалоговое окно выбора переменных в пакете STATISTICA

После этого в модуле **Comparing two independent samples (groups)** остается нажать на кнопку **Mann-Whitney U-test** (рис. 5.12), результаты будут выведены в табличной форме (табл. 5.4).



Рис. 5.12. Диалоговое окно модуля
«Сравнение двух независимых выборок»
(Comparing two independent samples (groups)) пакета STATISTICA

В таблице результатов достаточно найти столбец с фактическим p -уровнем значимости (p -level). В нашем примере он больше критического уровня ($0.086 > 0.05$) (табл. 5.4), что означает недостоверность отличий концентраций общего фосфора в озере Неро в разные годы измерений.

Результаты использования критерия Манна – Уитни

Variable	Rank Sum	Rank Sum	U	Z	p-level	Z	p-level	Valid N	Valid N	Rank Sum
Фосфор	141	112	34	1.71	0.086	1.71	0,086	10	12	0.093

5.6. Браковка выбросов и критерии исключения

В биологии и экологии часто встречаются ситуации, когда одно из полученных значений выборки сильно отличается от остальных. Эти отклонения могут возникнуть в результате неточности измерений, ошибок внимания, методических погрешностей и т. д. С другой стороны, отклоняющееся значение может отражать естественную вариабельность признака. Можно ли такие резко выделяющиеся значения использовать при дальнейших расчетах? Ответить на этот вопрос бывает затруднительно, одним из возможных методов принятия решения является применение критериев исключения.

Бражкой выбросов принято называть статистическую процедуру проверки сомнительных (выскакивающих) значений признака в выборке. Во многих случаях данный вид анализа необходим, поскольку отдельные нетипичные значения могут значительно смещать выборочные характеристики, например средние арифметические. Данный анализ наиболее актуален при малом объеме выборки, поскольку даже один выброс может значительно сместить значения выборочных характеристик. Статистические гипотезы при решении данной задачи формулируются следующим образом:

H_0 – сомнительные значения признака принадлежат к данной выборке, отличие выброса от других значений признака недостоверно.

H_a – сомнительные значения признака не принадлежат к данной выборке, отличие выброса от других значений признака достоверно.

Одним из программных средств решения данной задачи является модуль **Обработка выбросов** статистического пакета ATTESTAT (рис. 5.13).

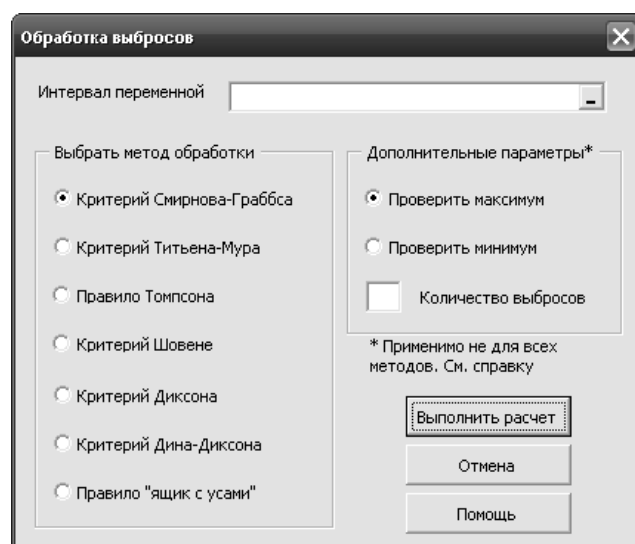


Рис. 5.13. Диалоговое окно модуля «Обработка выбросов» программы ATTESTAT

Достаточно ввести интервал выборки и выбрать из предложенного многообразия соответствующий критерий. При отклонении нулевой гипотезы выброс будет помечаться красным цветом в электронной таблице MS EXCEL.

Глава 6. Количественная оценка влияния фактора

6.1. Сущность метода

Наиболее распространенной задачей в биологии и экологии в общем виде является установление эффекта действия фактора, доказательство причинно-следственных механизмов изучаемого явления или процесса, отделение систематического влияния экологического фактора от случайных причин, оценка роли каждого из факторов в отдельности и их совместного воздействия на живые системы.

В предыдущей главе кратко были разобраны методы определения различий между двумя выборками по тем или иным показателям или признакам. По сути, в этом уже просматривается задача выявления причин, лежащих в основе наблюдаемых различий двух выборочных совокупностей. Важно не доказательство реальности существования различий между двумя выборками, а анализ тех причин, которые приводят к этим различиям.

Другими словами, если различия математически подтверждены, значит должен существовать фактор, определяющий их. Таким образом, проверка статистических гипотез на основе двух выборок, двух групп значений признака при грамотном планировании сравнительного анализа – это простейший инструмент количественной оценки влияния фактора на изучаемый показатель. К примеру, простейшие биологические и экологические эксперименты основаны на распределении изучаемых объектов на две группы – *контрольную* (без влияния фактора) и *опытную* (где вводится влияние фактора). В итоге решается задача обычного сравнения двух выборок с использованием критериев достоверности и, при корректном планировании эксперимента, исследователь способен с определенной вероятностью установить либо отсутствие, либо наличие эффекта действия фактора на изучаемый объект.

Существуют способы усложнения схемы исследований по оценке действия на признак интересующих исследователя факторов с целью извлечения дополнительной информации о механизмах и характере влияния фактора. В основе этого усложнения лежит методология, давно применяемая в экологических науках, – создание *градиента фактора*, т. е., образно говоря, ранжированное распределение во времени или в пространстве значений фактора, от минимального уровня действия до максимального воздействия. Если фактор влияет на изучаемый признак, то и значения признака обязательно должны каким-то образом изменяться в этом градиенте. Если фактор не оказывает воздействия, то какой бы широкий градиент фактора исследователь ни создавал, значения признака будут относительно постоянны при разных уровнях действия фактора.

Создание градиента фактора – это переход от задачи сравнения лишь двух выборок к задаче сравнения одновременно нескольких выборок, число которых будет определяться количеством уровней воздействия фактора, выбранных исследователем. Математически решить поставленную задачу позволяет так называемый *дисперсионный анализ*. Рассмотрим сущность данного метода на классическом примере: получены данные об урожайности озимой ржи при разных дозах внесения в почву минеральных удобрений. Необходимо установить достоверность эффекта влияния минеральных веществ на изучаемый показатель (урожайность).

Дозы минеральных веществ, кг/га	Урожайность по повторностям, ц/га			Средняя урожайность, ц/га
	№ 1	№ 2	№ 3	
15	8	8.4	8.6	8.3
20	8.2	9	10	9.0
25	11	12	13	12
30	7.5	8.5	7.4	7.8

В примере создан градиент фактора «минеральные удобрения» от 15 кг/га до 30 кг/га, выделены 4 уровня предполагаемого воздействия фактора, соответственно этому получены 4 выборки по одному и тому же признаку – «урожайность культуры», в каждой из которых по 3 повторности эксперимента. Рассчитаны средние значения в каждой выборке на основе 3-х повторностей, обозначим их как *групповые средние* значения признака. Сложив групповые средние значения и поделив их на число слагаемых (4), получим *общую групповую среднюю*, учитывающую все 4 выборки. Она будет равна:

$$\frac{8.3+9+12+7.8}{4} = 9.3 \text{ ц / га}$$

Анализ таблицы показывает, что с увеличением дозы минеральных веществ групповые средние значения урожайности не остаются постоянными, а неким образом варьируют или изменяются. При этом изменения средней урожайности не кажутся значительными, чтобы уверенно говорить о факте влияния минеральных веществ на этот показатель. С другой стороны, различаются между собой не только групповые средние значения урожайности, но и урожайность по повторностям внутри каждой из групп при постоянном уровне изучаемого фактора. Подобное изменение урожайности при неизменных значениях выделенного фактора свидетельствует о влиянии на изучаемый признак, помимо минеральных удобрений, других неизвестных или неучтенных в данном исследовании факторов. Таким образом, *сущность дисперсионного анализа* заключается в сравнении средних значений признака, связанных с соответствующими уровнями влияющего фактора: если средние значения признака сильно изменяются, значит фактор оказывает влияние на признак.

