РУП «ИНСТИТУТ РЫБНОГО ХОЗЯЙСТВА»
РУП «НАУЧНО-ПРАКТИЧЕСКИЙ ЦЕНТР НАЦИОНАЛЬНОЙ
АКАДЕМИИ НАУК БЕЛАРУСИ ПО ЖИВОТНОВОДСТВУ»

Мастицкий С. Э.

МЕТОДИЧЕСКОЕ ПОСОБИЕ
по использованию программы STATISTICA
при обработке данных биологических исследований

Рукопись

Минск
РУП «Институт рыбного хозяйства»
2009

УДК 57:519.24

Мастицкий С. Э. Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований. – Мн.: РУП «Институт рыбного хозяйства». – 76 С.

В пособии рассмотрены типовые задачи, с которыми сталкиваются исследователи-биологи в ходе статистической обработки результатов наблюдений и экспериментов (расчет параметров описательной статистики, сравнение двух и более групп, корреляционный и регрессионный анализы). Представлены пошаговые описания решений этих задач с использованием пакета прикладных программ STATISTICA. Пособие предназначено для студентов, аспирантов, и научных работников биологических специальностей.

Рецензент
канд. биол. наук, доцент кафедры общей экологии и методики преподавания биологии Белгосуниверситета *Т. А. Макаревич*

Рукопись

© Мастицкий С. Э.
РУП «Институт рыбного хозяйства»

Содержание

Предисловие 3
Глава 1. Ознакомление с интерфейсом программы. Создание и сохранение файла 5
1.1. Рабочее окно программы STATISTICA *5*
1.2. Создание и сохранение файлов *6*
Глава 2. Описательная статистика 8
2.1. Подготовка таблицы к вводу данных *8*
2.2. Полигон распределения *10*
2.3. Гистограммы *15*
2.4. Расчет параметров описательной статистики *17*
2.5. Диаграммы диапазонов *21*
2.4. Диаграммы размахов *23*
2.5. Круговые диаграммы *26*
Глава 3. Проверка соответствия анализируемых данных закону нормального распределения 28
3.1. О необходимости проверки нормальности распределения анализируемых данных *28*
3.2. Подгонка распределения *28*
3.3. Тесты Колмогорова-Смирнова и Шапиро-Уилка *30*
3.4. График нормальных вероятностей *32*
Глава 4. Сравнение двух групп 33
4.1. Случай независимых выборок *33*
4.2. Случай зависимых выборок *38*
4.3. Сравнение выборочной средней с константой *40*

Рукопись

Глава 5. Сравнение нескольких групп 43
5.1. Параметрический однофакторный дисперсионный анализ 43
5.2. Апостериорный анализ 47
5.3. Параметрический двухфакторный дисперсионный анализ 48
5.4. Дисперсионный анализ Фридмана 50
5.5. Дисперсионный анализ Крускала-Уоллиса..... 52
Глава 6. Корреляционный анализ..... 54
6.1. Коэффициент корреляции Пирсона 55
6.2. Сравнение двух коэффициентов корреляции Пирсона..... 57
6.3. Коэффициент корреляции Спирмена 59
6.4. Коэффициент ассоциации (связанности) 60
Глава 7. Регрессионный анализ..... 62
7.1. Оценка коэффициентов линейной регрессии 62
7.2. Трансформация нелинейно связанных признаков 68
7.3. Оценка коэффициентов уравнения нелинейной зависимости..... 72
Список использованных источников 75
Предметный указатель..... 76

Рукопись

Предисловие

Для того чтобы запечатлеть понравившуюся нам сцену можно использовать обычный пленочный фотоаппарат. Открывая его затвор, мы собираем и определенным образом фиксируем визуальную информацию на поверхности фотопленки. Однако этого не достаточно. Фотография не появится, если пленка не будет должным образом обработана. Есть и еще один этап: получившуюся фотографию можно снабдить красивой рамкой, чтобы она понравилась любому, кто на нее взглянет.

Через аналогичные стадии проходит любой современный ученый, в том числе и биолог, работая с числовой информацией. Сначала он *собирает* данные, например, измеряя линейные размеры какого-либо органа, определяя концентрацию глюкозы в крови, подсчитывая число организмов в учетной рамке. Дальше данные должны быть *обработаны* с применением методов статистики. На этом этапе становится ясно, о чем «говорят» полученные цифры. Чтобы «заставить» их «говорить», возможно, потребуется построить график или выполнить более сложный анализ. Наконец, извлеченную из данных информацию необходимо будет корректно *представить* тому, кто в ней заинтересован (коллегам-ученым, менеджерам, и т.п.).

Обработка числовой информации в наши дни немыслима без применения компьютера. Современный специалист-биолог обязан обладать навыками компьютерной обработки данных и иметь представление о программном обеспечении, с помощью которого ее можно выполнять. Сегодня существует большое количество специализированных приложений для статистического анализа. Одним из несомненных лидеров среди таких продуктов признана программа STATISTICA фирмы StatSoft, Inc., США. Помимо очень мощного набора процедур статистического и графического анализа, эта программа обладает весьма дружественным интерфейсом, что делает ее достаточно легкой для освоения и удобной в работе.

В последние годы было издано несколько руководств по работе с программой STATISTICA (Боровиков 1998, 2003; Боровиков, Ивченко 2000; Реброва 2003), среди которых, однако, практически нет пособий, учитывающих специфику

Рукопись

биологических исследований и получаемых в их ходе данных. Цель настоящего пособия – помочь широкому кругу биологов приступить к освоению программы STATISTICA и начать активно использовать ее в своей работе. Несмотря на то, что совсем недавно была выпущена уже 9-я версия программы, данное пособие посвящено описанию 6-й версии, как наиболее распространенной в отечественных университетах и исследовательских учреждениях. В связи с ограниченным объемом, пособие содержит примеры и пошаговые описания решений лишь *типовых задач*, возникающих в биологических исследованиях (описательная статистика, сравнение двух и более групп, корреляционный и регрессионный анализы). Кроме того, предполагается, что читатель

- прослушал как минимум вводный университетский курс биологической статистики,
- имеет навыки работы в среде Microsoft Office,
- уже установил STATISTICA 6.0 (в ее оригинальном англоязычном варианте) на своем компьютере с операционной системой Windows.

Некоторые факты из истории развития программы STATISTICA (по: Боровиков 1998, с дополнениями):

1991 г. – выход первой версии программы для DOS
1992 г. – выход первой версии программы для Macintosh
1993 г. – выходит первая версия STATISTICA для Windows
1994 г. – в результате сравнительного тестирования с пакетами BMDP 1.0, SPSS 6.1, Statgraphics 1.0 и Systat 5.01 STATISTICA получает первое место в нескольких научных и компьютерных изданиях (INSIGHT, MacWELT, C'T Magazine, WINDOWS Magazine)
1995 г. – программа включена в список 100 лучших программных продуктов (WINDOWS Magazine)
2001 – выход 6-й версии STATISTICA
Май 2009 – выход 9-й версии программы

РУКОПИСЬ

Глава 1. Ознакомление с интерфейсом программы. Создание и сохранение файла

1.1. Рабочее окно программы STATISTICA

Программа STATISTICA, являясь продуктом американской компании, имеет англоязычный интерфейс. Существующие русификаторы использовать не рекомендуется, поскольку в ряде случаев они грешат некорректным переводом статистических терминов. Учитывая то, что английский язык может вызвать затруднения у ряда читателей, все важные термины и опции меню будут при изложении материала сопровождаться переводом.

Запустив программу, вы увидите, что ее рабочее окно похоже на окна всех Windows-приложений. В самом верху слева находится заголовок окна в формате «*Statistica – Data*: Имя файла.*sta* (размер таблицы)». Далее следует строка основного меню, ряд разделов которого также стандартен для Windows-приложений: *File* (Файл), *Edit* (Правка), *View* (Вид), *Insert* (Вставка), *Format* (Формат), *Tools* (Инструменты), *Window* (Окно), *Help* (Помощь). Имеются, однако, и специфические разделы – *Statistics* (Статистические процедуры), *Graphs* (Графики), *Data* (Данные). За строкой меню следуют настраиваемая пользователем панель инструментов и рабочая область, занимающая большую часть окна программы (рис. 1.1).

Анализируемые данные хранятся в STATISTICA в виде электронной таблицы, подобно тому, как это происходит, например, в программе MS Excel. Однако таблица с данными в STATISTICA, которая носит название *Spreadsheet*, имеет свои особенности. В отличие от обычных электронных таблиц, в которых столбцы и строки равноправны, в таблице программы STATISTICA столбцы называются *Variables* (Переменные), а строки – *Cases* (Наблюдения). В качестве переменных выступают исследуемые признаки (например, рост, вес, концентрация, скорость и т.п.). Под наблюдениями же понимаются конкретные значения, которые принимают переменные. Важно отметить, что программа STATISTICA может обрабатывать не только числовые, но и текстовые данные, что очень удобно при работе с качественными признаками. Кроме того, таблицы *Spreadsheet* поддерживают

Рукопись

различные стандартные операции с ячейками, такие как выделение и перетаскивание диапазона, автозамена, копирование/вставка, импорт из других приложений (например, из MS Excel, Access) и др.

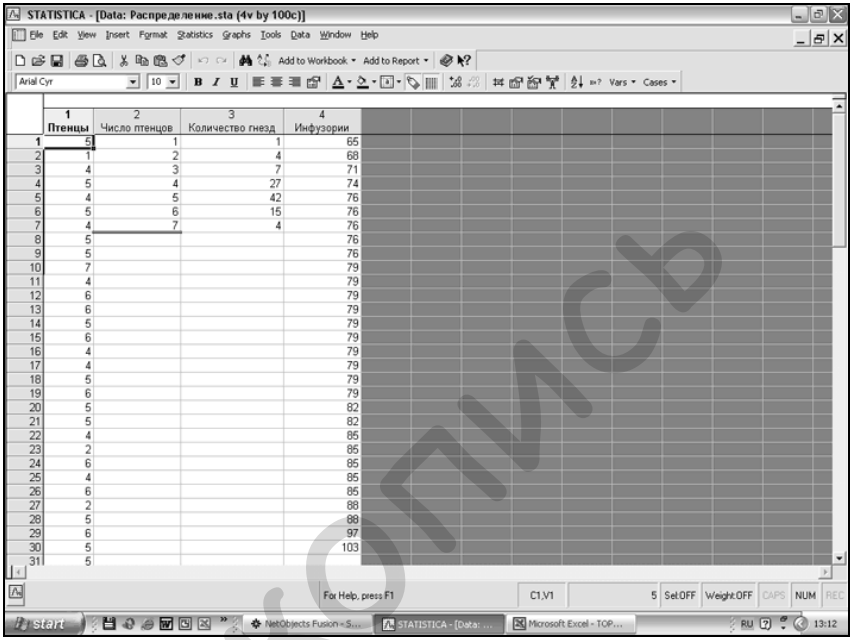


Рисунок 1.1. Внешний вид рабочего окна STATISTICA.

РукоПись

1.2. Создание и сохранение файлов

Запустите программу STATISTICA (из меню Windows «Пуск» или кликнув по соответствующему ярлыку на Рабочем столе). По умолчанию откроется последний файл, с которым выполнялась работа в ходе предыдущего сеанса (если таковой имеется). Закройте этот файл и создайте новый. Для этого можно воспользоваться одним из трех способов:

- В пункте основного меню *File* (Файл) выбрать *New* (Новый);
- Нажать кнопку [□] на панели инструментов;
- Применить сочетание клавиш «*Ctrl + N*».

В результате появится диалоговое окно создания нового документа (*Create new document*; рис. 1.2), в котором необходимо указать, какой именно документ создается. Мы создаем новую таблицу с данными, поэтому останемся на закладке *Spreadsheet*, которая по умолчанию предстает перед пользователем первой. Пусть в таблице будет 1 столбец и 100 строк. Чтобы сообщить об этом программе, в поле *Number of variables* (Количество переменных) выставим 1, а в поле *Number of cases* (Количество наблюдений) – 100. Остальные опции этой закладки оставим без изменений (поле *Placement* (Расположение): *As a stand-alone window* (Как самостоятельное окно)). После нажатия кнопки *OK* (или клавиши «Ввод» на клавиатуре) в рабочей области программы появится таблица с 1 столбцом и 100 строками. Сохраним созданный файл под именем «Распределение». Для этого можно воспользоваться тремя способами:

- В пункте основного меню *File* (Файл) выбрать *Save* (Сохранить);
- Нажать кнопку на панели инструментов;
- Применить сочетание клавиш «*Ctrl + S*».

При этом появится стандартное для Windows диалоговое окно, в котором необходимо указать имя нового файла, а также место, в котором он будет храниться.

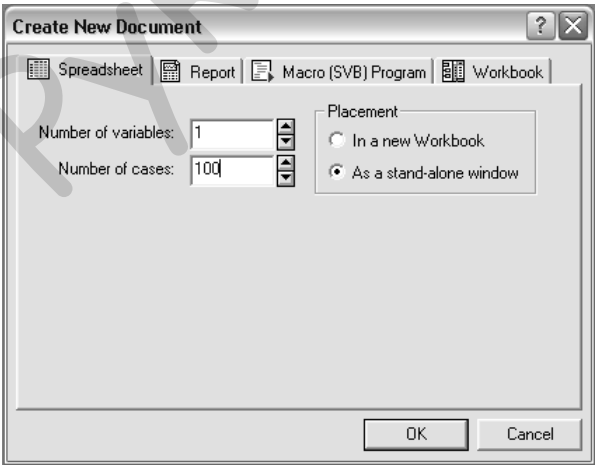


Рисунок 1.2. Диалоговое окно создания нового документа

Рукопись

Глава 2. Описательная статистика

2.1. Подготовка таблицы к вводу данных

Для быстрого ознакомления с полученными данными и выявления в них явных закономерностей на начальном этапе статистического анализа бывает полезно построить гистограмму или полигон распределения, которые представляют собой графики, отражающие связь между значениями изучаемого биологического признака и частотой встречаемости этих значений. Сейчас мы построим полигон распределения для представленных ниже данных о количестве птенцов в 100 гнездах лесной ласточки *Iridoprocne bicolor*:

5	4	5	5	4	5	4	3	5	4	7	5	6
1	6	4	4	4	5	5	3	5	5	5	5	5
4	6	2	3	4	5	5	5	5	5	5	4	4
5	5	6	4	6	2	5	5	3	5	3	7	3
4	6	4	5	5	5	5	5	5	5	6	4	
5	4	6	7	6	3	5	5	6	5	5	6	
4	4	2	4	4	6	2	6	5	4	6	4	
5	5	5	4	5	4	6	5	4	7	4	4	

Перед вводом этих данных необходимо выполнить определенную предварительную подготовку электронной таблицы. Воспользуемся уже созданным нами ранее файлом «Распределение.*sta*» (разд. 1.2). Обратите внимание на заголовок единственного столбца в таблице этого файла. Он выделен серым цветом и помимо порядкового номера содержит имя «*Var*1» (от англ. *Variable* – переменная). Когда в таблице есть лишь один столбец, никакой путаницы, конечно, не возникает. Однако если ничего не менять, она обязательно появится при большом количестве переменных. Чтобы этого избежать, столбцам полезно присваивать уникальные (не повторяющиеся) имена. Для переименования переменной подведите курсор мыши к заголовку соответствующего столбца и дважды кликните по нему. Появится окно, в котором осуществляется настройка свойств переменной (рис. 2.1). Имя переменной указывается в поле *Name*. Установите курсор в это поле и наберите слово «Птенцы». Формат надписи (шрифт, его размер

Рукопись

и т.п.) можно задать с помощью стандартных инструментов для форматирования текста, расположенных выше.

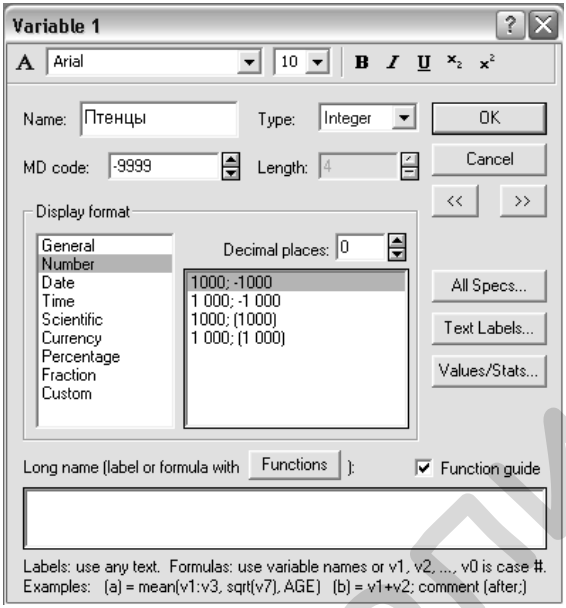


Рисунок 2.1. Окно настройки свойств переменной.