Почему же тогда анализ, основанный на сравнении нескольких средних значений между собой, был обозначен как *дисперсионный*? Здесь есть небольшая хитрость. Дело в том, что средние значения признака сравниваются между собой на основе разложения общей *дисперсии* признака на компоненты.

Примечание 1. При прочтении следующих абзацев автор советует читателю заглянуть в главу 2 (раздел 2.4) и вспомнить формулу расчета дисперсии.

Объективным статистическим показателем варьирования групповых средних значений и отдельных значений признака в рассматриваемой таблице является дисперсия, другими словами, именно расчет дисперсии способен показать степень изменения групповых средних значений урожайности в градиенте фактора. Общая дисперсия варьирования признака при оценке влияния фактора разлагается на *факториальную* (межгрупповую) дисперсию и *остаточную* (внутригрупповую) дисперсию. Разберем эти понятия подробнее.

Факториальная (межгрупповая) дисперсия связана с влиянием изучаемого фактора и рассчитывается как отношение суммы квадратов отклонений групповых средних значений признака от общей средней к числу степеней свободы:

$$S^2_{\text{факт.}} = \frac{\sum_{i=1}^n (\bar{X}_i - M)^2}{n-1},$$

где \bar{X}_i – групповое среднее значение признака,

M – общая групповая средняя,

n – объем выборки.

Чем существеннее фактор влияет на признак, тем сильнее групповые средние значения должны отклоняться от общей групповой средней, тем больше должна быть факториальная дисперсия, и наоборот. Таким образом, именно от величины факториальной дисперсии зависит вероятность достоверного влияния фактора, при незначительной факториальной дисперсии вероятность достоверного воздействия фактора будет небольшой.

Остаточная (внутригрупповая) дисперсия определяется влиянием случайных (т. е. неучтенных, неконтролируемых в исследовании) факторов и рассчитывается как отношение сум-

мы квадратов отклонений отдельных значений признака внутри группы от соответствующей групповой средней к числу степеней свободы:

$$S^2_{ост} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1},$$

где X_i – отдельное значение признака внутри группы,

\bar{X} – групповое среднее значение признака,

n – объем выборки.

Чем значительнее величины признака отклоняются друг от друга внутри каждой группы (т. е. при одинаковых уровнях изучаемого фактора), тем больше будет остаточная дисперсия и тем существеннее роль случайных факторов по сравнению с градиентом изучаемого фактора.

Наконец, *общую дисперсию* значений изучаемого признака можно рассчитать как отношение суммы квадратов отклонений отдельных значений признака внутри группы от общей средней к числу степеней свободы:

$$S^2_{общая} = \frac{\sum_{i=1}^n (X_i - M)^2}{n-1} \quad \text{или} \quad S^2_{общая} = S^2_{факт} + S^2_{ост}$$

О достоверности влияния фактора на признак судят по F-критерию Фишера, с помощью которого можно рассчитать вероятность эффекта действия фактора, при достоверном влиянии вероятность должна быть $P \geq 0.95$ (или $p \leq 0.05$). Формула расчета фактического значения критерия Фишера при дисперсионном анализе довольно проста – это отношение факториальной дисперсии к остаточной:

$$F = \frac{S^2_{факт}}{S^2_{ост}}.$$

Чем больше факториальная дисперсия, определяемая влиянием изучаемого фактора, по сравнению с остаточной диспер-

сией, не зависящей от изучаемого фактора, тем выше фактическое значение F-критерия Фишера и вероятность достоверного влияния изучаемого фактора на признак.

Перед расчетом фактического значения критерия Фишера (если используются статистические программы, то перед расчетом фактического p -уровня значимости) необходимо сформулировать статистические гипотезы:

H_0 – фактор недостоверно (случайно) влияет на признак;

H_a – фактор достоверно влияет на признак, в генеральной совокупности групповые средние значения признака не равны между собой.

Теперь обозначим круг задач, которые можно решать методами дисперсионного анализа:

1. Определение влияния одного или нескольких факторов на изучаемый признак или показатель.

2. Оценка влияния на признак не только каждого из факторов в отдельности, но и их совместного эффекта (взаимодействие факторов).

3. Установление силы влияния отдельных факторов и их совместного действия на изменчивость (вариабельность) изучаемого признака.

4. В рамках метода возможно производить множественные сравнения средних значений признака.

6.2. Базовая терминология дисперсионного анализа

1. Признаки, изменяющиеся под воздействием тех или иных причин, называют *результативными*.

Пример: урожайность озимой ржи – результативный признак.

2. Причины, вызывающие изменение результативного признака, называют *факторами*.

Пример: минеральное удобрение – фактор.

Факторы в дисперсионном анализе принято делить на 2 группы:

– *Регулируемые (контролируемые, организованные)* – факторы, влияние которых необходимо определить в исследовании.

– *Случайные (неучтенные, нерегулируемые)* – факторы, которые не контролируются в исследовании, но оказывают воздействие на величину результативного признака.

3. Значения (степень интенсивности) фактора принято называть *градациями или уровнями* данного фактора. Градации могут быть количественными и качественными.

Пример: градации фактора «минеральное удобрение» – дозы 15, 20, 25, 30 кг/га; градации фактора «природная зона» – тундра, северная тайга, средняя тайга, смешанный лес.

Примечание 2. Часто при планировании исследования от правильного выбора числа уровней фактора зависят результаты дисперсионного анализа. Чем шире размах уровней фактора, тем больше вероятность выявить эффект влияния фактора на результативный признак.

4. *Дисперсионный комплекс* – статистическая таблица, отражающая зависимость значений результативного признака от уровней фактора(ов). **Столбцы** – уровни фактора, **строки** – значения (повторности) результативного признака для каждого уровня.

В зависимости от числа факторов, влияние которых одновременно изучается, дисперсионные комплексы делятся на *однофакторные, двухфакторные и многофакторные*. Соответственно и дисперсионный анализ может быть *однофакторным, двухфакторным и многофакторным*.

6.3. Условия применимости и основные этапы дисперсионного анализа

Как и любой другой метод математической обработки данных, дисперсионный анализ с применением F-критерия Фишера имеет свои строгие границы применимости:

1. Корректное применение дисперсионного анализа предполагает нормальное или близкое к нормальному распределение результативного признака.

2. Желательно, чтобы группы в дисперсионном комплексе были одинакового или примерно одинакового объема. Несоблюдение этого условия влияет на мощность метода, особенно при двух- и многофакторном анализе.

3. Группы должны иметь примерно одинаковую остаточную дисперсию.

Если условия применимости метода не выполнены, необходимо использовать непараметрические приемы оценки влияния фактора.

Выделим условно **основные этапы** проведения дисперсионного анализа:

1. Создание структуры дисперсионного комплекса.
2. Проверка условий применимости дисперсионного анализа.
3. Определение того, влияет или нет изучаемый(ые) фактор(факторы) на результативный признак.
4. Проведение апостериорных (множественных) попарных сравнений групповых средних значений признака.
5. Заключительный этап – оценка силы влияния фактора на признак.

Разберем основные этапы дисперсионного анализа на основе использования прикладных статистических программ.

6.4. Однофакторный дисперсионный анализ в среде MS EXCEL и в пакете STATISTICA

Возможности пакета STATISTICA и ограничения табличного процессора MS EXCEL при проведении дисперсионного анализа продемонстрируем на классическом примере исследования урожайности озимой ржи в зависимости от разных доз минеральных удобрений.

Первый шаг – Создание структуры дисперсионного комплекса.

В MS EXCEL дисперсионный комплекс организуется по столбцам – это уровни фактора, строки – отдельные повторности урожайности (рис. 6.1).

В MS EXCEL однофакторный дисперсионный анализ выполняется с помощью одноименной статистической процедуры, входящей в Пакет анализа (рис. 6.2).

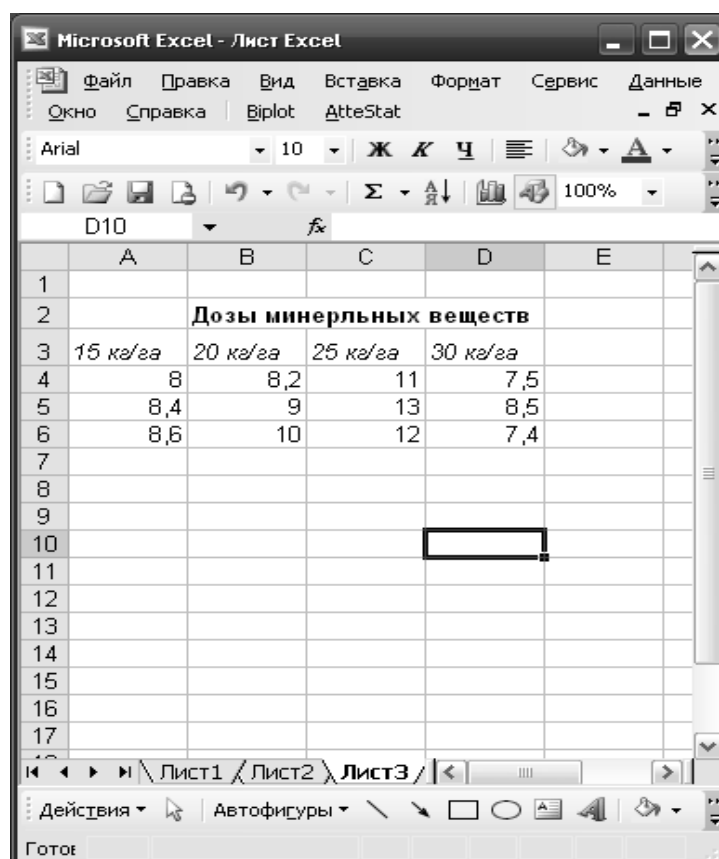


Рис. 6.1. Структура дисперсионного комплекса в среде MS EXCEL

В программе STATISTICA в один из столбцов заносятся все значения урожайности, а в соседнем столбце организуется группирующая переменная, разбивающая значения признака на 4 группы с помощью соответствующих кодов, например 15 кг/га, 20 кг/га, 25 кг/га, 30 кг/га (рис. 6.3).

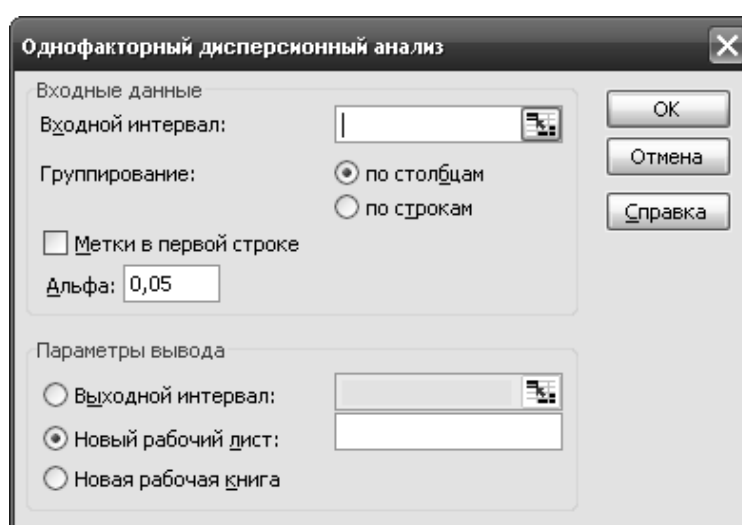


Рис. 6.2. Диалоговое окно процедуры «Однофакторный дисперсионный анализ» табличного процессора MS EXCEL

	1 Урожайность, ц/га	2 Доза удобрений, кг/га	3 Var3	4 Var4
1	8	15		
2	8,4	15		
3	8,6	15		
4	8,2	20		
5	9	20		
6	10	20		
7	11	25		
8	13	25		
9	12	25		
10	7,5	30		
11	8,5	30		
12	7,4	30		
13				
14				
15				

Рис. 6.3. Структура дисперсионного комплекса в электронной таблице пакета STATISTICA

Однофакторный дисперсионный анализ (Analysis of variance, сокращенно ANOVA) можно запустить либо через верхнее меню **Statistics** и выпадающую команду **ANOVA**, либо через команду **Basic Statistics / Tables**. В последнем случае нужно зайти в модуль **Breakdown & one-way ANOVA** (Группировка и однофакторный дисперсионный анализ) (рис. 6.4).

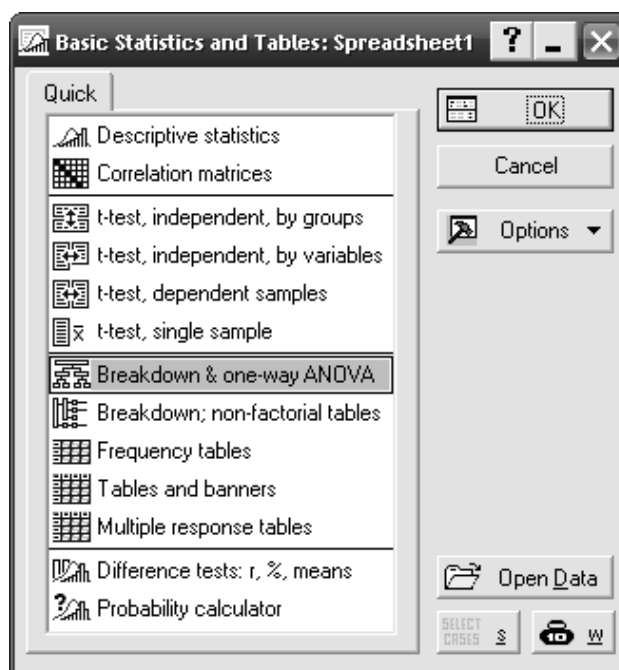


Рис. 6.4. Меню команды «Basic Statistics / Tables» с выбранным модулем «Breakdown & one-way ANOVA»

Второй шаг – Проверка условий применимости дисперсионного анализа.