В поле *Type* (Тип), расположенном справа от Name, указывается тип переменной. По умолчанию он выставлен на *Double* (Двойной), что подходит для случаев, когда значения переменной выражаются числами, лежащими в интервале $\pm 1{,}7\times 10^{308}$. Если анализируемые данные представляют собой только целые числа из интервала ±2 147 483 648, то следует выбрать тип *Integer* (Целое). Для переменных, которые выражаются целыми числами от 0 до 255 включительно можно установить специальный тип *Byte* (Байт). Наконец, если переменная содержит текстовые значения, то выбирают *Text*. Поскольку количество птенцов ласточки по определению может выражаться только целыми числами, в поле *Type* выберем *Integer*.

Более тонкая настройка типа переменной выполняется далее в поле *Display format* (Формат отображения). Поскольку мы намерены работать с числами, то выберем здесь *Number* (Число). Справа появится дополнительное поле *Decimal Places* (Десятичные знаки) – в нем указывается точность, с которой мы

Рукопись

хотим видеть в таблице наши данные, а также формат внешнего вида чисел. Установите количество десятичных знаков на 0.

Среди остальных элементов рассматриваемого окна особого внимания заслуживает поле *Long name* (Длинное имя). Его можно использовать как записную книжку, в которой бывает удобно оставить свои заметки о ходе анализа. Кроме того, здесь можно ввести формулу, в соответствии с которой будут пересчитаны значения переменной. Пример использования формул в программе STATISTICA будет рассмотрен ниже (разд. 7.2). После настроки переменной нажмите на кнопку *OK* и введите в таблицу приведенные выше даные о количестве птенцов ласточки (все 100 значений – в один имеющийся в таблице столбец).

2.2. Полигон распределения

В программе STATISTICA реализован графически-ориентированный подход к анализу данных. В связи с этим она обладает внушительным набором различных типов графиков, которые можно построить, обратившись к разделу главного меню *Graphs* или к соответствующим закладкам того или иного статистического модуля.

Продолжим работу с созданным ранее файлом «Распределение.*sta*», который уже содержит данные о количестве птенцов лесной ласточки (разд. 2.1). Для того чтобы программа смогла построить полигон распределения, из имеющихся данных нужно сформировать *вариационный ряд*, т.е. двойной ряд чисел, в котором содержатся значения анализируемого признака и частоты их встречаемости. Перед тем как сделать это, добавим два столбца в нашу таблицу. Подведите курсор к заголовку столбца «Птенцы» и нажмите правую клавишу мыши. В появившемся контекстном меню выберите пункт *Add variables* (Добавить переменные). Далее появится диалоговое окно, в котором нужно указать количество добавляемых переменных (поле *How many* – выставляем «2») и их положение в таблице (поле *After* (После) – наберите слово «Птенцы»). Остальные настройки можно оставить без изменений. В таблице появятся два новых столбца с названиями «*NewVar1*» и «*NewVar2*». Переименуйте их самостоятельно в

Рукопись

«Число птенцов» и «Количество гнезд» соответственно (разд. 2.1).

Теперь сформируем вариационный ряд. Для этого в разделе основного меню *Statistics* (Статистические процедуры) выберем модуль *Basic Statistics/Tables* (Основные статистические показатели/Таблицы), а в нем – опцию *Frequency tables* (Таблицы частот). В появившемся диалоговом окне необходимо указать, какую именно переменную мы собираемся анализировать. Для этого служит кнопка *Variables* (Переменные) – она будет очень часто встречаться нам в дальнейшем (рис. 2.2). При нажатии на нее появится еще одно окно (*Select the variables for the analysis*), основная часть которого занята списком имеющихся в таблице переменных. Дважды кликните по пункту «Птенцы», а затем нажмите либо кнопку *Summary: Frequency tables* (Результат: Таблицы частот), либо *Summary* (Результат), либо просто клавишу «Ввод» на клавиатуре.

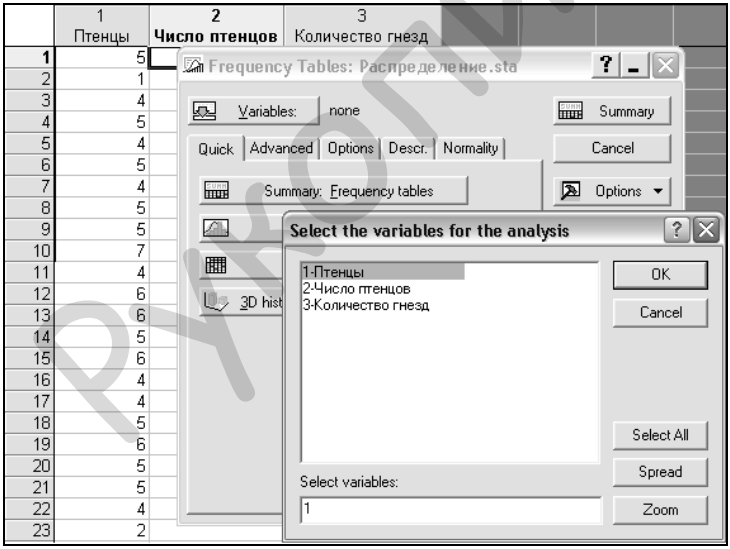


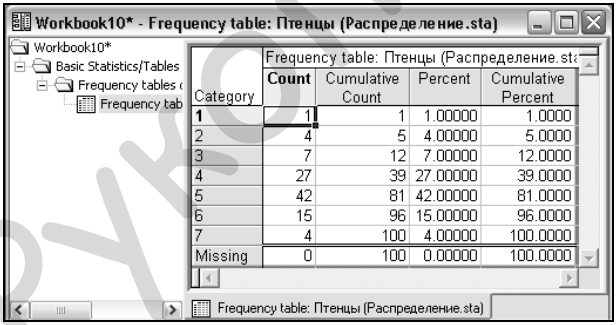
Рисунок 2.2. Выбор переменной для расчета таблицы частот.

В итоге программа сформирует таблицу, представляющую собой «расширенный» вариант вариационного ряда (рис. 2.3). В этой таблице имеются следующие столбцы:

РУКОПИСЬ

- *Category* (Категория): содержит ранжированные значения анализируемой переменной, отмеченные в выборке. В случае с нашим примером видим, что число птенцов в гнездах лесной ласточки изменялось от 1 до 7.
- *Count* (Счет): здесь приведены частоты, с которыми встречались отмеченные значения переменной (так, было найдено только 1 гнездо с 1 птенцом, 4 гнезда с 2 птенцами, 7 гнезд с 3 птенцами и т.д.).
- *Cumulative count*: накопленные частоты.
- *Percent*: процент, который составляет каждая из частот от общего числа наблюдений.
- *Cumulative percent*: накопленные процентные доли частот.

Последняя строка итоговой таблицы называется *Missing* (Отсутствующие) – она имеет отношение к неотмеченным в быборке значениям переменной. Таковых в нашем примере нет (встречались все возможные значения числа птенцов – от 1 до7), в связи с чем на пересечении столбца *Count* и строки *Missing* видим 0.



Category	Count	Cumulative Count	Percent	Cumulative Percent
1	1	1	1.00000	1.0000
2	4	5	4.00000	5.0000
3	7	12	7.00000	12.0000
4	27	39	27.00000	39.0000
5	42	81	42.00000	81.0000
6	15	96	15.00000	96.0000
7	4	100	4.00000	100.0000
Missing	0	100	0.00000	100.0000

Рисунок 2.3. Таблица частот для данных о числе птенцов в гнездах лесной ласточки.

Обратите внимание: итоговая таблица анализа *Frequency Tables* является частью окна с заголовком *Workbook* (Рабочая книга). Такая форма вывода результатов очень удобна и является характерной особенностью программы STATISTICA. Результаты любого анализа, который в дальнейшем применялся бы к данным открытого в текущий момент файла, заносился бы в эту же рабочую книгу на отдельный лист. Каталог выполненных анализов отображается в отдельной области слева.

Рукопись

Рабочую книгу можно сохранить в виде самостоятельного файла (с расширением *.stw*) и при необходимости вернуться к ней просмотра результатов выполненного анализа.

Внесите «вручную» необходимые значения из полученной таблицы частот в нашу исходную таблицу с данными (это можно сделать и путем копирования, однако описание этой процедуры не приводится).

Теперь у нас есть все необходимое, чтобы построить полигон распределения. В разделе главного меню *Graphs* (Графики) выберите *2D Graphs* (Двухмерные графики) > *Line plots (Variables)* (Линейные графики (по переменным)).

В появившемся диалоговом окне (рис. 2.4) выберите закладку *Advanced* (Расширенные настройки). На ней в поле *Graph type* (Тип графика) выберите *XY Trace*, а в выпадающем меню *Display points* (Отображать точки) – *On* (Включить). Наконец, откройте закладку *Options* 1, разыщите на ней выпадающее меню *Case labels* (Подписи наблюдений) и выберите пункт *Off* (Отключить). Вы можете не выполнять последнюю операцию, если хотите, чтобы значения количества птенцов на графике отображались.

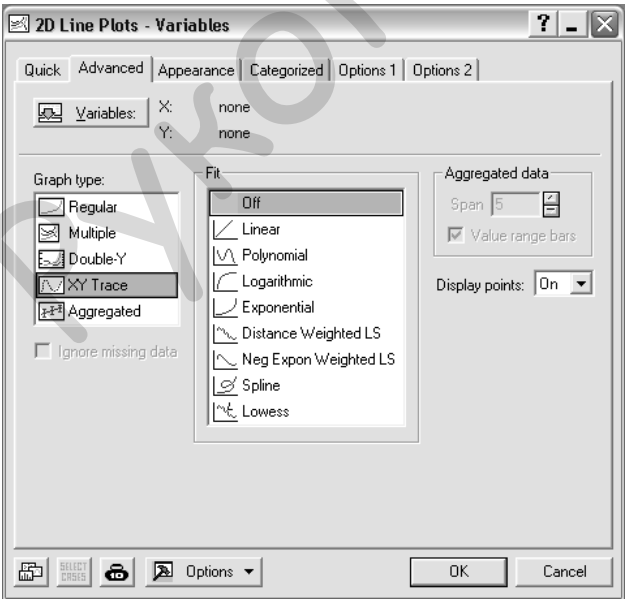


Рисунок 2.4. Диалоговое окно модуля *2D Line plots* на закладке *Advanced*.

Теперь необходимо «объяснить» программе, какой из столбцов таблицы с данными соответствуют числу птенцов (ось *X*), а какой – частоте встречаемости (ось *Y*). Для этого вернитесь на закладку *Advanced* и нажмите уже знакомую вам кнопку *Variables* (Переменные). В результате появится окно с двумя списками переменных. В левом списке выделите пункт «Число птенцов», а в правом – «Количество гнезд». Далее нажимаем кнопку *OK*, затем еще раз *OK*, и получаем долгожданный график (рис. 2.5). Заметьте: полученный график является составной частью рабочей книги *Workbook*, как это ранее было с итоговой таблицей анализа *Frequency Tables*.

Если кликнуть один раз по полученному графику правой кнопкой мыши и из контекстного меню выбрать опцию *Copy Graph*, можно скопировать график в буфер обмена и затем вставить в документ практически любого другого Windows-приложения, например, MS Word или Excel. Кроме того, график можно сохранить как самостоятельный файл с расширением *.stg*. Для этого необходимо выделить иконку графика в каталоге рабочей книги и, удерживая нажатой левую клавишу мыши, перетащить ее за пределы рабочей книги (рис. 2.5). В результате рисунок окажется в отдельном окне. Теперь, кликнув по нему правой кнопкой мыши, можно применить команду *Save Graph* (Сохранить график).

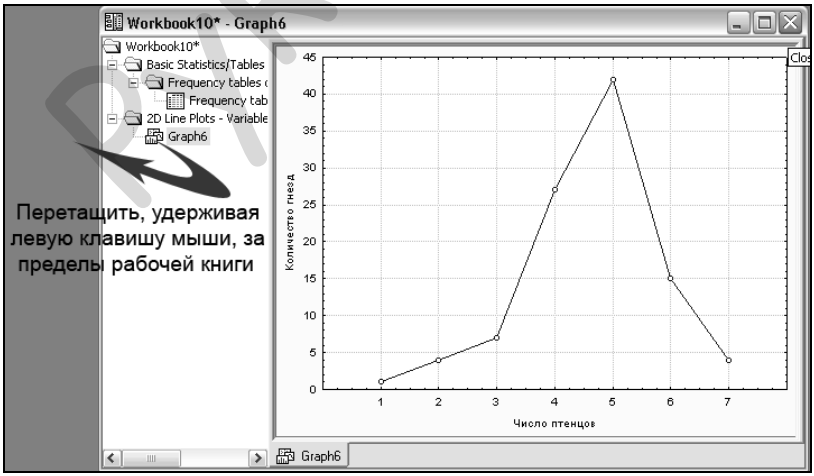


Рисунок 2.5. Полигон распределения для данных о числе птенцов лесной ласточки. См. объяснения в тексте.

Программа STATISTICA предоставляет широкие возможности для придания графикам необходимого внешнего вида. Достаточно кликнуть по интересующему вас элементу, и появится диалоговое окно с множеством опций по его настройке (заголовки, оси и их названия, маркеры и их форма, цвет и размер, и т.п.).

2.3. Гистограммы

В таблице ниже представлены результаты измерений длины клеток (в мкм) инфузории-комменсала *Conchophthirus acuminatus* из мантийной полости двустворчатого моллюска *Dreissena polymorpha*:

65	68	71	74	76	76	76	76	76	79	79	79	79	79	79
79	79	79	79	82	82	85	85	85	85	85	88	88	97	103

Размах значений длины клеток достаточно велик (103 - 65 = 38 мкм). Кроме того, в этом ряду некоторые значения отсутствуют (например, не встречены клетки с длиной 66, 78, 83 мкм). Для графического изображения частотного распределения в данном случае лучше подходит *гистограмма*, а не полигон распределения.

Создайте новый файл данных с одной переменной (назовите ее, например, «Инфузории») и 30 наблюдениями (см. разд. 1.2), присвойте ему имя (например, «Клетки.*sta*») и сохраните. Введите приведенные выше значения длины клеток в столбец этой таблицы данных. Для построения гистограммы достаточно выполнить следующие действия:

- В основном меню программы выбрать *Graphs* > 2*D Graphs* > *Histograms* (Гистограммы).
- В появившемся окне (рис. 2.6) выбрать закладку *Advanced*. Нажав на кнопку *Variables*, выбрать для анализа переменную «Инфузории». В поле *Fit type* (Тип подгонки) выбрать *Off*, а в выпадающем меню *Y axis* (Ось Y) – *%*. Остальные настройки можно оставить без изменений.
- Нажмите на кнопку *OK*. В результате у вас должен получиться график, подобный приведенному на рис. 2.7.

РУКОПИСЬ

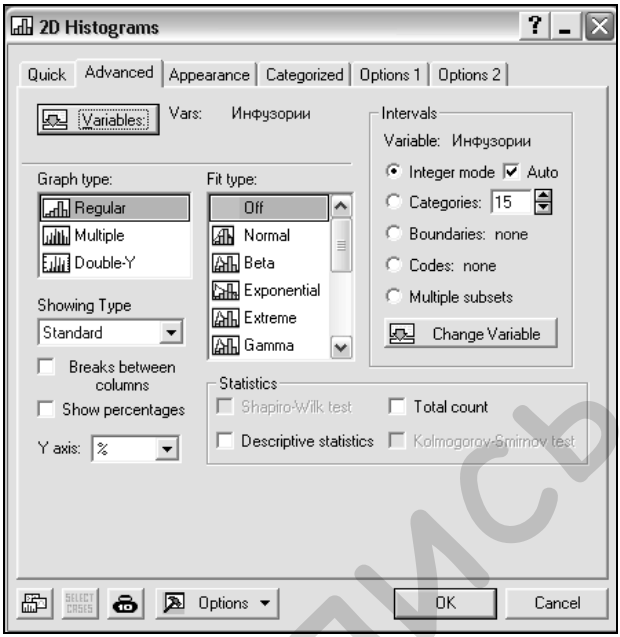


Рисунок 2.6. Диалоговое окно модуля *2D Histograms* на закладке *Advanced*.

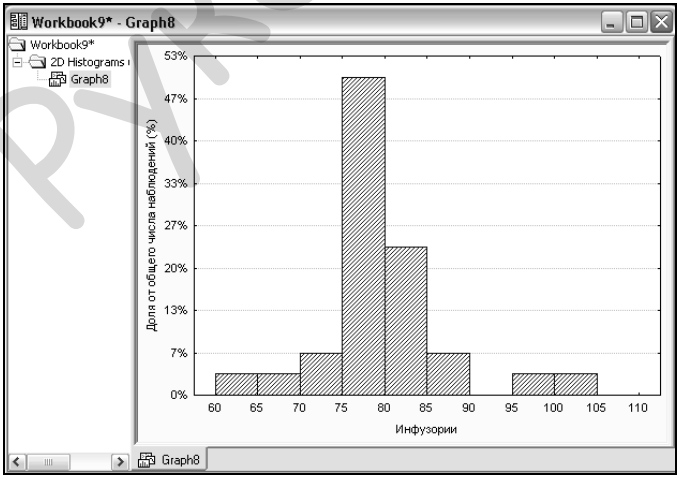


Рисунок 2.7. Частотное распределение значений длины клеток инфузории *C. acuminatus*.