В нашем примере количество повторностей в каждой группе одинаково, значит первое условие о равенстве выборок по каждому уровню фактора выполнено.

В MS EXCEL отсутствуют средства проверки применимости дисперсионного анализа, в частности однородности дисперсий и нормальности распределения результативного признака.

В пакете STATISTICA проверить нормальность распределения результативного признака можно с использованием критериев согласия. Поскольку число повторностей невелико и общий объем выборки небольшой, то наиболее подходящим является критерий Шапиро – Уилка. Алгоритм проверки нормальности распределения признака был разобран в главе 5, поэтому здесь приводим фактический p -уровень значимости критерия Шапиро – Уилка. Он оказался равен $p = 0.067$, что больше критического (0.05), следовательно нулевая гипотеза, гласящая о случайности отличия эмпирического распределения от нормального закона, принимается. Второе условие выполнено.

Третье условие об однородности (примерном равенстве) дисперсий проверяется с помощью специально разработанных критериев, к примеру критерия Левена (*Levene test*) и критерия Брауна – Форсайта (*Brown-Forsythe test*). Нулевая гипотеза формулируется известным способом: дисперсии групп различаются недостоверно (дисперсии однородны). Для проверки нулевой гипотезы необходимо зайти в модуль **Breakdown & one-way ANOVA**, указать программе зависимую и группирующую переменные и после нажатия кнопки **Ok** в открывшемся диалоговом окне выбрать третью закладку **ANOVA & tests**, как показано на рисунке (рис. 6.5).

Остается нажать на кнопку либо **Levene tests**, либо **Brown-Forsythe tests** (рис. 6.5), пример таблицы результатов по критерию Левена показан ниже (табл. 6.1).

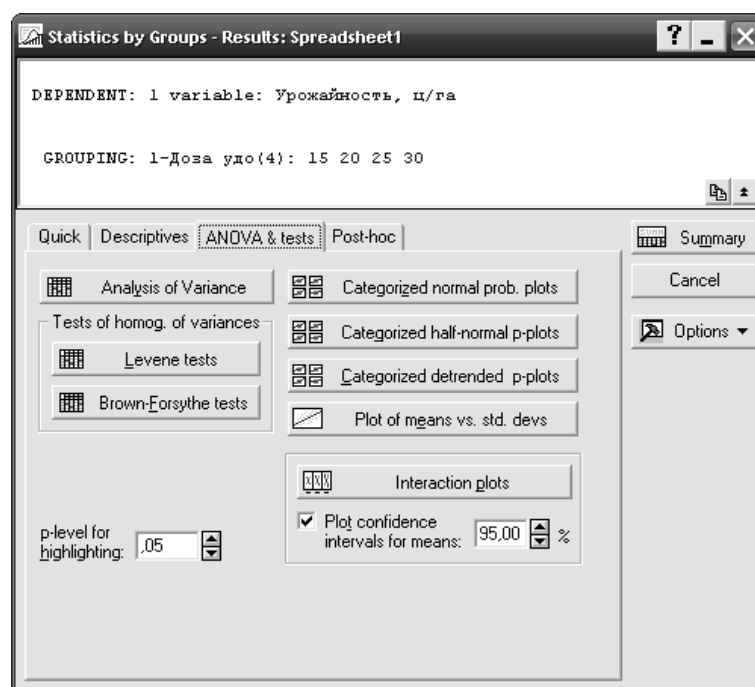


Рис. 6.5. Диалоговое окно модуля «Breakdown & one-way ANOVA», закладка «ANOVA & tests»

Таблица 6.1

Результаты использования критерия Левена

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Урожайность ц/га	0.36	3	0.12	1.25	8	0.15	0.76	0.54

Фактический р-уровень значимости (0.54) значительно превышает критический (0.05), следовательно нулевая гипотеза принимается, дисперсии групп различаются недостоверно. Третье условие выполнено.

Третий шаг – Установление достоверности влияния фактора на результативный признак.

Вначале обозначим статистические гипотезы:

H_0 – дозы удобрений недостоверно влияют на урожайность озимой ржи, различия групповых выборочных средних случайны;

H_a – дозы удобрений достоверно влияют на урожайность озимой ржи, различия групповых выборочных средних неслучайны.

В MS EXCEL в диалоговом окне процедуры Однофакторный дисперсионный анализ (рис. 6.2) указателем мыши необходимо выделить диапазон ячеек с дисперсионным

комплексом, поставить флажок напротив **Метки** в первой строке и нажать кнопку **Ок** (рис. 6.2). Результаты появятся в виде таблицы (табл. 6.2).

Таблица 6.2

Результаты дисперсионного анализа в среде MS EXCEL

<i>Источник вариации</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-значение</i>	<i>F критическое</i>
Между группами	31.58667	3	10.52889	18.49878	0.000587987	4.066180557
Внутри групп	4.553333	8	0.569167			
Итого	36.14	11				

В таблице результатов содержатся все необходимые статистические выкладки, в частности рассчитана факториальная (между группами) дисперсия ($MS = 10.5$) и остаточная (внутри групп) дисперсия ($MS = 0.57$), приводится фактическое значение F-критерия Фишера (18.49) и соответствующий этому значению р-уровень значимости (0.000587). Поскольку фактический р-уровень значительно меньше критического уровня (0.05), нулевая гипотеза отвергается и статистически доказывается влияние минерального удобрения на урожайность озимой ржи с вероятностью $P = 99.99994$.

Аналогичные результаты получаем в пакете STATISTICA: для этого в модуле **Breakdown & one-way ANOVA**, в закладке **ANOVA & tests**, необходимо нажать на кнопку **Analysis of variance** (рис. 6.5). Результат будет представлен в табличной форме (табл. 6.3).

Таблица 6.3

Результаты дисперсионного анализа в системе STATISTICA

<i>Variable</i>	<i>SS Effect</i>	<i>df Effect</i>	<i>MS Effect</i>	<i>SS Error</i>	<i>df Error</i>	<i>MS Error</i>	<i>F</i>	<i>p</i>
Урожайность, ц/га	31.58	3	10.52	4.55	8	0.569	18.498	0.000588

Четвертый шаг – Апостериорные (множественные) сравнения групповых средних значений признака.

После того как достоверно установлено влияние регулируемого фактора на результативный признак, при необходимости прибегают к множественному сравнению групповых средних друг с другом или с какой-либо другой величиной, например контрольным вариантом эксперимента.

В табличном процессоре MS EXCEL проведение данного этапа дисперсионного анализа невозможно. В пакете прикладных программ STATISTICA для этих целей имеются специально разработанные критерии достоверности.

В чем суть множественных попарных сравнений групповых средних? Установление достоверности влияния фактора на результативный признак означает, что групповые средние не равны между собой. Однако принятие данного факта в результате отклонения нулевой гипотезы не говорит о том, какие именно групповые средние значения признака достоверно отличаются между собой, а какие недостоверно. Нулевая гипотеза о случайном влиянии фактора на признак может отклоняться и в случае, когда все групповые средние достоверно отличаются друг от друга, и в случае, когда лишь одно среднее значение достоверно отличается от всех остальных, различающихся между собой лишь случайно. Результат в обоих случаях один и тот же – отклонение нулевой гипотезы, но характер действия фактора на признак разный. Поэтому во многих случаях подобный анализ даёт исследователю важную дополнительную информацию о самом характере воздействия фактора на результативный признак, позволяет установить уровень фактора, оказывающий наибольшее воздействие.