РУКОПИСЬ

2.4. Расчет параметров описательной статистики

Расчет параметров описательной статистики в программе STATISTICA выполняется при помощи модуля *Descriptive statistics* (Описательная статистика). Для его запуска выполните одно из следующих действий (*Примечание*: любой анализ в программе STATISTICA можно запустить только если предварительно был открыт файл с данными, например, «Распределение.*sta*» из разд. 1.2):

- Войдите в раздел *Statistics* основного меню и выберите в нем пункт *Basic statistics/Tables* (см. разд. 2.2). В появившемся окне дважды кликните по пункту *Descriptive statistics* (рис. 2.8).
- В разделе *View* (Вид) основного меню выберите *Toolbars* (Инструменты) > *Statistics*. В верхней части рабочего окна появится дополнительная панель инструментов, содержащая кнопки быстрого запуска практически всех типов статистического анализа, реализованных в программе. Для запуска *Basic statistics* нажмите кнопку , после чего дважды кликните по пункту *Descriptive statistics* в открывшемся окне.

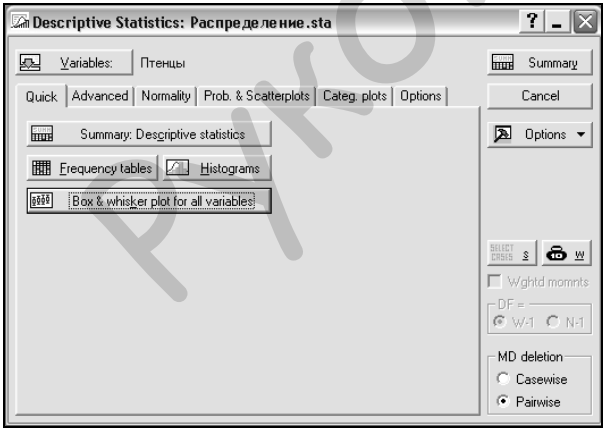


Рисунок 2.8. Модуль *Descriptive Statistics* на закладке *Quick.*

В диалоговом окне модуля *Descriptive statistics* (рис. 2.8) присутствует ряд элементов, встречающиеся в большинстве модулей программы, например:

Рукопись

- кнопка *Variables*, с помощью которой выбираются анализируемые переменные;
- кнопка *Summary* (Результат) – выводит результаты анализа;
- кнопка *Options* (Опции) – позволяет настроить внешний вид программы и окон вывода результатов анализа;
- стандартная для Windows кнопка *Cancel* (Отмена).

Кроме того, это окно имеет несколько закладок. По умолчанию перед пользователем первой предстает закладка *Quick* (Быстро). Находясь на ней, можно выполнить следующие операции:

- Рассчитать показатели описательной статистики – кнопка *Summary: Descriptive statistics*. Перечень рассчитываемых показателей определяется настройками, заданными на другой закладке окна – *Advanced*.
- Получить таблицу частот встречаемости каждого значения анализируемой переменной – кнопка *Frequency tables* (см. разд. 2.2);
- Построить частотное распределение значений анализируемой переменной в виде гистограммы – кнопка *Histograms*. Автоматически вместе с гистограммой программа нарисует теоретически ожидаемую нормальную кривую, глядя на которую можно заключить, подчиняются ли анализируемые данные нормальному закону распределения.
- Построить для выбранной переменной (или для нескольких переменных одновременно) т.н. диаграмму размаха (см. разд. 2.4) – кнопка *Box & whisker plot for all variables*.

Для расчета подробного перечня показателей описательной статистики следует воспользоваться другой закладкой модуля – *Advanced* (рис. 2.8). Основную часть этой закладки занимает список статистических показателей:

- *Valid N* – объем выборки;
- *Mean* – арифметическая средняя;
- *Sum* – сумма значений анализируемой переменной;
- *Median* – медиана;
- *Mode* – мода;
- *Geom. mean* – геометрическая средняя;
- *Harm. mean* – гармоническая средняя;
- *Standard Deviation* – стандартное отклонение;
- *Variance* – дисперсия;

РУКОПИСЬ

- *Std. err. of mean* – стандартная ошибка средней;
- *Conf. limits for means: Interval %* – доверительные пределы для средних: ширина доверительного интервала;
- *Skewness* – коэффициент асимметрии;
- *Std. err., Skewness* – стандартная ошибка коэффициента асимметрии;
- *Kurtosis* – коэффициент эксцесса;
- *Std. err., Kurtosis* – стандартная ошибка коэффициента эксцесса;
- *Minimum & maximum* – минимальное и максимальное значения;
- *Lower & upper quartiles* – нижний и верхний квартили;
- *Percentile boundaries: First & Second*: первый и второй процентили;
- *Range* – размах;
- *Quartile range* – интерквартильный размах.

На закладке *Advanced* имеются также следующие кнопки:

- *Select all stats* – позволяет выбрать для расчета сразу все имеющиеся статистические показателей;
- *Reset* – сброс «галочек» у всех показателей;
- *Save settings as default* – используя эту кнопку, можно сохранить определенный набор показателей, которые программа будет предлагать для расчета по умолчанию при каждом запуске модуля *Descriptive Statistics*.

Следующей за *Advanced* идет закладка *Normality* (Нормальность) (рис. 2.9). Это важная составляющая модуля описательной статистики, которой вам предстоит пользоваться очень часто. Здесь можно определить, насколько статистически значимо частотное распределение анализируемых данных отличается от нормального распределения. Наиболее важными элементами этой закладки являются:

- Уже известные вам кнопки *Frequency tables* и *Histograms*;
- Поле *Categorization* (Категоризация): воспользовавшись опцией *Number of intervals*, можно задать количество «столбиков» на гистограмме. Эта опция используется в случаях, когда анализируемый биологический признак является непрерывным. Если же он дискретен, т.е. выражается только целыми числами, следует выбрать опцию *Integral intervals* (Categories).

Рукопись

- Опция *Normal expected frequencies* (Ожидаемые нормальные частоты): при ее выборе и последующем нажатии на кнопку *Frequency tables* программа выдаст таблицу, которая помимо фактических частот численных значений переменной, будет содержать также теоретически ожидаемые нормальные частоты.
- Тесты, применяемые для проверки соответствия анализируемых данных закону нормального распределения – *Kolmogorov-Smirnov & Lilliefors test for normality* и *Shapiro-Wilk's W test*. Подробнее эти тесты будут рассмотрены в разд. 3.3.

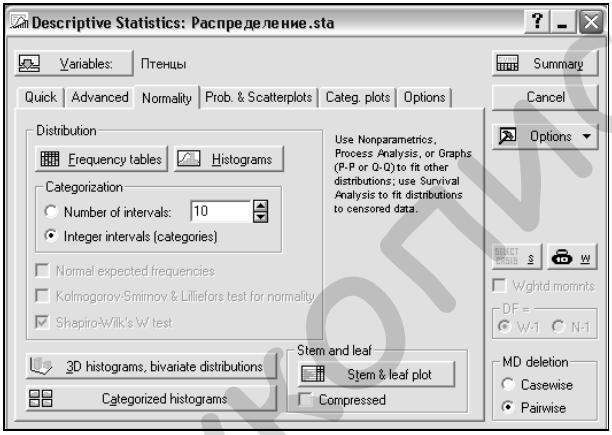


Рисунок 2.9. Модуль *Descriptive Statistics* на закладке *Normality.*

В ряде случаев полезной может оказаться и закладка *Prob. & Scatterplots* (Вероятностные графики и диаграммы рассеяния), следующая за закладкой *Normality*. В частности, с ее помощью можно построить двух- и трехмерные графики зависимости между анализируемыми переменными, а также проверить данные на нормальность с использованием графика нормальных вероятностей (*Normal probability plot*) (см. разд. 3.4).

Результаты практически любого статистического анализа воспринимаются гораздо легче, когда они представлены в графической форме. Как уже отмечалось ранее, графические возможности программы STATISTICA весьма обширны. Сейчас мы рассмотрим несколько типов графиков, которые наиболее часто используются в биологических исследованиях.

Рукопись

2.5. Диаграммы диапазонов

Диаграммы диапазонов (*wisker plots*) удобны для описания временной динамики или пространственного градиента исследуемых величин. Точки на таких графиках чаще всего соответствуют средней арифметической или, реже, медиане анализируемого признака. Отличительной особенностью является наличие у точек т.н. «усов» (от англ. «*whiskers*») – вертикально или горизонтально отходящих линий, длина которых соответствует величине выбранного исследователем показателя разброса данных (минимум и максимум, стандартное отклонение, дисперсия, квартили) или точности оценки генеральных параметров (стандартная ошибка, доверительный интервал).

Data: Температу...

	1 Сезон	2 Температура
1	Май	12.9
2	Май	13.0
3	Май	13.0
4	Июнь	19.0
5	Июнь	19.5
6	Июнь	20.4
7	Июль	25.6
8	Июль	24.0
9	Июль	26.0
10	Август	20.1
11	Август	19.0
12	Август	20.1
13	Сентябрь	15.0
14	Сентябрь	14.8
15	Сентябрь	14.0

Рисунок 2.10. Пример оформления данных для построения диаграммы диапазонов

Рассмотрим следующий пример. В течение 5 месяцев – с мая по сентябрь – в озере на постоянной станции выполняли измерение температуры воды на глубине 1 м. Измерения проводились в полдень три раза в месяц. Полученные данные приведены на рис. 2.10 (создайте аналогичный файл данных и сохраните его – он потребуется нам также при рассмотрении следующего раздела). Изобразим графически динамику среднемесячной температуры в водоеме.

Обратите внимание на то, каким образом данные располагаются в таблице (рис. 2.10). Чтобы программа «поняла», какие из наблюдений относятся к конкретному месяцу, был введен дополнительный столбец «Сезон». В нем перечислены названия месяцев, а в соседнем столбце – «Температура» – приведены сами значения исследуемой переменной. Такой способ оформления данных характерен для многих видов статистического анализа, реализованных в STATISTICA, и будет неоднократно встречаться нам в дальнейшем. Столбец «Сезон» в

Рукопись

терминах программы называется «группирующей переменной» (*Grouping variable*), а столбец, в котором непосредственно находятся значения исследуемого признака – зависимой переменной (*Dependent variable*). Последнее название является очень подходящим, т.к. указывает на то, что, например, температура воды *зависит* от сезона года.

Для построения диаграммы диапазонов необходимо в разделе *Graphs* основного меню выбрать 2*D Graphs*, а затем – *Means w/Error plots* (Графики средних с ошибками). Вешний вид появляющегося в результате этого окна представлен на рис. 2.11.

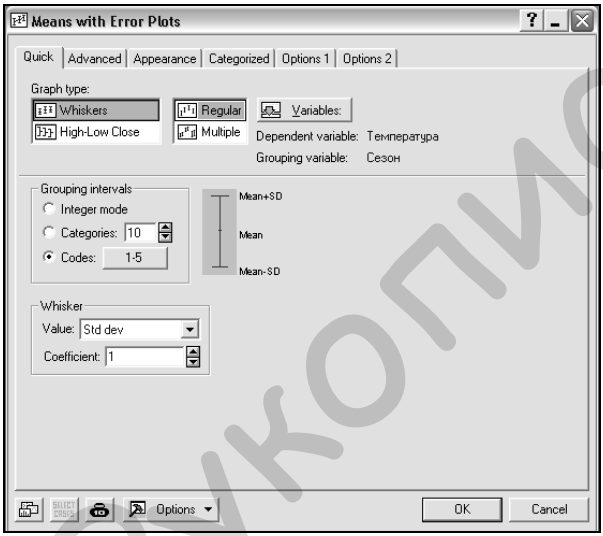


Рисунок 2.11. Окно настройки параметров диаграммы диапазонов.

Рукопись

Как обычно, начинать следует с указания переменных, которые будут участвовать в анализе – для этого необходимо воспользоваться кнопкой *Variables*. Появится диалоговое окно (*Select variables for means with error plots*) с двумя списками имеющихся в таблице переменных. В левом списке необходимо выбрать зависимую переменную (в нашем случае это «Температура»), а в правом – группирующую переменную («Сезон»). После этого – нажать кнопку *OK*.

Далее в поле *Grouping intervals* (Группирующие интервалы) нужно указать программе, на какие интервалы ей следует разбить ось X. В нашем примере вдоль оси X должны располагаться названия месяцев. Чтобы сообщить это

программе, нажимаем кнопку *Codes* (Коды), а на появляющейся панели – кнопки *All* (Все) и *OK* (*Пояснение*: в качестве кодов в нашем примере выступают названия месяцев. Поскольку мы хотим, чтобы на графике были отображены данные для всех месяцев, в течение которых выполнялись измерения температуры, необходимо нажать кнопку *All*).

Осталось указать, чему на графике будут соответствовать «усы», отходящие от точек. Для этого служит поле *Whisker*. Предположим, мы хотим, чтобы длина «усов» была бы равна одному стандартному отклонению. В выпадающем меню *Value* (Значение) выбираем *Std dev* (Стандартное отклонение), а в поле *Coefficient* (Коэффициент) ставим 1 (рис. 2.11). Теперь все основные настройки завершены. После нажатия на кнопку OK можно будет увидеть график, подобный приведенному на рис. 2.12.

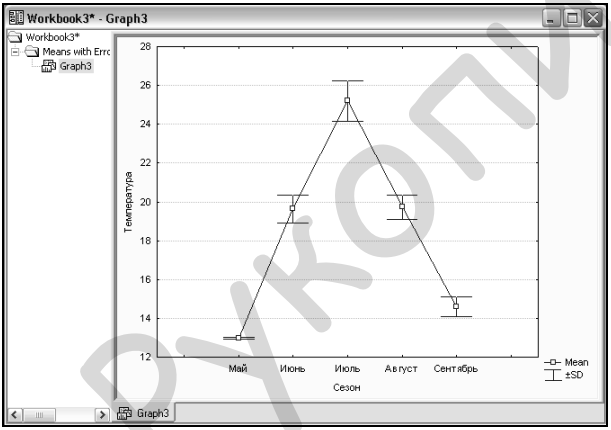


Рисунок 2.12. Диаграмма диапазонов, построенная по данным о температуре воды в озере.

Рукопись

2.4. Диаграммы размахов

Диаграммы размаха, или «ящики с усами» (от англ. «*box-whisker plots*»), получили свое название за характерный вид: точку, соответствующую средней арифметической или медиане, окружает вертикально расположенный прямоугольник («ящик»), длина которого соответствует одному из показателей разброса или точности оценки генерального параметра. Дополнительно от этого прямоугольника отходят «усы», также соответствующие

по длине одному из показателей разброса или точности. Таким образом, графики этого типа позволяют дать очень полную статистическую характеристику для каждой анализируемой выборки. Диаграммы размаха можно использовать для визуальной экспресс-оценки разницы между двумя или более группами (например, между датами отбора проб, экспериментальными группами, участками пространства и т.п.).

Для построения диаграммы диапазонов необходимо в разделе *Graphs* основного меню выбрать 2*D Graphs*, а затем – *Box plots*. На рис. 2.13 представлен внешний вид данного модуля, открытый на закладке *Advanced*.

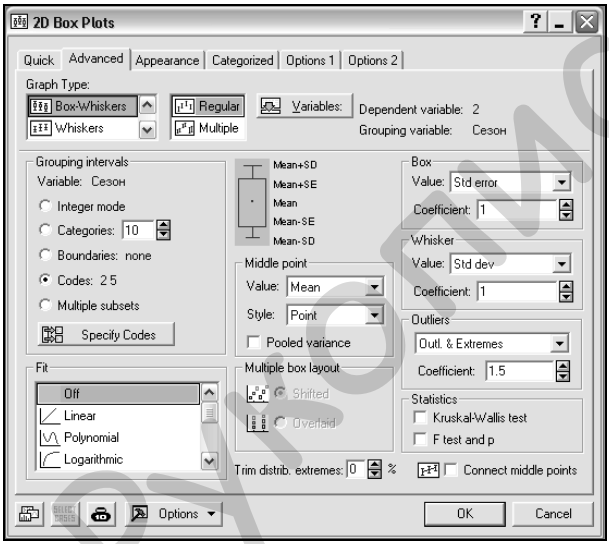


Рисунок 2.13. Окно настройки параметров диаграммы размахов.

Вернемся к примеру, рассмотренному ранее в разделе 2.2 (откройте сохраненный файл с данными по температуре воды). Предположим, что мы хотим визуально сравнить, различается ли среднемесячная температура воды в озере в июне и сентябре. Для построения графика необходимо установить следующие настройки (см. рис. 2.13):

- На закладке *Advanced* нажать на кнопку *Variables* и указать, какая из переменных является зависимой (*Dependent*) («Температура»), а какая – группирующей (*Grouping*) («Сезон»).

РУКОПИСЬ

- В поле *Grouping intervals* выбрать опцию *Codes*, а затем нажать кнопку *Specify codes* (Определить коды), чтобы указать программе, какие именно месяцы будут участвовать в анализе. В появившемся окне ввести через пробел слова «Июнь» и «Сентябрь».
- В выпадающем меню *Value* поля *Middle point* (Средняя точка) выбрать *Mean* (Арифметическая средняя). Так мы сообщим программе, что на графике в качестве точек ей следует изображать средние значения температуры.
- В выпадающем меню *Value* поля *Box* выбрать, статистический показатель, который будет изображен в виде «ящика» (например, *Std error* – Стандартная ошибка). *Coefficient* выставить на 1.
- В выпадающем меню *Value* поля *Whisker* выбрать статистический показатель, который будет изображен в виде «усов» (например, *Std dev* – Стандартное отклонение).
- В поле *Outliers* (Выбросы) выбрать *Off* (Отключить). (*Пояснение*: в результате этого действия программа не будет изображать на графике точки-выбросы, т.е. значения признака, которые слишком велики или слишком малы по сравнению с остальными значениями в выборке.)

Остальные настройки можно оставить без изменений. После нажатия на кнопку *OK* появится график, подобный приведенному на рис. 2.14. На полученном графике хорошо видно, что среднемесячная температура воды в июне была значительно выше, чем в сентябре.

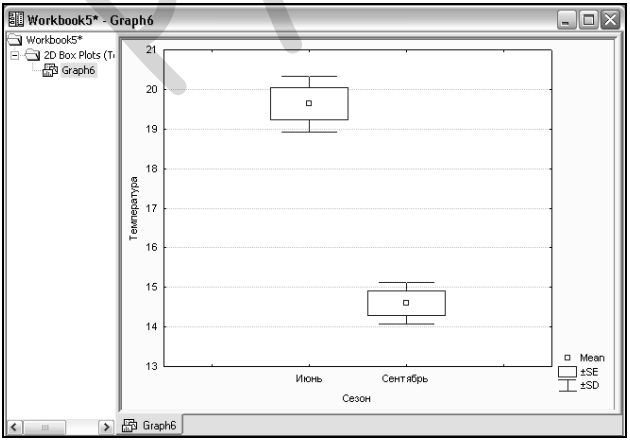
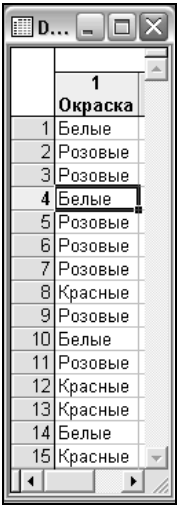


Рисунок 2.14. Диаграмма размахов, построенная по данным о температуре воды в озере за июнь и сентябрь.

РУКОПИСЬ

2.5. Круговые диаграммы

Круговые диаграммы (*pie charts*) удобны при анализе качественных признаков. Например, такой график хорошо подойдет для описания соотношения растений с разной окраской цветков в изучаемой популяции. Допустим, при обследовании 15 растений получены результаты, приведенные на рис. 2.15.



	1 Окраска
1	Белые
2	Розовые
3	Розовые
4	Белые
5	Розовые
6	Розовые
7	Розовые
8	Красные
9	Розовые
10	Белые
11	Розовые
12	Красные
13	Красные
14	Белые
15	Красные

Рисунок 2.15. Пример оформления данных для построения круговой диаграммы

Для построения круговой диаграммы, секторы которой были бы пропорциональны долям каждого из вариантов окраски цветков, необходимо выполнить следующее:

- В разделе *Graphs* главного меню выбрать *2D Graphs* > *Pie chart*.
- В появившемся окне перейти на закладку *Advanced* (рис. 2.16).
- В поле *Graph type* (Тип графика) выбрать *Pie chart – Counts* (Кругова диаграмма – Счет). Данная опция позволяет построить график на основе исходных данных – программа сама подсчитает, сколько в анализируемой совокупности было растений с белыми, розовыми и красными цветками. Если бы мы ввели в таблицу предварительно рассчитанные доли каждой из окрасок, то в поле *Graph type* следовало бы выбрать *Pie chart – Values* (Кругова диаграмма – Значения). Однако при этом пришлось бы оформить таблицу с данными несколько по-иному (здесь не рассматривается).
- В поле *Frequency intervals* (Интервалы частот) выбрать опцию *Codes* и нажать кнопку *Specify codes*. На появившейся панели нажать кнопку *All*, а затем *OK*.
- В поле *Pie legend* (Легенда диаграммы) выбрать вариант того, как будут подписаны сегменты круговой диаграммы. Например, при выделении *Text and Percent* будут отображены названия вариантов окраски цветков и частота (%), с которой встречается каждый вариант в популяции.

РУКОПИСЬ

Остальные настройки можно оставить неизменными. Нажатие на кнопку *OK* приведет к построению графика, подобного приведенному на рис. 2.17.

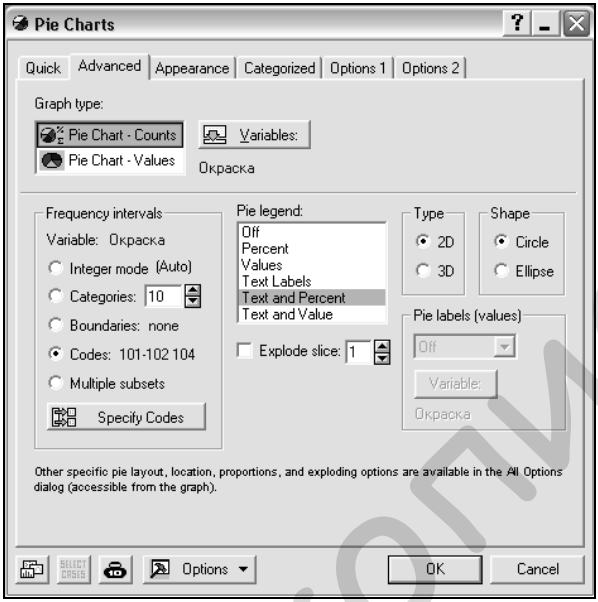


Рисунок 2.16. Окно настройки параметров круговой диаграммы.

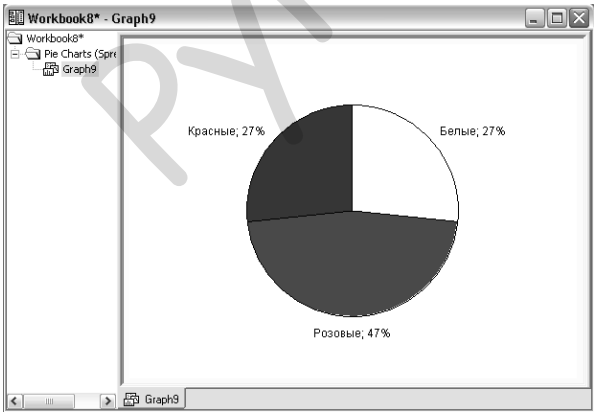


Рисунок 2.17. Круговая диаграмма, построенная по данным об окраске цветков.

Рукопись

Глава 3. Проверка соответствия анализируемых данных закону нормального распределения

3.1. О необходимости проверки нормальности распределения анализируемых данных

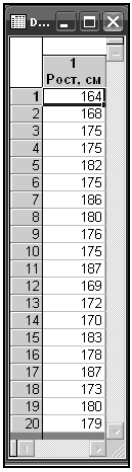
Как известно, существующие методы статистического анализа можно подразделить на две группы – параметрические и непараметрические. Важным условием, определяющим возможность применения параметрических методов, является подчинение анализируемых данных закону нормального (Гауссова) распределения, которое имеет характерный колоколообразный вид. В то же время непараметрические методы выполнения этого условия не требуют. Установлено, что в подавляющем большинстве случаев (около 75%; Леонов 2007) распределения биологических признаков существенно отличаются от нормального. Тем не менее, очень многие исследователи-биологи совершают ошибку, применяя параметрические методы анализа для ненормально распределенных признаков. Часто это приводит к выводам, не соответствующим действительности (Гланц 1999; Леонов 2007). Во избежание указанной ошибки, любой анализ биологических признаков должен сопровождаться проверкой нормальности их распределения. Для этого существует достаточно широкий набор методов. Мы рассмотрим три подхода, реализованные в программе STATISTICA.

3.2. Подгонка распределения

На рис. 3.1 представлены результаты измерения роста у 20 студентов мужского пола. Необходимо установить, распределены ли эти данные по нормальному закону.

В программе STATISTICA имеется специальный модуль – *Distribution fitting* (Подгонка распределения), позволяющий проверить соответствие анализируемых данных целому ряду математических распределений. Его можно запустить из раздела главного меню *Statistics*, или нажав кнопку на дополнительной панели инструментов (см. разд. 2.4). Внешний вид окна этого модуля приведен на рис. 3.2.

Рукопись



	1 Рост, см
1	164
2	168
3	175
4	175
5	182
6	175
7	186
8	180
9	176
10	175
11	187
12	169
13	172
14	170
15	183
16	178
17	187
18	173
19	180
20	179

Рисунок 3.1. Данные о росте 20 студентов (см). Объяснения в тексте.

Поскольку нам необходимо проверить, подчиняются ли данные о росте студентов нормальному распределению, в списке непрерывных распределений (*Continuous distributions*) выбираем *Normal*, после чего нажимаем кнопку *OK* (рис. 3.2). Далее появится еще одно диалоговое окно (рисунок не приводится), где необходимо указать, какую именно переменную мы хотим проанализировать, и как. Переменная для анализа задается путем нажатия кнопки *Variables*.

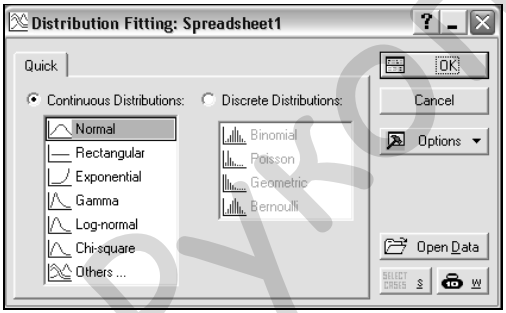


Рисунок 3.2. Диалоговое окно модуля *Distribution fitting.*

Остальные настройки можно оставить без изменений. Нажав кнопку *Plot of observed and expected distributions* (График наблюдаемого и ожидаемого распределений), получим гистограмму распределения данных о росте студентов и колоколообразную красную кривую, соответствующую ожидаемому нормальному распределению (у этого ожидаемого распределения те же средняя арифметическая и стандартное отклонение, что и в анализируемой совокупности данных) (рис. 3.3). Глядя на полученный рисунок, можно сказать, что в целом распределение значений роста студентов соответствует нормальному (столбики гистограммы образуют

РУКОПИСЬ

колоколообразную фигуру). Это заключение, основанное на визуальном анализе распределения, имеет и более строгое подтверждение в виде результатов теста χ^2 (*Chi-square test*, см. в верхней части графика на рис. 3.3). В данном случае этот тест проверяет нулевую гипотезу о том, что наблюдаемое распределение анализируемого признака не отличается от теоретически ожидаемого нормального распределения. Поскольку вероятность ошибиться, отклонив эту гипотезу оказалась намного больше 0,05 (P = 0,49448), мы принимаем, что гипотеза действительно верна. Иными словами, распределение значений роста студентов статистически *не* отличается от нормального распределения.

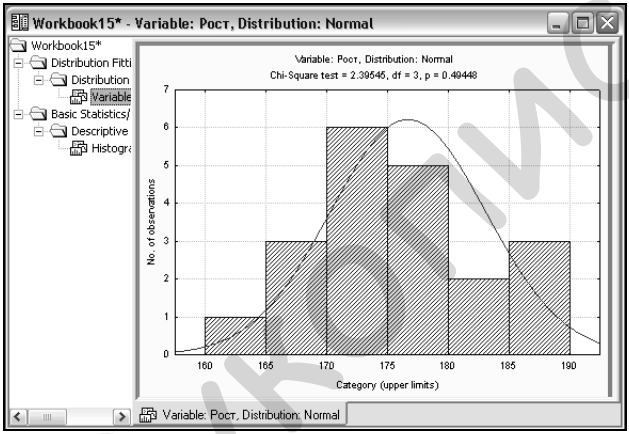


Рисунок 3.3. Результат анализа, выполненного с помощью модуля *Distribution fitting*.

3.3. Тесты Колмогорова-Смирнова и Шапиро-Уилка

Следует отметить, что мощность теста хи-квадрат при проверке нормальности распределения анализируемых данных относительно невысока (другими словами, его применение достаточно часто приводит к ошибочному выводу о нормальности распределения). Поэтому лучше воспользоваться другими тестами. Их можно найти в уже рассмотренном выше модуле *Descriptive Statistics* (*Описательная статистика*) (разд. 2.4). После запуска этого модуля необходимо открыть закладку *Normality* и в поле *Distribution* (Распределение) разыскать опции *Kolmogorov-Smirnov and Lilliefors test for normality*

РУКОПИСЬ

(Тест Колмогорова-Смирнова и Лиллифорса на нормальность) и *Shapiro-Wilk's W test* (W-тест Шапиро-Уилка) (рис. 2.9). Равно как и критерий χ^2, эти тесты проверяют нулевую гипотезу об отсутствии различий между наблюдаемым распределением признака и теоретическим ожидаемым нормальным распределением. Наиболее предпочтительным, особенно при небольших выборках (n = 3 ÷ 50) является использование W-критерия Шапиро-Уилка, поскольку он обладает наибольшей мощностью в сравнении со всеми перечисленными критериями (т.е. чаще выявляет различия между распределениями в тех случаях, когда они действительно есть). Для выбора того или иного теста, достаточно поставить флажок рядом с его названием. После выбора анализируемой переменной (кнопка *Variables*) и нажатия кнопки *Histograms* программа создаст гистограмму распределения значений признака и ожидаемую нормальную кривую (рис. 3.4). Результаты выбранных тестов на нормальность автоматически располагаются в заголовке этого графика. При P > 0,05 можно заключить, что анализируемое распределение *не отличается* от нормального. В примере с данными о росте студентов для теста Шапиро-Уилка получаем P = 0,7979 (рис. 3.4), что подтверждает сделанный ранее вывод о нормальности распределения этих данных.

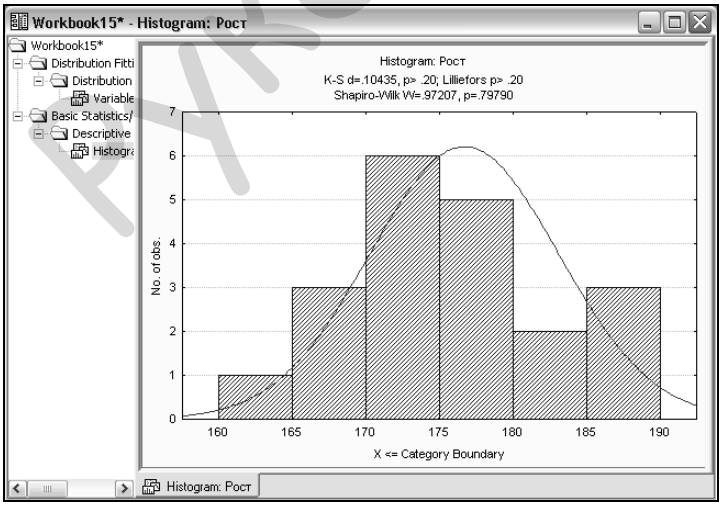


Рисунок 3.4. Результат проверки нормальности распределения данных, выполненной при помощи модуля *Descriptive Statistics*.

3.4. График нормальных вероятностей

В модуле *Descriptive Statistics* реализован еще один способ проверки данных на нормальность распределения. Он заключается в использовании графика нормальных вероятностей, или т.н. «вероятностной бумаги». Такой график изображает зависимость ожидаемых нормальных частот значений признака от их реальных частот. Очевидно, что если между наблюдаемым и ожидаемым распределениями нет никакой разницы, точки на этом графике выстроятся строго вдоль прямой. Иначе они образуют фигуру отличную от прямой. На этом принципе и основано применение «вероятностной бумаги». Для построения графика такого типа необходимо в модуле *Descriptive Statistics* перейти на закладку *Prob. & Scatterplots* (Вероятностные графики и диаграммы рассеяния) и нажать на кнопку *Normal probability plot* (График нормальных вероятностей). В результате появится график, подобный приведенному на рис. 3.5. Точки на этом рисунке плотно выстраиваются вдоль теоретически ожидаемой прямой, что еще раз подтверждает нормальность распределения данных о росте студентов.