Следующий вопрос: можно ли применять при множественном сравнении средних значений наиболее популярный и известный читателю t-критерий Стьюдента? Не вдаваясь в подробности, следует предупредить, что применение t-критерия Стьюдента для решения поставленной задачи неправомерно, поскольку данный метод разработан для сравнения средних значений 2-х выборок, при дисперсионном анализе исследователь имеет дело с несколькими выборками. При многократном использовании t-критерия Стьюдента для попарного сравнения многих средних значений, связанных между собой, увеличивается вероятность ошибочного обнаружения достоверных отличий между средними значениями,

когда их на самом деле нет! В статистике данная проблема получила обозначение «*эффект множественных сравнений*». Что делать в этом случае? Либо вводить поправки для t-критерия Стьюдента, снижая критический уровень значимости с учетом числа попарных сравнений, либо использовать критерии множественных сравнений:

1. *Критерий наименьшей значимой разности* (LSD test) – аналог t-критерия Стьюдента, разработанный для попарного сравнения большого числа выборок.

2. *Критерий Тьюки* (Tukey test) – применяется при сравнении групповых средних дисперсионного комплекса при равном числе наблюдений в каждой группе.

3. *Критерий Шеффе* (Scheffe test) – используется при наличии как неравночисленных, так и равных по объему групп в дисперсионном комплексе.

Перед проведением множественных попарных сравнений групповых средних значений признака рекомендуется провести графический анализ, позволяющий предварительно оценить характер различий между средними значениями. Для этого в модуле **Breakdown & one-way ANOVA** надо зайти в первую закладку **Quick** (Быстрая) и нажать на кнопку **Interaction plots** (Графики взаимодействия). В результате появится знакомая читателю диаграмма размаха, на которой точками обозначены средние значения урожайности для каждой из 4-х доз минеральных удобрений, а отрезками – 95% доверительный интервал для генеральной средней (рис. 6.6).

Из графика видно, что в градиенте фактора от наименьших до наибольших доз урожайность озимой ржи изменяется нелинейно, вначале возрастает, а потом снижается (рис. 6.6). При этом средняя урожайность при дозах фактора 15 кг/га, 20 кг/га и 30 кг/га различается незначительно, а перекрывание доверительными интервалами самих выборочных средних значений свидетельствует в первом приближении о недостоверном отличии этих 3-х средних значений друг от друга. Напротив, средняя урожайность озимой ржи при дозе минеральных удобрений 25 кг/га значительно отличается от остальных средних, при этом доверительный интервал не накрывает ни одну из выборочных средних (рис. 6.6).

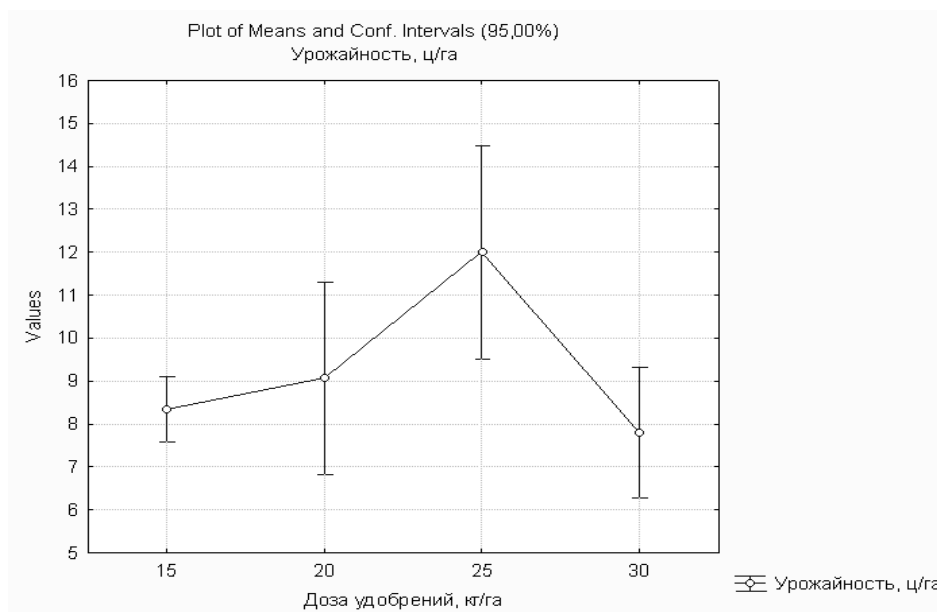


Рис. 6.6. Диаграмма размаха

Проверим наши рассуждения с помощью критерия наименьшей значимой разности: нулевая гипотеза состоит в том, что попарно средние значения урожайности отличаются друг от друга недостоверно. Необходимо в том же самом модуле **Breakdown & one-way ANOVA** зайти в четвертую закладку **Post-hoc** (Апостериорные критерии) и нажать на кнопку **LSD test** (рис. 6.7).

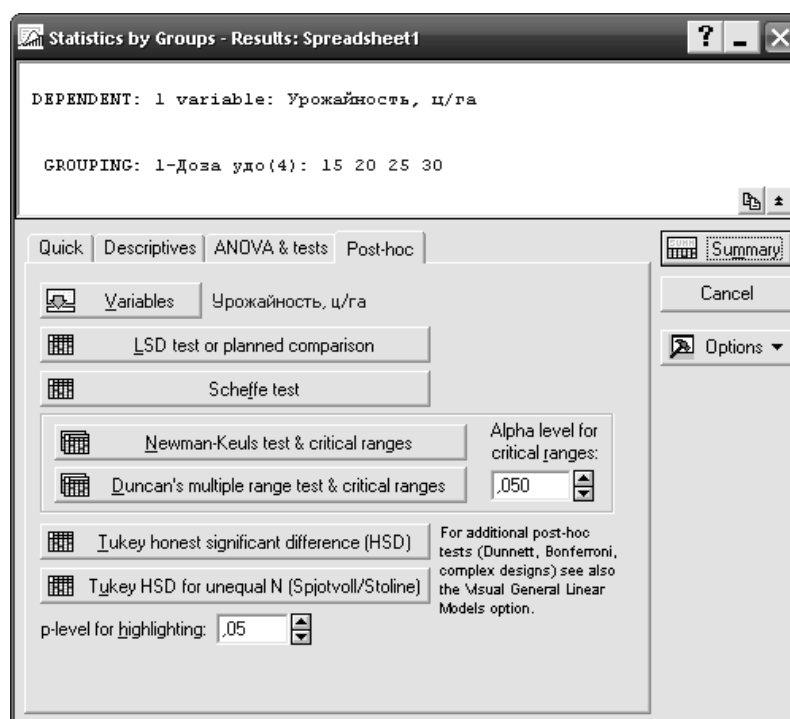


Рис. 6.7. Диалоговое окно модуля «Breakdown & one-way ANOVA», закладка «Post-hoc»

В таблице результатов рассчитаны фактические р-уровни значимости для каждого попарного сравнения всех 4-х средних значений урожайности (табл. 6.4). Цветом обозначены р-уровни, меньшие 0.05, следовательно, именно эти пары средних и отличаются друг от друга достоверно (табл. 6.4). Анализ таблицы показывает, что достоверные отличия отмечаются только между 3-й средней (урожайность озимой ржи при дозе 25 кг/га) и всеми остальными средними (табл. 6.4). Первая, вторая и четвертая средние (урожайность озимой ржи при дозах 15 кг/га, 20 кг/га и 30 кг/га) отличаются недостоверно.