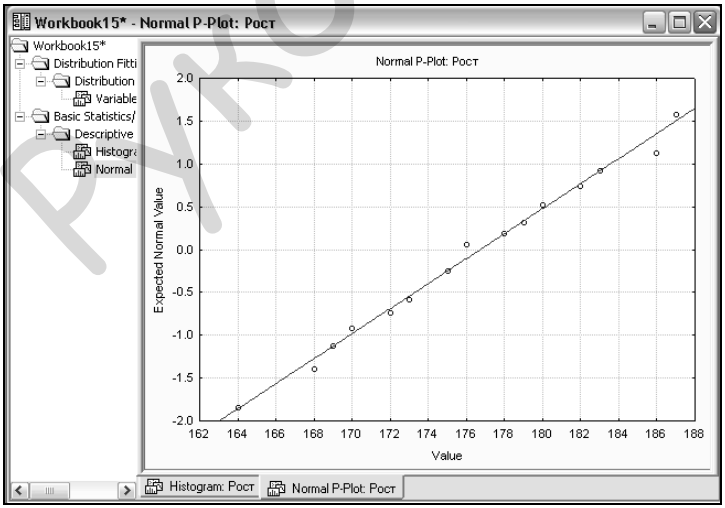


Рисунок 3.5. Проверка нормальности распределения данных о росте студентов с использованием графика нормальных вероятностей.

РУКОПИСЬ

Глава 4. Сравнение двух групп

4.1. Случай независимых выборок

Одной из обычных задач в биологических исследованиях является сравнение арифметических средних двух групп (например, экспериментальной и контрольной). Классическим методом, позволяющим ее решать, является *t-тест Стьюдента*, или просто «*t-тест*». Нулевая гипотеза, проверяемая в ходе данного теста, заключается в том, что *обе группы происходят из одной генеральной совокупности*; другими словами, что наблюдаемые различия между средними значениями сравниваемых выборок *случайны* и *не* вызваны действием изучаемого фактора. Тест Стьюдента относится к группе параметрических методов анализа. Его корректное применение требует выполнения трех условий:

- обе выборки должны быть *независимыми*, т.е. свойства одной из них никак не должны быть связаны со свойствами другой (известно, например, что женщины в среднем ниже мужчин, однако это не является результатом того, что мужчины оказывают какое-то особое влияние на рост женщин – дело здесь в генетических особенностях пола);
- обе выборки должны подчиняться *нормальному закону распределения*;
- между дисперсиями выборок не должно быть статистически значимой разницы (*однородность дисперсий*).

К сожалению, многие исследователи-биологи игнорируют перечисленные условия при выполнении теста Стьюдента, что часто приводит к ошибочным результатам (Гланц 1999; Леонов 2007). Наиболее «опасным» является несоблюдение требования о нормальности распределения значений признака в сравниваемых группах. Способы проверки нормальности распределения описаны в предыдущей главе.

Рассмотрим, как тест Стьюдента можно выполнить при помощи программы STATISTICA. Считается, что животные, обитающие в северных широтах, обладают более короткими придатками тела, нежели животные из южных широт. На рис. 3.6 приведены данные о длине крыльев (мм) птиц одного вида, пола и возраста, обитающих в разных широтах. Применим *t*-тест для сравнения средних значений этих двух независимых

Рукопись

выборок. (*Примечание*: в учебных целях допустим, что обе выборки распределены нормально, и что их дисперсии различаются незначительно).

Data: Birds...

	1 Широта	2 Длина
1	Северная	120
2	Северная	113
3	Северная	125
4	Северная	118
5	Северная	116
6	Северная	114
7	Северная	119
8	Южная	116
9	Южная	117
10	Южная	121
11	Южная	114
12	Южная	116
13	Южная	118
14	Южная	123
15	Южная	120

Рисунок 4.1. Пример оформления данных для выполнения *t*-теста для независимых выборок.

Обратите внимание на то, как оформлены данные на рис. 4.1. Как и при построении графиков типа *Whisker plot* или *Box-whisker plot* (разд. 2.3-2.4), в таблице имеются две переменные. Одна из них – *группирующая* (*Grouping variable*) («Широта») – содержит коды, указывающие принадлежность данных о длине крыльев к конкретной группе. Другая – т.н. *зависимая переменная* (*Dependent variable*) («Длина») – содержит собственно данные. Возможен и другой вариант оформления – данные для каждой группы («Северная» и «Южная») можно было бы просто внести в отдельные столбцы, не используя группирующую переменную.

Для выполнения *t*-теста для независимых выборок необходимо выполнить следующие действия:

- Запустить соответствующий модуль (рис. 4.2) из меню *Statistics* > *Basic statistics/Tables* > *t-test, independent, by groups* (если в таблице с данными есть группирующая переменная) или *t-test, independent, by variables* (если данные внесены в самостоятельные столбцы). (*Примечание*: мы рассмотрим вариант теста, при котором группирующая переменная присутствует; рис. 4.1). Вместо использования меню *Statistics* можно нажать кнопку на дополнительной панели инструментов.
- В открывшемся окне нажать кнопку *Variables* и указать программе, какая из переменных является группирующей, а какая – зависимой (рис. 4.3).
- Нажать на кнопку *Summary: T-tests* (рис. 4.2).

Рукопись

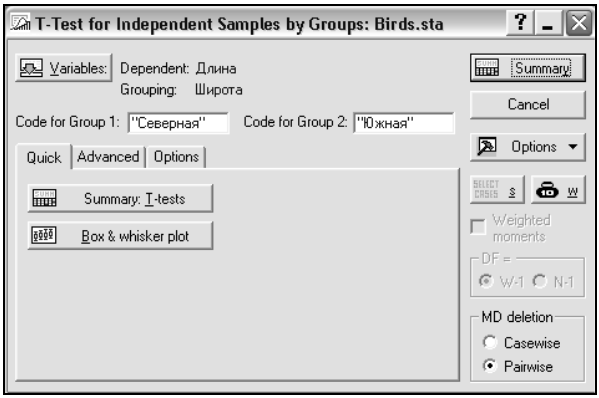


Рисунок 4.2. Модуль *t*-теста для независимых выборок.

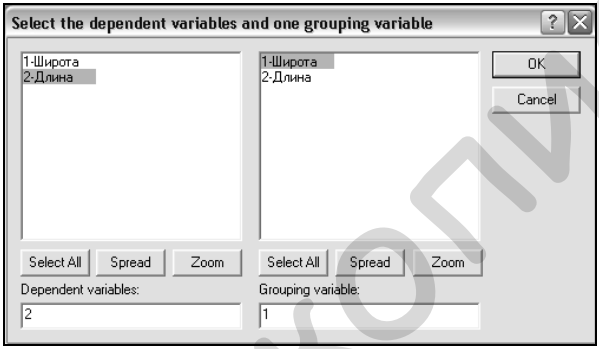


Рисунок 4.3. Выбор переменных для включения в *t*-тест.

РУКОПИСЬ

В итоге программа создаст рабочую книгу, содержащую таблицу с результатами *t*-теста. Эта таблица имеет несколько столбцов (рис. 4.4):

- *Mean* (Северная): среднее значение длины крыльев у птиц в группе «Северная»;
- *Mean* (Южная): среднее значение длины крыльев у птиц в группе «Южная»;
- *t-value*: значение рассчитанного программой *t*-критерия Стьюдента;
- *df:* число степеней свободы;
- *P*: вероятность ошибочно отвергнуть нулевую гипотезу об отсутствии различий между средними (см. выше). Фактически, это самый главный интересующий нас результат анализа. В рассматриваемом примере $P > 0{,}05$, на основании

чего можно сделать вывод об отсутствии статистически значимых различий между средними значениями длины крыльев птиц из разных широт.

- *Valid N* (Северная): объем выборки «Северная»;
- *Valid N* (Южная): объем выборки «Южная»;
- *Std. dev.* (Северная): стандартное отклонение выборки «Северная»;
- *Std. dev.* (Южная): стандартное отклонение выборки «Южная»;
- *F-ratio, Variances*: значение *F*-критерия Фишера, с помощью которого проверяется гипотеза о равенстве дисперсий в сравниваемых выборках (см. выше условия применения теста Стьюдента);
- *P, Variances*: вероятность ошибки для *F*-теста Фишера. Поскольку в нашем случае $P > 0{,}05$, можно заключить, что дисперсии сравниваемых выборок не различаются (т.е. условие однородности дисперсий выполняется).

T-tests; Grouping: Широта (Spreadsheet1)
Group 1: Северная
Group 2: Южная

Variable	**Mean Северная**	Mean Южная	t-value	df	p	Valid N Северная	Valid N Южная	Std.Dev. Северная	Std.Dev. Южная	F-ratio Variances	p Variances
Длина	117,8571	118,1250	-0,146732	13	0,885595	7	8	4,059087	2,997022	1,834327	0,445585

Рисунок 4.4. Результаты выполнения *t*-теста для независимых выборок.

Если значения признака в двух сравниваемых группах распределены ненормально, применение параметрического *t*-теста для их сравнения будет часто приводить к искаженным результатам. В таких случаях следует воспользоваться соответствующим непараметрическим аналогом теста Стьюдента. Для сравнения двух независимых ненормально распределенных выборок используется *U-тест Манна-Уитни* (*Mann-Whitney U-test*). В программе STATISTICA этот тест выполняется следующим образом:

- В меню *Statistics* выбрать *Nonparametrics*, а затем *Comparing two independent samples* (Сравнение двух независимых выборок). Альтернативный вариант запуска – нажатие кнопки на дополнительной панели инструментов.

РУКОПИСЬ

- В появившемся окне (рис. 4.5) нажать на кнопку *Variables* и выбрать зависимую и независимую переменные (для примера воспользуемся теми же данными по длине конечностей у птиц из разных широт, см. рис. 4.1).

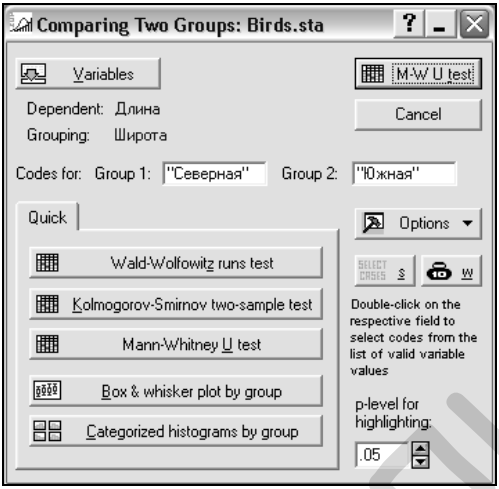


Рисунок 4.5. Окно *Comparing two groups* модуля *Nonparametrics* (при сравнении независимых выборок).

- Нажать на кнопку *Mann-Whitney U-test* или *M-W U test*. Внешний вид появляющегося после этого окна представлен на рис. 4.6. Самое главное, на что следует обратить внимание в итоговой таблице теста – это величина вероятности ошибки *P*. При большом числе наблюдений в выборках (20 и более) значение *P* необходимо искать в 5-м столбце таблицы (вслед за «Z»), иначе – в 7-м (вслед за «*Z-adjusted*»). При $P < 0{,}05$ делается вывод о наличии статистически значимой разницы между сравниваемыми выборками (*Примечание*: в отличие от *t*-теста, тест Манна-Уитни сравнивает не средние значения выборок, а суммы рангов по каждой из них. Ранг – положение определенного значения изучаемого признака в упорядоченном по убыванию или возрастанию ряду).

Mann-Whitney U Test (Spreadsheet3)
By variable Широта
Marked tests are significant at p <,05000

variable	**Rank Sum Северная**	Rank Sum Южная	U	Z	p-level	Z adjusted	p-level	Valid N Северная	Valid N Южная	2*1sided exact p
Длина	53,50000	66,50000	25,50000	-0,289319	0,772338	-0,291144	0,770941	7	8	0,778866

Рисунок 4.6. Результаты выполнения теста Манна-Уитни.

Рукопись

4.2. Случай зависимых выборок

С зависимыми выборками исследователь имеет дело каждый раз, когда измерения значений изучаемого признака выполняются на одних и тех же объектах. Рассмотрим следующий пример. Для выяснения эффективности нового удобрения последнее внесли в одинаковом количестве на 9 одинаковых по площади участках и в конце вегетационного сезона измерили урожайность некоторой культуры. На следующий год по совершенно аналогичной схеме (на *тех же* участках) выполнили еще один эксперимент, однако со старым удобрением (рис. 4.7). Необходимо выяснить, различается ли средняя урожайность культуры в зависимости от используемого удобрения.

Data: Sp...

	1 Новое	2 Старое
1	2250	1920
2	2410	2020
3	2260	2060
4	2200	1960
5	2360	1960
6	2320	2140
7	2240	1980
8	2300	1940
9	2090	1790

Рисунок 4.7. Пример оформления данных для выполнения *t*-теста для зависимых выборок.

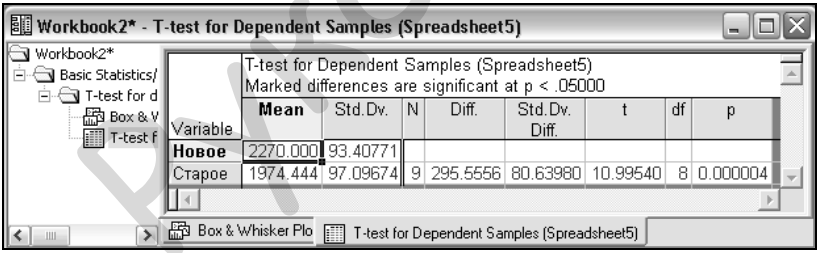
Поскольку удобрения вносились на одни и те же участки, то выборки, полученные в результате двух описанных экспериментов, являются зависимыми. Это объясняется тем, что урожайность культуры во второй год исследований вполне могла испытывать определенное последействие нового удобрения, т.е. она *зависела* от того, что происходило с опытными участками ранее. Чтобы сравнить средние урожайности воспользуемся *t*-тестом для зависимых выборок. (*Примечание*: в учебных целях допустим, что условие о нормальности распределения данных выполняется). Для выполнения этого варианта *t*-теста необходимо:

- Запустить соответствующий модуль из меню *Statistics* > *Basic statistics/Tables* > *t-test, dependent samples*. Вместо использования меню *Statistics* можно нажать кнопку на дополнительной панели инструментов.
- В открывшемся окне нажать на кнопку *Variables* и указать программе первую (*First variable*) и вторую (*Second variable*) переменные, участвующие в анализе.
- Нажать на кнопку *Summary: T-tests*.

РУКОПИСЬ

В результате появится таблица с результатами, очень похожая на ту, что мы уже видели при выполнении *t*-теста для независимых выборок. Она содержит следующие столбцы (рис. 4.8):

- *Mean* – средние значения урожайности для каждой из сравниваемых групп;
- *Std. dv.* – стандартные отклонения для каждой группы;
- *N* – число наблюдений;
- *Diff.* – средняя разница урожайности (о том, как «работает» *t*-тест для зависимых выборок, см., например, Гланц 1999);
- *Std. dv. diff.* – стандартное отклоение для средней разницы;
- *t* – значение *t*-критерия;
- *df* – число степеней свободы;
- *P* – вероятность ошибочно отвергнуть нулевую гипотезу о том, что средние величины урожайности в сравниваемых группах не различаются. Поскольку в нашем случае $P << 0{,}05$, можно смело заключить, что средние значения урожайности при использовании нового и старого удобрений значительно различаются (*Примечание:* при наличии различий, результаты анализа в STATISTICA обычно (но не во всех модулях!) выделяются красным цветом).



Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
Новое	2270.000	93.40771						
Старое	1974.444	97.09674	9	295.5556	80.63980	10.99540	8	0.000004

Рисунок 4.8. Результаты выполнения *t*-теста для зависимых выборок.

Если две зависимые выборки распределены ненормально, то для их сравнения следует применить *тест Уилкоксона* (*Wilcoxon matched pair test*), который можно найти там же, где и тест Манна-Уитни (*Statistics* > *Nonparametrics* > *Comparing dependent samples*, или нажать кнопку на дополнительной панели инструментов). Далее необходимо:

РУКОПИСЬ

- В появившемся окне (рис. 4.9) нажать кнопку *Variables* и задать переменные для анализа (для примера используем данные, представленные на рис. 4.7).
- Нажать кнопку *Wilcoxon matched pair test*.
- В итоговой таблице (рис. 4.10) найти величину *P*. При $P < 0{,}05$ можно сделать вывод о наличии статистически значимой разницы между сравниваемыми выборками.