Таблица 6.4

Результаты использования критерия наименьшей значимой разности в системе STATISTICA

<i>Дозы удобрений</i>	<i>{1}</i>	<i>{2}</i>	<i>{3}</i>	<i>{4}</i>
15 кг/га {1}		0.267977	0.000341	0.411804
20 кг/га {2}	0.267977		0.001423	0.073782
25 кг/га {3}	0.000341	0.001423		0.000135
30 кг/га {4}	0.411804	0.073782	0.000135	

Таким образом, применение критерия наименьшей значимой разности подтвердило предположения, сделанные при графическом анализе. Это означает, что оптимальной или наиболее эффективной дозой минеральных удобрений в нашем примере является 25 кг/га, дальнейшее увеличение количества минеральных удобрений нецелесообразно, поскольку не вызывает соответствующей прибавки в урожайности озимой ржи. Кроме того, множественные апостериорные сравнения доказали нелинейный эффект влияния минеральных удобрений на урожайность озимой ржи в изученном градиенте фактора.

Примечание 3. Методы множественного сравнения следует применять только после того, как с помощью дисперсионного анализа отвергнута нулевая гипотеза о равенстве всех средних. В качестве ещё одного параметрического критерия для множественных сравнений можно использовать критерий Ньюмена – Кейлса (Гланц, 1999). Иногда задача заключается в том, чтобы сравнить несколько групп с единственной – контрольной. В этом случае можно воспользоваться критерием Даннета, предназна-

ченным для сравнения нескольких выборок с одной контрольной группой (Гланц, 1999). Технически данная процедура может быть реализована в пакете прикладных программ STATISTICA.

Пятый шаг – Оценка силы влияния фактора на признак.

Последнее, что может установить исследователь относительно эффекта действия фактора на признак, – это доля вариации значений признака, которая зависит от регулируемого фактора, может быть объяснена влиянием этого фактора. Также можно рассчитать, какая часть разброса значений признака определяется воздействием случайных неизвестных факторов.

Показатель силы влияния фактора представляет собой отношение факториальной дисперсии к общей дисперсии и выражается формулой:

$$h^2 = \frac{S^2_{\text{факт}}}{S^2_{\text{общая}}} \cdot 100\%$$

Поскольку факториальная дисперсия определяется влиянием регулируемого фактора, то её доля от общей вариации значений признака является количественной характеристикой силы действия фактора на признак при данных условиях. Рассчитаем показатель силы влияния фактора для нашего примера, воспользовавшись таблицей 6.2.

$$h^2 = \frac{10.52}{(10.52 + 0.569)} \cdot 100\% = 94.9\%$$

Таким образом, доля объясненной влиянием минеральных удобрений вариации в урожайности озимой ржи составляет 94.9%. Оставшиеся 5.1% вариации признака определяются действием случайных факторов.

6.5. Непараметрический однофакторный дисперсионный анализ

Данный вариант метода приходится использовать в случае, когда не выполняются требования к проведению параметрического дисперсионного анализа, в частности предположение о нормальности распределения результативного признака, или когда используемые данные «низкого качества» (порядковые).

Непараметрической альтернативой дисперсионного однофакторного анализа является метод Краскела – Уоллиса (ANOVA Kruskal – Wallis). При использовании непараметрического дисперсионного анализа фактически проверяемые гипотезы те же самые (фактор влияет или не влияет на признак), но метод расчета иной: сравниваются не групповые средние значения признака, а суммы рангов для разных групп, если они значительно отличаются друг от друга, значит фактор оказывает влияние на признак. Технически непараметрических анализ можно провести в пакете STATISTICA или в программе анализа данных ATTESTAT.

Для примера оценим влияние микробиологического препарата, содержащего азотфиксирующие бактерии, на прудовой фитопланктон, состоящий главным образом из зеленых хлорококковых водорослей. В результате проведения эксперимента *in situ* были получены следующие данные о влиянии различных разведений препарата (1 часть препарата : n частей прудовой воды) на концентрации хлорофилла «а» в прудовой воде (табл. 6.5). Эксперимент проведен в 3-х повторностях для каждого варианта.

Таблица 6.5

Изменение концентраций хлорофилла «а» (мкг/л) в вариантах эксперимента без добавления и с добавлением в разных разведениях микробиологического препарата

<i>Контроль (без добавки препарата)</i>	<i>Добавка препарата в разведении 1:40 000</i>	<i>Добавка препарата в разведении 1:4000</i>	<i>Добавка препарата в разведении 1:400</i>
34.6	32.1	54.6	78.5
39.3	34.3	53.6	110.0
34.9	39.3	56.0	85.0

Таблицу результатов можно перенести в табличный процессор MS EXCEL и воспользоваться программой ATTESTAT и модулем Дисперсионный анализ (рис. 6.8).

В строку Интервал данных необходимо ввести указателем мыши интервал ячеек из электронной таблицы MS EXCEL с исходными числовыми данными (столбцы с повторностями концентраций хлорофилла «а»). В Интервале вывода следует

указать любую свободную ячейку в электронной таблице MS EXCEL, начиная с которой будут выведены вычисленные статистические показатели. Далее в группе **Непараметрические методы** ставим флажок напротив **Критерий Краскела – Уоллиса** и нажимаем на кнопку **Выполнить расчет** (рис. 6.7). Результаты появятся в ячейках электронной таблицы MS EXCEL в текстовой форме (рис. 6.8).