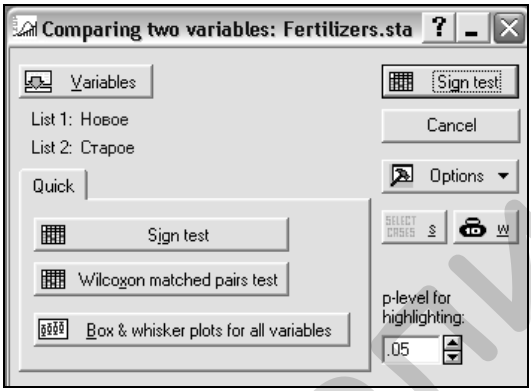
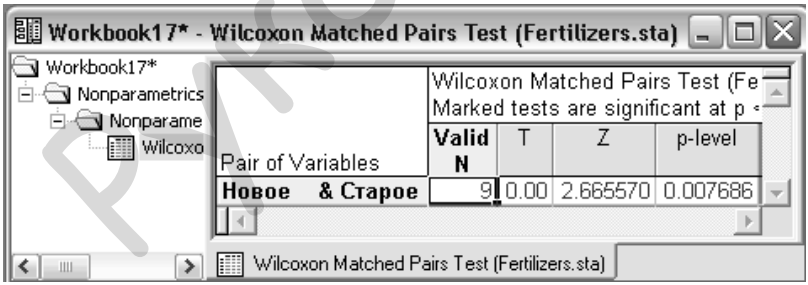


Рисунок 4.9. Окно *Comparing two groups* модуля *Nonparametrics* (при сравнении зависимых выборок)



4.10. Результаты выполнения теста Уилкоксона.

4.3. Сравнение выборочной средней с константой

В ряде случаев возникает необходимость сравнить выборочную среднюю не с другой выборочной средней, а с определенной константой. Допустим, что, согласно

государственному стандарту, ПДК некоторого загрязнителя составляет 100 единиц. При замерах содержания этого вещества в 10 пробах городской почвы получены следующие значения: 108, 99, 112, 100, 101, 98, 95, 105, 90, 102. Необходимо установить, превышает ли среднее содержание загрязнителя в исследованных образцах почвы предельно допустимую концентрацию?

Для ответа на поставленный вопрос необходимо воспользоваться анализом, который в программе STATISTICA называется *t-test for single means* (*t*-тест для средних, рассчитанных по одной выборке). Его можно найти в меню *Statistics* > *Basic Statistics/Tables* > *t-test, single sample*, или нажав кнопку на дополнительной панели инструментов. В результате появится окно, внешний вид которого представлен на рис. 4.11.

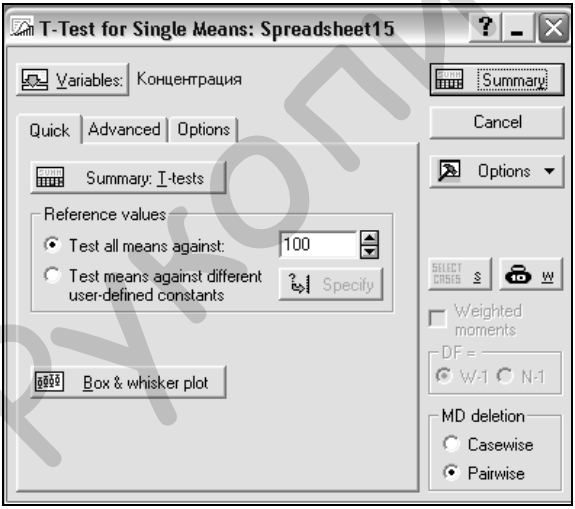


Рисунок 4.11. Окно настройки *t*-теста для средних, рассчитанных по одной выборке

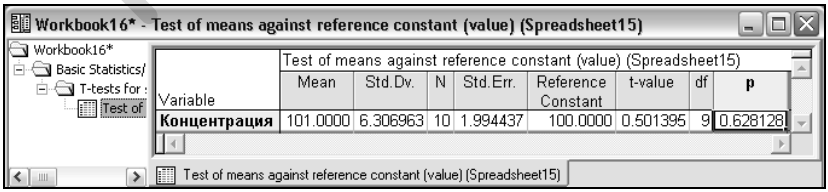
Нажав на кнопку *Variables*, необходимо выбрать столбец, который содержит анализируемые данные. Таких переменных можно ввести несколько (например, если бы аналогичные отборы проб почвы производились в разные месяцы, то в анализ можно было бы включить и эти данные). В таком случае

Рукопись

программа сравнит все выборки с одним «контрольным» значением. Последнее задается в поле *Reference values* (Контрольные значения) (рис. 4.11). В нашем примере контролем является ПДК = 100 – его и следует внести напротив опции *Test all means against...* (Сравнить все средние с...). При необходимости, включенные в анализ переменные можно сравнить с несколькими контрольными значениями (это достигается путем активации опции *Test means against user-defined constants* (Сравнить средние с константами, заданными пользователем)).

После нажатия на кнопку *Summary* появится рабочая книга, содержащая таблицу с результатами анализа (рис. 4.12). В этой таблице имеются следующие столбцы:

- *Mean* – среднее значение, рассчитанное на основе выборочных данных (в нашем случае – средняя концентрация загрязнителя в 10 пробах почвы);
- *Std dv.* – стандартное отклонение;
- *N* – объем выборки;
- *Std. err.* – стандартная ошибка;
- *Reference constant* – контрольное значение;
- *df* – число степеней свободы;
- *P* – вероятность ошибочно отвергнуть нулевую гипотезу о том, что выборочная средняя не отличается от контрольной величины. В нашем случае $P > 0{,}05$. Таким образом, несмотря на некоторое превышение средней концентрации загрязнителя в некоторых пробах, в целом это превышение это является незначительным.



Variable	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
Концентрация	101.0000	6.306963	10	1.994437	100.0000	0.501395	9	0.628128

Рисунок 4.12. Результаты сравнения выборочной средней с контрольным значением.

Рукопись

Глава 5. Сравнение нескольких групп

Тест Стьюдента и его непараметрические аналоги, рассмотренные в предыдущей главе, предназначены для сравнения исключительно *двух выборок*. Однако очень часто исследователи допускают ошибку, используя *t*-тест для попарных сравнений более двух выборок (подробнее об этой проблеме см., например, в книге Гланц 1999). Во избежание данной ошибки необходимо использовать дисперсионный анализ (или «ANOVA» – от англ. *analysis of variance*).

5.1. Параметрический однофакторный дисперсионный анализ

Data: Fertilizers ...

	1 Удобрение	2 Урожайность
1	Старое	1920
2	Старое	2020
3	Старое	2060
4	Старое	1960
5	Старое	1960
6	Старое	2140
7	Старое	1980
8	Старое	1940
9	Старое	1790
10	Новое 1	2250
11	Новое 1	2410
12	Новое 1	2260
13	Новое 1	2200
14	Новое 1	2360
15	Новое 1	2320
16	Новое 1	2240
17	Новое 1	2300
18	Новое 1	2090
19	Новое 2	2252
20	Новое 2	2400
21	Новое 2	2261
22	Новое 2	2205
23	Новое 2	2368
24	Новое 2	2321
25	Новое 2	2249
26	Новое 2	2296
27	Новое 2	2100

Рисунок 5.1. Пример оформления данных для выполнения однофакторного дисперсионного анализа.

Для проверки эффективности двух новых удобрений был выполнен следующий эксперимент. На опытном поле случайным образом были выбраны 27 одинаковых по площади участков. Весной в 7 из них внесли «старое» удобрение, в 7 –новое удобрение 1, а в оставшиеся 7 – новое удобрение 2. В конце года была определена урожайность культуры, использованной в эксперименте. Полученные данные приведены на рис. 5.1. Вопрос: различается ли средняя урожайность культуры в зависимости от типа удобрения?

Воспользуемся *однофакторным дисперсионным анализом* (поскольку проверяется влияние лишь одного фактора – типа удобрения). Для его выполнения необходимо:

- Запустить модуль *One-way ANOVA* (рис. 5.2) из меню *Statistics* > *ANOVA*. Можно также нажать кнопку на дополнительной панели инструментов.

Рукопись

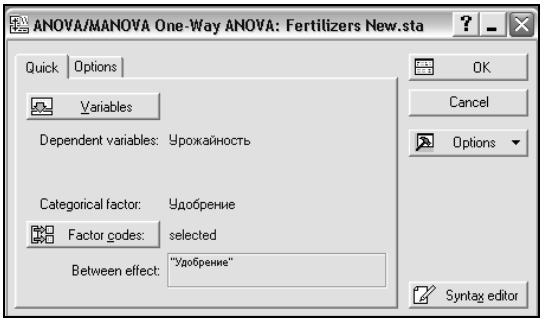


Рисунок 5.2. Окно модуля однофакторного дисперсионного анализа.

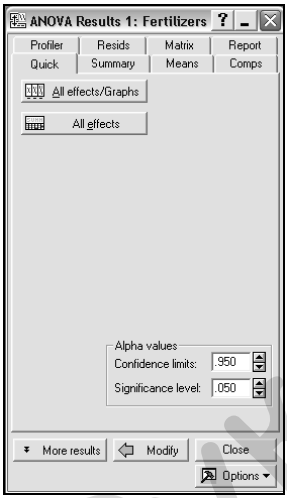


Рисунок 5.3. Окно выбора результатов дисперсионного анализа.

- Нажать на кнопку *Variables* и выбрать зависимую («Урожайность») и группирующую переменные («Удобрение»). Нажать на кнопки: *Factor codes* > *All* (так мы укажем программе, что в анализе должны участвовать все задействованные в эксперименте группы) > *OK* > *OK*. В итоге появится окно с 8 закладками (рис. 5.3). Автоматически программа откроет его на закладке *Quick* (Быстро).

Результаты анализа можно получить уже на данном этапе, если нажать на кнопку *All effects* (Все эффекты). Однако рассматриваемый вариант дисперсионного анализа является параметрическим, т.е. предполагает выполнение следующих обязательных условий в отношении данных: 1) в каждой из сравниваемых групп значения анализируемого признака распределяются *нормально*; 2) групповые *дисперсии однородны* (т.е. между ними нет статистически значимой разницы). Кроме того, все сравниваемые выборки должны быть независимыми. Поэтому перед получением результатов анализа следует проверить, выполняются ли указанные условия, и поступаем ли мы корректно, используя данный вариант дисперсионного анализа.

Рукопись

Для проверки условий ANOVA необходимо выполнить следующее:

- Нажать на кнопку *More results* (Дополнительные результаты), расположенную в нижней части окна *ANOVA Results*. В результате этого появится окно, представленное на рис. 5.4.

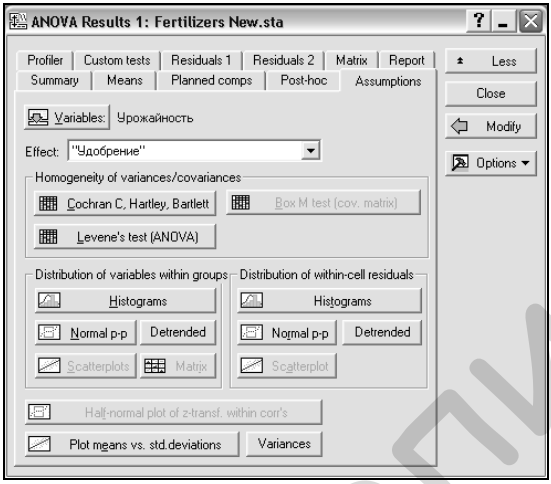


Рисунок 5.4. Окно дополнительных результатов дисперсионного анализа на закладке *Assumptions*.

- Открыть закладку *Assumptions* (Допущения). Для проверки однородности групповых дисперсий в поле *Homogeneity of variances/covariances* нажать на кнопку *Levene's test* (тест Левена). Если результат этого теста указывает на отсутствие различий между дисперсиями (P > 0,05), то применение параметрического варианта дисперсионного анализа является обоснованным. В нашем примере различий действительно нет (P = 0,993) (проверьте самостоятельно).
- Для проверки нормальности распределения анализируемых данных необходимо воспользоваться одной из опций, доступных в поле *Distribution of variables within groups* (Распределение переменных внутри групп). *Примечание*: если число наблюдений в сравниваемых группах невелико, лучше использовать график нормальных вероятностей (кнопка *Normal p-p*; см. также разд. 3.4). Если же их много, то можно оценить характер распределений, построив гистограммы (кнопка *Histograms*). При нажатии на одну их этих кнопок программа предложит список групп, участвующих в анализе. Пример графика, построенного на «вероятностной бумаге»

Рукопись

для «Старого» удобрения, приведен на рис. 5.5. Видно, что точки-наблюдения тесно укладываются вдоль теоретически ожидаемой прямой. Аналогичная ситуация характерна и для остальных двух групп из рассматриваемого примера. Таким образом, анализируемые данные по урожайности удовлетворяют обоим условиям *ANOVA*.

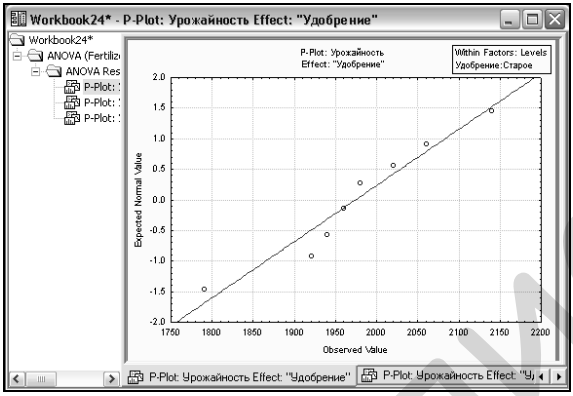
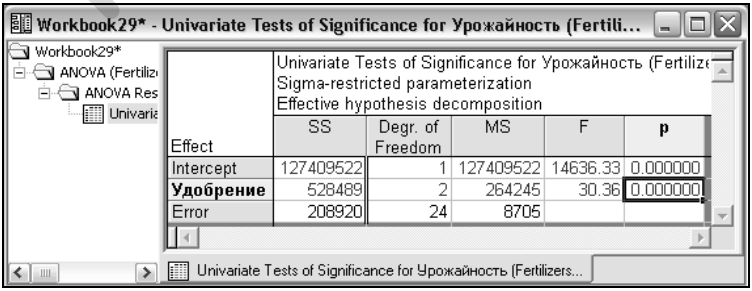


Рисунок 5.5. Окно дополнительных результатов дисперсионного анализа на закладке *Assumptions*.

- Наконец, необходимо на закладке *Summary* (рис. 5.4) нажать кнопку *Test all effects* (Проверить все эффекты). В появившейся таблице результатов (рис. 5.6) необходимо разыскать ячейку с величиной ошибки *P* для нулевой гипотезы об отсутствии связи между урожайностью и типом удобрения (строка «Удобрение»). Поскольку в нашем примере *P* << 0,05, можно заключить, что средняя урожайность культуры статистически значимо различается в зависимости от использованного удобрения.



Effect	SS	Degr. of Freedom	MS	F	p
Intercept	127409522	1	127409522	14636.33	0.000000
Удобрение	528489	2	264245	30.36	0.000000
Error	208920	24	8705		

Рисунок 5.6. Результаты однофакторного дисперсионного анализа.

Рукопись

5.2. Апостериорный анализ

Важно помнить, что дисперсионный анализ позволяет проверить лишь гипотезу об отсутствии различий между сравниваемыми группами в целом. Однако с его помощью невозможно узнать, какие именно группы различаются между собой. Для выяснения этого необходимо воспользоваться методами множественных сравнений, являющихся частью т.н. апостериорного анализа (*Post-hoc analysis*). Механизм их работы заключается в проведении попарных сравнений средних значений всех групп, включенных в дисперсионный анализ.

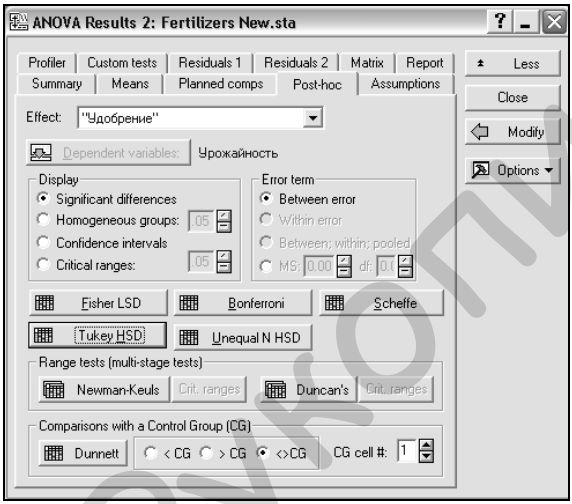


Рисунок 5.7. Окно дополнительных результатов дисперсионного анализа на закладке *Post-hoc*.