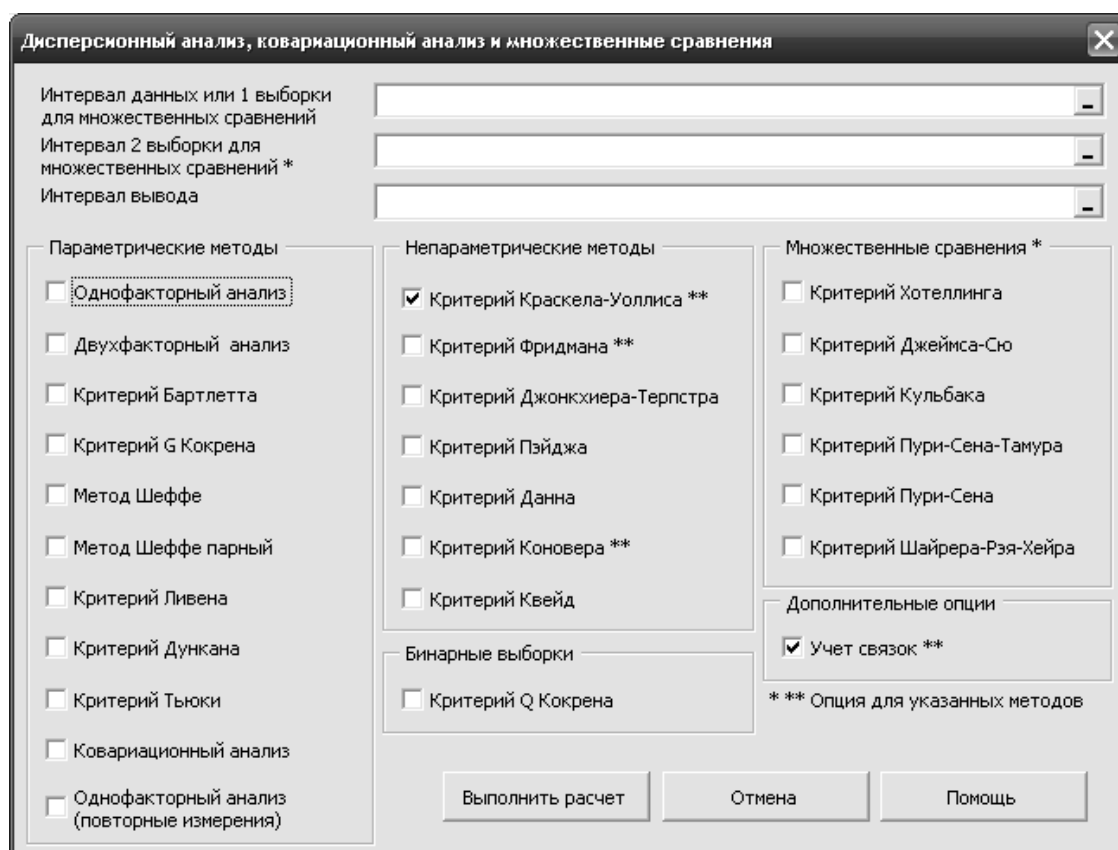


Рис. 6.7. Диалоговое окно модуля «Дисперсионный анализ» программы анализа данных ATTESTAT

Поскольку фактический p -уровень значимости (0.021) меньше критического (0.05), следовательно добавление микробиологического препарата достоверно повлияло на концентрации хлорофилла «а» в прудовой воде (рис. 6.8). Как видно из таблицы 6.5, воздействие препарата в целом оказало стимулирующий эффект на прудовой фитопланктон.

	F	G	H
74	Выдача включает:		
75	[Номер], [номер], статистика, Р-значение, [степени свободы]		
76	Ранговый однофакторный анализ Краскела-Уоллиса		
77	С учетом связей		
78	9,666666667	0,021245675	3
79			
80			
81			
82			
83			
84			
85			
86			
87			

Рис. 6.8. Результаты непараметрического дисперсионного анализа Краскела – Уоллиса

Примечание 4. В случае непараметрического дисперсионного анализа при отклонении нулевой гипотезы и необходимости проведения апостериорных множественных сравнений групповых средних значений признака можно применить непараметрический критерий Данна (Dunn's multiple comparison post-test) (Гланц, 1999).

Суть более сложных схем дисперсионного анализа (двухфакторный и многофакторный), позволяющих оценивать не только влияние каждого фактора по отдельности, но и их взаимодействие (сочетанное действие), значительно не отличается от рассмотренного в данном пособии алгоритма однофакторного дисперсионного анализа.

Заключение

В пособии сделана попытка изложить базовые понятия биометрии и элементарные методы количественного анализа данных наблюдений и экспериментов применительно к биологическим и экологическим исследованиям в доступной для читателя форме. Усвоение студентами таких понятий, как стандартная ошибка, доверительный интервал, дисперсия, доверительная вероятность, уровень значимости, статистические гипотезы, критерии достоверности, законы распределения, является ключевым этапом для последующего самостоятельного освоения более сложных методов одномерной и многомерной статистики и планирования экспериментов. Во многих аналогичных пособиях этому уделяется меньше внимания, как правило, приводятся готовые формулировки без подробного разъяснения сути базовых понятий. Кроме того, приведенные примеры использования компьютерных программ анализа данных при решении исследовательских задач являются элементарной основой для практического освоения учащимися методов биометрии. И помните: освоение биометрии лишь инструмент, позволяющий количественно измерить и увидеть в собранном материале скрытые от исследователя закономерности. Главное остается в том, что измерять и на что смотреть при решении научно-исследовательских задач, т. е. в первую очередь нужно быть вдумчивым исследователем и только потом умелым статистиком или математиком!

Вопросы к экзамену

1. Основные понятия биометрии (статистическая совокупность, единица наблюдения, признак, варьирование признаков и их причины). Ошибки измерений.

2. Типы экологических данных. Статистические ряды и их графики.

3. Выборочный и сплошной методы исследования, преимущества и недостатки. Понятие генеральной совокупности и выборки, примеры.

4. Репрезентативность выборок. Способы взятия выборок из генеральной совокупности.

5. Степенные и структурные средние величины, формулы расчета и значение при обработке экологических данных.

6. Показатели вариации, формулы расчета и значение при обработке экологических данных.

7. Понятие вероятности. Априорная и апостериорная вероятность, примеры. Закон нормального распределения признаков, параметры нормального распределения.

8. Правило 3-х сигм, его практическое применение. Эмпирическое и теоретическое распределение признаков (экологических показателей). Понятие асимметрии и эксцесса эмпирического распределения.

9. Статистическое оценивание генеральных параметров. Точечные и интервальные оценки. Понятие доверительной вероятности и уровня значимости при расчете доверительных интервалов для выборочных средних значений.

10. Основные задачи, решаемые при статистических сравнениях. Понятие достоверности выборочной разности. Нулевая и альтернативная гипотезы. Понятие критерия достоверности.

11. Понятие уровня значимости применительно к критериям достоверности. Классификации критериев достоверности. Преимущества и недостатки параметрических и непараметрических критериев достоверности.

12. Способы проверки нормальности эмпирического распределения признака. Критерии согласия (нормальности), условия их применимости.

13. Параметрические критерии различий: t-критерий Стьюдента и F-критерий Фишера. Область использования, формулы расчета, условия применимости.

14. Непараметрические критерии различий: критерий Манна–Уитни, критерий Вилкоксона, критерий знаков и критерий серий Вальда–Вольфовица. Область использования, условия применимости.

15. Дисперсионный анализ. Сущность метода. Основные понятия и термины (результативный признак, фактор, градации фактора, дисперсионный комплекс, их виды). Нулевая и альтернативная гипотезы в дисперсионном анализе.

16. Основные этапы дисперсионного анализа.

17. Виды дисперсионного анализа. Условия применимости классического параметрического дисперсионного анализа. Непараметрический дисперсионный анализ.

18. Понятия «функциональная связь» и «корреляция», примеры. Основные этапы корреляционного анализа.

19. Значение коэффициента корреляции, виды, градация, условия применимости. Понятие коэффициента детерминации.

20. Определение достоверности коэффициента корреляции. Корреляция и причинно-следственная зависимость. Понятие ложной и частной корреляции.

21. Понятие о регрессии. Сущность регрессионного анализа и область его применения. Основные этапы регрессионного анализа.

22. Виды регрессионных связей, уравнения, графические модели регрессии.

23. Определение достоверности параметров регрессионного уравнения и адекватности уравнения регрессии. Регрессия и выбросы. Регрессия и неоднородность выборки.

24. Анализ временных рядов, основные этапы.

25. Множественная регрессия, сущность, уравнение. Методы пошаговой регрессии.

26. Отличие многомерных методов анализа от одномерных. Сущность кластерного, дискриминантного и факторного анализов.

27. Основы теории планирования экспериментов. Виды экспериментов. Проблема мнимых повторностей. Схема полного факторного эксперимента. Статистические методы обработки экспериментальных данных.