Для выполнения множественных сравнений необходимо открыть закладку *Post hoc* (рис. 5.7) в окне дополнительных результатов дисперсионного анализа (*More results*). Программа STATISTICA предлагает ряд тестов для множественных сравнений, несколько различающихся по мощности: *Fisher LSD*, *Bonferroni*, *Scheffe*, *Tukey HSD*, *Newman-Keuls*, *Duncan's*, *Dunnet*. Наиболее часто используемыми являются тесты Тьюки (*Tukey HSD*) и Ньюмена-Кейлса (*Newman-Keuls*). Нажатие на кнопку соответствующего теста приводит к появлению рабочей книги с матрицей значений *P*. Из рис. 5.8, например, видно, что статистически значимая разница в урожайности существует между парами удобрений «Старое – Новое 1» и

РУКОПИСЬ

«Старое – Новое 2» ($P < 0{,}05$), тогда как оба новых удобрения по эффективности не различаются ($P > 0{,}05$) (выполнен тест Тьюки для выборок с одинаковыми объемами).

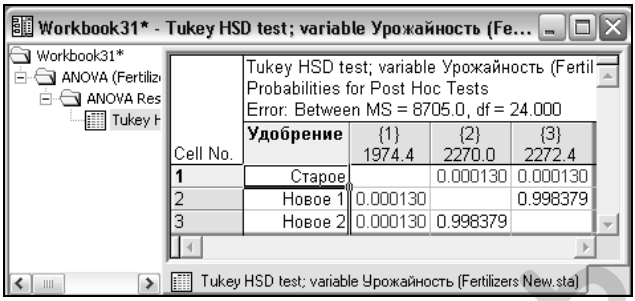


Рисунок 5.8. Результат выполнения теста Тьюки.

5.3. Параметрический двухфакторный дисперсионный анализ

В примере с тремя типами удобрений (предыдущий раздел) мы не обращали внимания на то, на каких опытных участках произрастала культура. Участки выбирались случайным образом, а это значит, что просто в силу случая они могли оказаться очень разными по своим физико-химическим свойствам, что, в свою очередь, также могло сказаться на урожайности. Теперь мы учтем это обстоятельство и выполним двухфакторный дисперсионный анализ (фактор 1 – тип удобрения, фактор 2 – тип почвы опытного участка). Поскольку в анализ включен дополнительный фактор, в таблицу с данными необходимо добавить еще одну группирующую переменную, которая будет содержать коды типов почвы (рис. 5.9). Для выполнения двухфакторного дисперсионного анализа в программе STATISTICA необходимо выполнить следующее:

- Запустить модуль *Factorial ANOVA* (Факторный дисперсионный анализ) из меню *Statistics* > *ANOVA*. Можно также нажать кнопку на дополнительной панели инструментов.
- В появившемся окне (рисунок не приводится) нажать кнопку *Variables* и выбрать зависимую («Урожайность») и

группирующие переменные («Удобрение» и «Тип почвы»). Нажать на кнопки: *Factor codes* > *All* > *OK* > *OK*. В итоге появится уже знакомое вам по рис. 5.3 окно с 8 закладками. (*Примечание*: при изложении дальнейшего материала предполагается, что анализируемые данные успешно прошли проверку на нормальность распределения и однородность групповых дисперсий – см. выше).

Data: Fertilizers New.sta* (3v ...

	1 Удобрение	2 Тип почвы	3 Урожайность
1	Старое	т	1920
2	Старое	п	2020
3	Старое	т	2060
4	Старое	п	1960
5	Старое	т	1960
6	Старое	т	2140
7	Старое	п	1980
8	Старое	п	1940
9	Старое	п	1790
10	Новое 1	т	2250
11	Новое 1	п	2410
12	Новое 1	п	2260
13	Новое 1	т	2200
14	Новое 1	т	2360
15	Новое 1	т	2320
16	Новое 1	п	2240
17	Новое 1	т	2300
18	Новое 1	п	2090
19	Новое 2	п	2252
20	Новое 2	т	2400
21	Новое 2	т	2261
22	Новое 2	т	2205
23	Новое 2	т	2368
24	Новое 2	п	2321
25	Новое 2	т	2249
26	Новое 2	п	2296
27	Новое 2	п	2100

Рисунок 5.9. Пример оформления данных для выполнения двухфакторного дисперсионного анализа. Обозначения: «Старое», «Новое 1», и «Новое 2» - типы удобрений; «п» и «т» - песчаная и торфянистая почва.

- Нажать на кнопку *All effects*. Это приведет к появлению таблицы с результатами дисперсионного анализа (рис. 5.10). Самое главное для нас в этой таблице – это вторая и третья строки. В конце этих строк приведены вероятности ошибок для нулевых гипотез об отсутствии влияния типа удобрения и типа почвы на урожайность. Видно, что лишь в случае с типом удобрения $P < 0{,}05$. Это говорит о значительном влиянии данного фактора на урожайность. Доказать подобный же эффект для типа почвы нам в данном эксперименте не удалось ($P > 0{,}05$). Строка «Удобрение × Тип почвы» касается взаимного влияния исследуемых факторов на урожайность. Как видим, взаимодействие между этими факторами также отсутствует ($P > 0{,}05$).

Аналогичным образом в программе STATISTICA можно выполнить дисперсионный анализ и с большим количеством факторов.

Рукопись

Workbook37* - Univariate Tests of Significance for Урожайность (Fertilizers New.sta)

Univariate Tests of Significance for Урожайность (Fertiliz
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	125818651	1	125818651	14318.19	0.000000
Удобрение	497392	2	248696	28.30	0.000001
Тип почвы	22003	1	22003	2.50	0.128504
Удобрение*Тип почвы	2383	2	1192	0.14	0.873949
Error	184534	21	8787		

Рисунок 5.10. Результаты двухфакторного дисперсионного анализа.

5.4. Дисперсионный анализ Фридмана

Рассмотренные выше варианты дисперсионного анализа помимо обязательных условий о нормальности и однородности групповых дисперсий предполагали также, что сравниваемые группы являются независимыми. В случае с зависимыми выборками необходимо воспользоваться *дисперсионным анализом Фридмана* (*Friedman ANOVA*). Следует отметить, что, являясь непараметрическим методом, анализ Фридмана не требует нормальности распределения данных и однородности дисперсий.

Data: Shell Length.sta (3v by 10c)

	1 Июнь	2 Июль	3 Август
Моллюск 1	3	5	7
Моллюск 2	4	6	7
Моллюск 3	4	5	6
Моллюск 4	5	6	8
Моллюск 5	2	4	6
Моллюск 6	3	6	8
Моллюск 7	4	6	8
Моллюск 8	4	5	7
Моллюск 9	3	5	8
Моллюск 10	5	7	8

Рисунок 5.11. Пример оформления данных для выполнения дисперсионного анализа Фридмана.

На рис. 5.11 представлены данные, полученные в ходе наблюдений за изменениями длины раковины моллюска *Sphaerium* sp. в летние месяцы. Десять особей были посажены в специальный садок, который затем установили в озере в типичном для моллюска биотопе. Каждый месяц измеряли длину раковины у всех 10 моллюсков. Необходимо выяснить, произошли ли существенные изменения средней длины раковины к концу опыта.

Рукопись

Поскольку измерения раковины выполнялись на одних и тех же особях *Sphaerium*, три полученные выборки являются зависимыми. Сравним их с помощью дисперсионного анализа Фридмана. Для этого необходимо:

- Запустить модуль анализа (рис. 5.12) из меню *Statistics* > *Nonparametrics* > *Comparing multiple dependent samples* (Сравнение нескольких зависимых выборок). Можно также нажать кнопку на дополнительной панели инструментов.
- Нажать кнопку *Variables* и выбрать переменные, которые должны участвовать в анализе.
- Нажать кнопку *Summary: Friedman ANOVA and Kendall's concordance* (*Результат: ANOVA по Фридману и критерий согласованности Кендалла*).
- В появившейся таблице с результатами необходимо отыскать величину ошибки *P* для нулевой гипотезы о том, что в течение трех месяцев наблюдений длина раковины у моллюсков существенно не изменилась. Эта величина находится в заголовке таблицы (рис. 5.13). При $P < 0{,}05$ (как в нашем случае) можно сделать вывод о наличии статистически значимых различий между группами. В этом же заголовке приводится т.н. коэффициент согласованности Кендалла. Он рассчитывается путем усреднения коэффициентов корреляции Спирмена (см. разд. 6.3) для каждой пары участвующих в анализе групп. Чем больше различия между группами, тем ближе коэффициент Кендалла к 1.

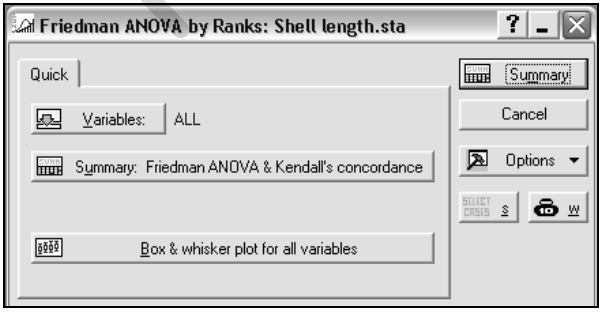


Рисунок 5.12. Модуль дисперсионного анализа Фридмана.

РУКОПИСЬ

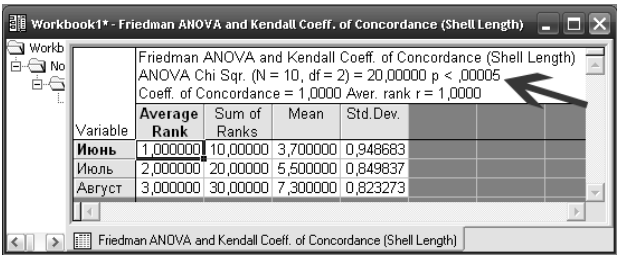


Рисунок 5.13. Результаты дисперсионного анализа Фридмана. Вероятность ошибки *P* указана стрелкой.

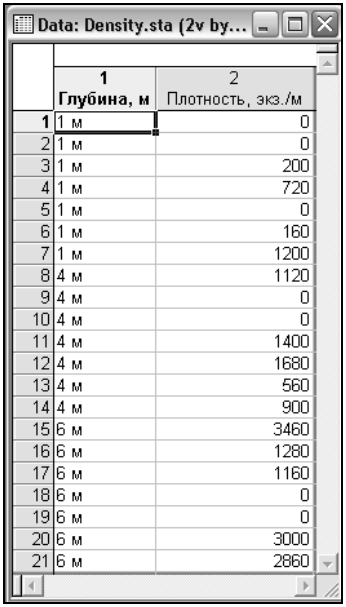
5.5. Дисперсионный анализ Крускала-Уоллиса

Как уже отмечалось в главе 3, в биологических исследованиях данные распределены по нормальному закону достаточно редко. И даже если выборочные единицы отбираются из нормально распределенных генеральных совокупностей, объем выборок часто оказывается слишком малым для того, чтобы вообще сделать какие-либо выводы относительно вида распределения. Это делает параметрический дисперсионный анализ неприменимым. Выходом становится использование непараметрического *дисперсионного анализа Крускала-Уоллиса* (или *H*-теста) (*Kruskal-Wallis ANOVA*), хотя он и обладает несколько меньшей мощностью в сравнении с параметрическим вариантом.

Неподчинение данных нормальному распределению особенно часто встречается в экологических исследованиях, например, при определении плотности популяций животных или растений методом учетных площадок. На рис. 5.14 представлены данные о плотности популяции моллюска *Dreissena polymorpha* на разных глубинах озера. Как видно, число наблюдений для каждой из глубин невелико (n = 7). При этом даже если объединить все имеющиеся данные в одну совокупность, окажется, что их распределение далеко от нормального (проверьте самостоятельно).

Чтобы выяснить, различается ли плотность популяции моллюска на разных глубинах, применим дисперсионный анализ Крускала-Уоллиса. Для этого в программе STATISTICA необходимо выполнить следующее:

Рукопись



Data: Density.sta (2v by...

	1 Глубина, м	2 Плотность, экз./м
1	1 м	0
2	1 м	0
3	1 м	200
4	1 м	720
5	1 м	0
6	1 м	160
7	1 м	1200
8	4 м	1120
9	4 м	0
10	4 м	0
11	4 м	1400
12	4 м	1680
13	4 м	560
14	4 м	900
15	6 м	3460
16	6 м	1280
17	6 м	1160
18	6 м	0
19	6 м	0
20	6 м	3000
21	6 м	2860

Рисунок 5.14. Пример оформления данных для выполнения дисперсионного анализа Крускала-Уоллиса.

- Запустить модуль анализа (рис. 5.15) из меню *Statistics* > *Nonparametrics* > *Comparing multiple independent samples* (Сравнение нескольких независимых выборок). Можно также нажать кнопку на дополнительной панели инструментов.
- Нажать кнопку *Variables* и выбрать зависимую («Плотность») и группирующую («Глубина») переменные. Нажать на кнопки: *Factor codes* > *All* > *OK* > *OK*.
- Нажать на кнопку *Summary: Kruskal-Wallis ANOVA and Median test* (Результат: ANOVA по Крускалу-Уоллису и медианный тест).

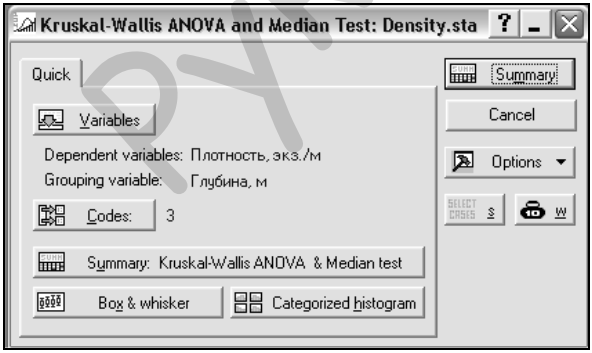


Рисунок 5.15. Модуль дисперсионного анализа Крускала-Уоллиса.

- В появившейся таблице с результатами (рис. 5.16) необходимо отыскать величину ошибки *P* для нулевой гипотезы о том, что плотность популяции дрейссены на исследованных глубинах не различается. Если *P* > 0,05 (как в

РУКОПИСЬ

нашем примере), то следует вывод об отсутствии различий между сравниваемыми группами. Вместе с результатом *Kruskal-Wallis ANOVA* на отдельном листе в рабочей книге программа выдает также результаты т.н. *медианного теста*. Этот тест проверяет ту же нулевую гипотезу, что и *H*-тест Крускала-Уоллиса, однако является менее мощным.

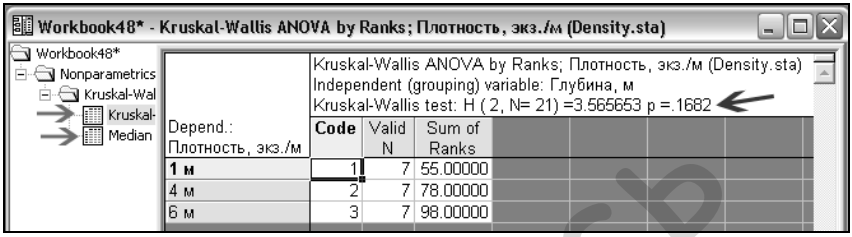


Рисунок 5.16. Результаты дисперсионного анализа Крускала-Уоллиса. Величина ошибки *P* указана темной стрелкой справа. На отдельном листе (см. светлые стрелки слева) программа выдает также результаты медианного теста.

Глава 6. Корреляционный анализ

Ответьте на вопрос: «Есть ли связь между ростом и весом тела человека?»

Правильно, есть, и она положительна: в целом, чем человек выше, тем он тяжелее. Но насколько сильна данная связь? Ответить на этот второй вопрос уже сложнее, т.к. далеко не все высокие люди много весят, равно как и не все низкие люди худые. Однако выход есть – его дает область статистики, называемая корреляционным анализом. Корреляционный анализ позволяет сделать заключение не только о том, какова связь между двумя признаками по направлению (прямая или обратная), но и, что очень важно, выразить ее количественно при помощи *коэффициента корреляции* – величины, изменяющейся от -1 до +1. Чем ближе коэффициент к 1 (по модулю), тем сильнее связь между признаками. Знак коэффициент указывает на направлении зависимости. В этой главе мы рассмотрим, как выполнить корреляционный анализ при помощи программы STATISTICA.

Рукопись

6.1. Коэффициент корреляции Пирсона

Использование *коэффициента корреляции Пирсона* для оценки степени связи между двумя признаками предполагает выполнение следующих двух обязательных условий:

- значения обоих анализируемых признаков распределены *нормально*;
- связь между признаками является *линейной*.

Способы проверки данных на нормальность распределения мы рассмотрели ранее (глава 3). О том, как установить линейность зависимости между признаками, изложено ниже.

Предположим, необходимо выяснить наличие связи между длиной крыла и длиной хвоста у некоторого вида птицы. Для этого были выполнены соответствующие измерения у 12 особей. Полученные данные приведены на рис. 5.1.

Data: Bird Correlation.sta (2v ...