28. Моделирование как метод исследования сложных систем. Классификация моделей. Статистическое, аналитическое и имитационное моделирование.

29. Математические модели роста популяций, моделирование экосистем.

Рекомендуемая литература

1. Апостолов, Л. Г. Математические методы в экологии / Л. Г. Апостолов, А. В. Ивашов. – Симферополь : СГУ, 1981.
2. Баканов, А. И. Основные источники ошибок в гидробиологических и ихтиологических исследованиях / А. И. Баканов, М. М. Сметанин, Н. М. Шихова // Биология внутренних вод. – 2001. – № 4. – С. 79–87.
3. Бейли, Н. Статистические методы в биологии / Н. Бейли. – М. : Изд-во иностр. лит-ры, 1962.
4. Боровиков, В. Statistica для профессионалов: искусство анализа данных на компьютере / В. Боровиков. – СПб. : Питер, 2001.
5. Вадзинский, Р. Статистические вычисления в среде Excel / Р. Вадзинский. – СПб. : Питер, 2008.
6. Василевич, В. И. Статистические методы в геоботанике / В. И. Василевич. – Л. : Наука, 1969.
7. Владимирский, Б. М. Математические методы в биологии / Б. М. Владимирский. – Ростов н/Д. : Изд-во Ростовского ун-та, 1983.
8. Гайдышев, И. Анализ и обработка данных: специальный справочник / И. Гайдышев. – СПб. : Питер, 2001.
9. Гланц, С. Медико-биологическая статистика / С. Гланц. – М. : Практика, 1999.
10. Гмурман, В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. – М. : Высшая школа, 2001.
11. Ивантер, Э. В. Введение в количественную биологию / Э. В. Ивантер, А. В. Коросов. – Петрозаводск : ПетрГУ, 2003.
12. Ивантер, Э. В. Элементарная биометрия / Э. В. Ивантер, А. В. Коросов. – Петрозаводск : ПетрГУ, 2005.
13. Лакин, Г.Ф. Биометрия / Г.Ф. Лакин. – М. : Высшая школа, 1990.
14. Малета, Ю. С. Непараметрические методы статистического анализа в биологии и медицине / Ю. С. Малета, В. В. Тарасов. – М. : Изд-во Московского ун-та, 1982.
15. Платонов, А. Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы / А. Е. Платонов. – М. : РАМН, 2000.
16. Плохинский, Н. А. Математические методы в биологии / Н. А. Плохинский. – М. : Изд-во Московского ун-та, 1978.

17. Поморский, Ю. Л. Методы биометрических исследований / Ю. Л. Поморский. – Л., 1935.
18. Пузаченко, Ю. Г. Математические методы в экологических и географических исследованиях / Ю. Г. Пузаченко. – М. : Академия, 2004.
19. Реброва, О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. – М. : МедиаСфера, 2002.
20. Рокицкий, П. Ф. Биологическая статистика / П. Ф. Рокицкий. – Минск : Вышэйшая школа, 1967.
21. Сиделев, С. И. О задачах, проблемах и методике преподавания дисциплины «Математические методы в биологии и экологии» студентам Ярославского государственного университета / С. И. Сиделев, А. А. Зубишина // Современные проблемы биологии, экологии, химии: материалы Межд. науч.-практ. конф. – Ярославль, 2011. – С. 326–333.
22. Терентьев, П. В. Практикум по биометрии / П. В. Терентьев, Н. С. Ростова. – Л. : Наука, 1977.
23. Тихонов, С. В. Практические занятия по математическим методам в биологии и экологии / С. В. Тихонов. – Ярославль : ЯрГУ, 2003.
24. Тюрин, Ю. Н. Статистический анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. – М. : ИНФРА, 2003.
25. Урбах, В. Ю. Статистический анализ в биологических и медицинских исследованиях / В. Ю. Урбах. – М. : Медицина, 1975.
26. Урбах, В. Ю. Биометрические методы (статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине) / В. Ю. Урбах. – М. : Наука, 1964.
27. Халафян, А. А. STATISTICA 6. Статистический анализ данных / А. А. Халафян. – М. : Бином-Пресс, 2007.
28. Шитиков, В. К. Количественная гидроэкология: методы системной идентификации / В. К. Шитиков, Г. С. Розенберг, Т. Д. Зинченко. – Тольятти : ИЭВБ РАН, 2003.
29. Шмидт, В. М. Математические методы в ботанике / В. М. Шмидт. – Л. : Изд-во Ленинградского ун-та, 1984.

Оглавление

Введение	3
Глава 1. Общие вопросы применения количественных методов в биологии и экологии	5
1.1. Роль статистических методов в биологии и экологии.....	5
1.2. Программное обеспечение анализа данных	10
1.3. Несколько слов о терминологии	18
1.4. Характер биологических и экологических данных	20
1.5. Выборочный метод исследования	22
Глава 2. Приемы первичной статистической обработки данных.....	32
2.1. Статистические ряды	34
2.2. Графический анализ	36
2.3. Таблицы	41
2.4. Статистические характеристики выборочной совокупности, или как сжато описать данные	45
Глава 3. Законы распределения биологических и экологических переменных.....	59
3.1. Вероятность события	59
3.2. Закон распределения	61
3.3. Нормальное распределение	65
3.4. Понятие асимметрии и эксцесса распределения.....	69
3.5. Биномиальное распределение	71
3.6. Другие типы теоретических распределений.....	73
Глава 4. Статистические оценки генеральных параметров, или насколько точно данные выборки соответствуют реальности	76
4.1. Стандартная ошибка среднего значения.....	77
4.2. Доверительный интервал для среднего значения	81
Глава 5. Проверка статистических гипотез.....	87
5.1. Достоверность выборочной разности. Нулевая и альтернативная гипотезы. Понятие критерия достоверности	89
5.2. Классификация критериев достоверности	93

5.3. Проверка нормальности распределения в пакете STATISTICA	99
5.4. Использование параметрических критериев в MS EXCEL	102
5.5. Использование непараметрических критериев в пакете STATISTICA	107
5.6. Браковка выбросов и критерии исключения	110
Глава 6. Количественная оценка влияния фактора	111
6.1. Сущность метода	111
6.2. Базовая терминология дисперсионного анализа.....	116
6.3. Условия применимости и основные этапы дисперсионного анализа	117
6.4. Однофакторный дисперсионный анализ в среде MS EXCEL и в пакете STATISTICA	118
6.5. Непараметрический однофакторный дисперсионный анализ.....	128
Заключение	132
Вопросы к экзамену	133
Рекомендуемая литература	135

Учебное издание

Сиделев Сергей Иванович

**Математические методы
в биологии и экологии:
введение в элементарную биометрию**

Учебное пособие

Редактор, корректор М. Э. Левакова
Верстка Е. Л. Шелехова

Подписано в печать 08.02.12. Формат 60×84 ¹/₁₆.
Бум. офсетная. Гарнитура "Times New Roman".
Усл. печ. л. 8,14. Уч.-изд. л. 7,05.
Тираж 50 экз. Заказ

Оригинал-макет подготовлен
в редакционно-издательском отделе Ярославского
государственного университета им. П. Г. Демидова.

Отпечатано на ризографе.

Ярославский государственный университет им. П. Г. Демидова.
150000, Ярославль, ул. Советская, 14.

С. И. Сиделев

**Математические методы
в биологии и экологии:
введение в элементарную биометрию**