	1 Длина крыла, см	2 Длина хвоста, см
1	10.4	7.4
2	10.8	7.6
3	11.1	7.9
4	10.2	7.2
5	10.3	7.4
6	10.2	7.1
7	10.7	7.4
8	10.5	7.2
9	10.8	7.8
10	11.2	7.7
11	10.6	7.8
12	11.4	8.3

Рисунок 6.1. Пример оформления данных для расчета коэффициента корреляции Пирсона.

Для расчета коэффициента корреляции Пирсона необходимо выполнить следующее:

- Запустить модуль анализа из меню *Statistics* > *Basic Statistics/Tables* > *Correlation Matrices* (Корреляционные матрицы). Можно воспользоваться также кнопкой на дополнительной панели инструментов.
- В появившемся окне выбрать переменные, которые должны участвовать в анализе. Для этого нужно нажать либо кнопку *One variable list* (Один список переменных) либо *Two lists* (*rect. matrix*) (Два списка (прямоугольная матрица)). В первом случае анализируемые переменные последовательно выбираются из одного списка, а во втором – из двух (рис. 6.2).
- Проверить условия применимости коэффициента Пирсона (см. выше). Для визуальной оценки выполнения этих условий можно нажать кнопку *Scatterplot matrix for selected variables*

Рукопись

(Диаграмма рассеяния для выбранных переменных). В результате программа построит точечный график, по осям которого будут отложены значения соответствующих переменных (рис. 6.3). Диагональная линия на этом графике служит для оценки линейности связи между анализируемыми признаками. Если точки-наблюдения укладываются вдоль этой линии на близком расстоянии, можно говорить о существовании прямолинейной зависимости. Вместе с диаграммой рассеяния программа строит также распределения значений анализируемых признаков в виде гистограмм, по форме которых можно проверить условие о нормальности распределения. (*Примечание*: в рассматриваемом примере условие нормальности распределения явно не выполняется, однако в учебных целях мы продолжим расчет коэффициента корреляции Пирсона).

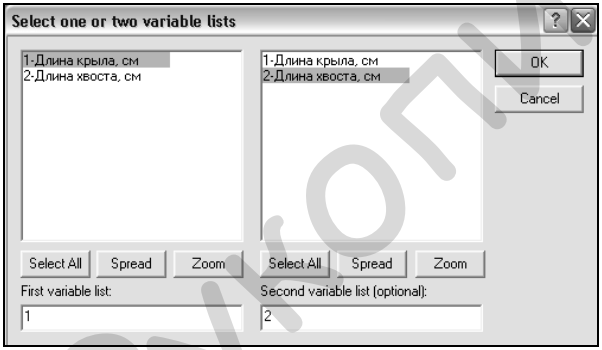


Рисунок 6.2. Выбор переменных для расчета коэффициента корреляции Пирсона.

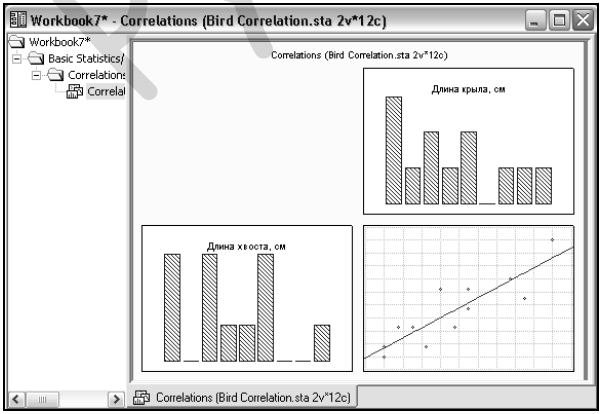
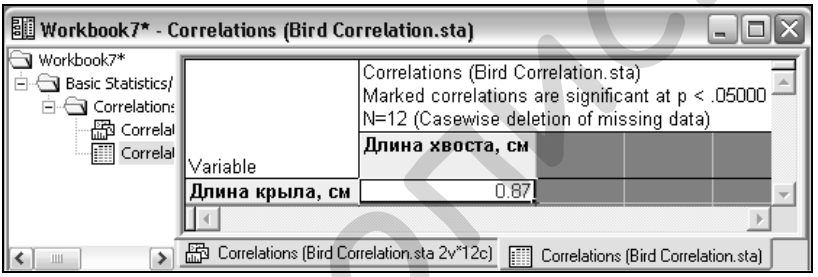


Рисунок 6.3. Визуальная оценка условий применимости коэффициента корреляции Пирсона

Рукопись

- Нажать на кнопку *Summary: Correlation matrix* (Результат: Корреляционная матрица). В результате появится таблица, содержащая рассчитанный программой коэффициент корреляции (рис. 6.4). В нашем случае коэффициент оказался очень высоким (r = 0,87), что указывает на существование тесной связи между длиной крыла и длиной хвоста у исследуемого вида птиц. Одновременно с расчетом коэффициента программа оценивает и его статистическую значимость, т.е. проверяет нулевую гипотезу о том, что в действительности связь между признаками отсутствует. Статистически значимые коэффициенты корреляции Пирсона в программе STATISTICA выделяются красным цветом (P < 0,05).



6.4. Результат расчета коэффициента Пирсона.

6.2. Сравнение двух коэффициентов корреляции Пирсона

В ряде случаев возникает необходимость сравнить два коэффициента корреляции, рассчитанных на основе разных выборок. Рассмотрим следующий пример. В ходе обследования 98 рыб одного вида выяснилось, что коэффициент корреляции между общей длиной тела и длиной головы у них составляет 0,78. У другой популяции того же вида рыб (обследовано 95 особей) эта же зависимость выражалась несколько более высоким коэффициентом – 0,84. Случайна ли разница между этими двумя коэффициентами?

Нулевая гипотеза, которую нам предстоит проверить, заключается в том, что разница действительно случайна. Для ее проверки воспользуемся модулем программы, который

Рукопись

называется *Difference tests* (Тесты на различие). Он находится в меню *Statistics* > *Basic Statistics/Tables* > *Difference tests: r, %, means*. Внешний вид модуля приведен на рис. 6.5.

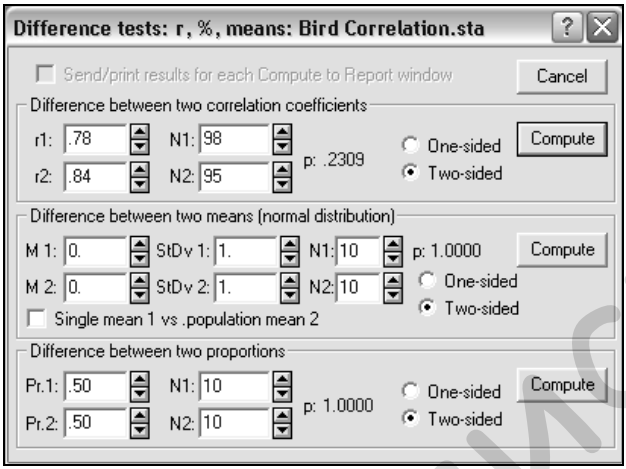


Рисунок 6.5. Окно модуля *Difference tests*.

Рассмотрите рис. 6.5 внимательно. В верхней части изображенного на нем диалогового окна имеется поле *Difference between two correlation coefficients* (Различие между двумя коэффициентами корреляции) с четырьмя ячейками, в которые необходимо ввести соответствующие данные для выполнения анализа. В ячейки *r*1 и *r*2 вводятся значения сравниваемых коэффициентов корреляции, в *N*1 и *N*2 – объемы выборок, на основе которых эти коэффициенты были рассчитаны. Рядом с ячейками нужно указать, какой вариант теста мы собираемся выполнять – одно- (*One-sided*) или двухсторонний (*Two-sided*). Поскольку мы просто хотим выяснить наличие разницы между коэффициентами, оставим *Two-sided*. Если бы стояла необходимость проверить гипотезу о том, что один из коэффициентов значительно больше другого, то нужно было бы выбрать опцию *One-sided*. Наконец, следует нажать кнопку *Compute* (Рассчитать). В том же поле программа покажет, чему равна вероятность справедливости нулевой гипотезы. В нашем примере P = 0,2309, что больше обычно принимаемого уровня «0,05». Следовательно, наблюдаемая разница между коэффициентами корреляции является случайной.

РУКОПИСЬ

6.3. Коэффициент корреляции Спирмена

При расчете коэффициента корреляции Пирсона для длины крыла и хвоста птиц (рис. 6.1) мы столкнулись с тем, что значения этих признаков не были распределены нормально (рис. 6.3). В подобных ситуациях применение коэффициента Пирсона может приводить к выводам, не соответствующим действительности. Вместо него следует воспользоваться одним из непараметрических коэффициентов корреляции. Из последних наиболее обычен *ранговый коэффициент корреляции Спирмена*. Рассчитаем его для тех же данных о длине крыла и хвоста у птиц (рис. 6.1). Для этого необходимо выполнить следующее:

- Запустить модуль *Nonparametric correlations* (Непараметрические корреляции) из меню *Statistics* > *Nonparametrics* > *Correlations* (*Spearman, Kendall tau, gamma*) (Корреляции (Спирмена, тау Кендалла, гамма)).
- В появившемся окне (рис. 6.6) нажать на кнопку *Variables* и выбрать столбцы, содержащие необходимые данные.

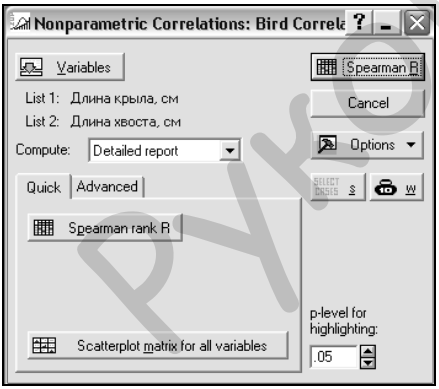


Рисунок 6.6. Модуль непараметрического корреляционного анализа.

- Нажать кнопку *Spearman R* или *Spearman rank R*. Появится таблица с результатами анализа (рис. 6.7), которая содержит столбцы *Valid N* (число наблюдений), *Spearman R* (коэффициент корреляции Спирмена), $t(N\text{-}2)$ (значение критерия Стьюдента для числа степеней свободы n-2), и P (вероятность ошибки для нулевой гипотезы об отсутствии связи между признаками).

В нашем примере коэффициент корреляции Спирмена оказался несколько ниже рассчитанного ранее коэффициента Пирсона (0,85 против 0,87). При этом он является в высокой степени статистически значимым ($P << 0{,}05$).

Рукопись

Рисунок 6.7. Результат расчета коэффициента корреляции Спирмена.

6.4. Коэффициент ассоциации (связанности)

Коэффициенты корреляции Пирсона и Спирмена применяются для оценки связи между количественными признаками. Однако многие биологические исследования сопряжены с изучением качественных величин (окраска, пол, наличие или отсутствие определенного состояния, и т.п.). Для таких признаков также можно определить степень «связанности». Один из показателей, который позволяет это сделать – *коэффициент ассоциации* φ, который изменяется от 0 до 1. Чем ближе φ к 1, тем сильнее связь.

Рассмотрим пример из иммунологии. Совокупность из 111 мышей разделили на две группы: по 57 и 54 мыши. Первой группе мышей сделали инъекцию патогенных бактерий при последующем введении сыворотки с антителами. Животные из второй группы служили контролем – им сделали только инъекции бактерий. После некоторого времени инкубации оказалось, что всего погибло 38 мышей, а выжило 73. Из погибших 13 принадлежали первой группе, а 25 – ко второй (контрольной). Вопрос: есть ли связь между введением сыворотки и выживаемостью мышей?

Данные описанного эксперимента удобно представить в виде т.н. таблицы сопряженности размером 2×2 (= четырехпольная таблица, или таблица с двумя входами):

	Погибло	Выжило
Бактерии + сыворотка	13	44
Только бактерии	25	29

РУКОПИСЬ

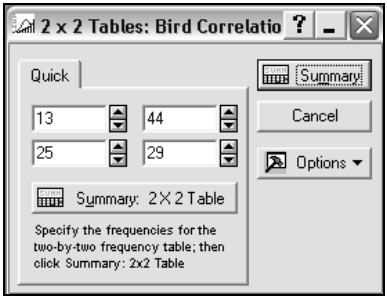
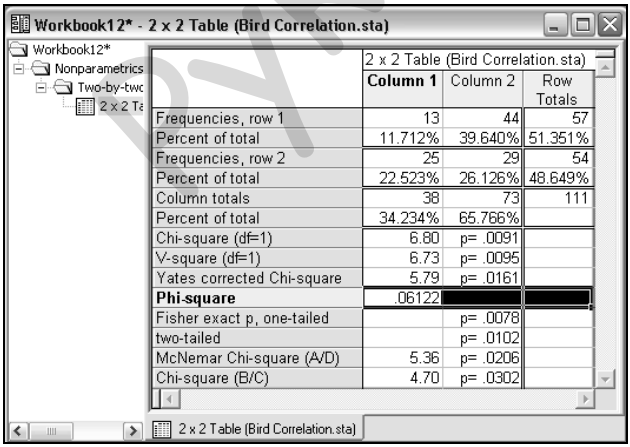


Рисунок 6.8. Модуль анализа таблиц сопряженности 2×2.

Для расчета коэффициента ассоциации необходимо:

- Запустить модуль анализа четырехпольных таблиц (рис. 5.11) из меню *Statistics* > *Nonparametrics* > 2×2 *Tables*.
- В ячейки появившейся панели (рис. 6.8) ввести данные о численности мышей в каждой из экспериментальных групп (в соответствии с приведенной выше таблицей).
- Нажать кнопку *Summary*. В результате этого появится таблица с большим набором статистических показателей (рис. 6.9). Нам необходима строка *Phi-square*, в которой находится значение коэффициента ассоциации, возведенное в квадрат для придания ему положительного значения. В нашем случае фи-квадрат равен 0,061. После извлечения корня получаем φ = 0,247. Таким образом, полученное значение фи указывает на достаточно слабую связь между введением сыворотки и выживаемостью зараженных мышей.



Workbook12* - 2 x 2 Table (Bird Correlation.sta)

Workbook12*
Nonparametrics
Two-by-two
2 x 2 Ta

2 x 2 Table (Bird Correlation.sta)

	Column 1	Column 2	Row Totals
Frequencies, row 1	13	44	57
Percent of total	11.712%	39.640%	51.351%
Frequencies, row 2	25	29	54
Percent of total	22.523%	26.126%	48.649%
Column totals	38	73	111
Percent of total	34.234%	65.766%	
Chi-square (df=1)	6.80	p= .0091	
V-square (df=1)	6.73	p= .0095	
Yates corrected Chi-square	5.79	p= .0161	
Phi-square	.06122		
Fisher exact p, one-tailed		p= .0078	
two-tailed		p= .0102	
McNemar Chi-square (A/D)	5.36	p= .0206	
Chi-square (B/C)	4.70	p= .0302	

2 x 2 Table (Bird Correlation.sta)

Рисунок 6.9. Результат анализа таблицы сопряженности 2×2.

РУКОПИСЬ

Заметьте, что модуль 2×2 *Tables* можно использовать также для сравнения частот бинарных качественных признаков в двух группах. Так, в таблице на рис. 6.9 помимо коэффициента ассоциации приведены значения: *критерия* χ^2 без поправки Йетса и с ней; *точного критерия Фишера*; *критерия Мак-Нимара* для зависимых групп. Анализируя, например, результат теста χ^2, можно заключить, что несмотря на установленный нами слабый эффект от введения сыворотки с антителами, выживаемость мышей в контрольной и экспериментальной группах все же статистически значимо различается (P = 0,0091; см. строку *Chi-square (df = 1)* на рис. 6.9).

Глава 7. Регрессионный анализ

В предыдущей главе мы научились оценивать направление и степень связи между двумя признаками с помощью корреляционного анализа. Хотя коэффициент корреляции и позволяет количественно охарактеризовать степень связи, с его помощью невозможно предсказать, чему в среднем будет равно значение одного признака при заданном значении другого признака. Решить эту задачу позволяет регрессионный анализ.

7.1. Оценка коэффициентов линейной регрессии

Достаточно часто связь между двумя биологическими признаками имеет линейный характер, что, как известно, можно выразить в виде уравнения

$$y = a + bx, \text{ где}$$

- y и x – анализируемые признаки;
- a – *свободный член* уравнения; при b = 0 получаем $y = a$, т.е. a – это точка, в которой линия регрессии пересекается с осью OY (эту точку называют также «y-пересечением», или «*Intercept*»);
- b – *коэффициент регрессии*, отражающий угол наклона линии регрессии. Чем больше b отличается от 0, тем сильнее связь между анализируемыми признаками.

Даже если связь между биологическими признаками носит нелинейный характер (например, экспоненциальный),

Рукопись

практически всегда можно выделить участки, хорошо аппроксимируемые линейной регрессией. Поэтому именно с нее мы и начнем рассмотрение регрессионного анализа.

Приведенное выше уравнение можно использовать для описания связи между двумя признаками лишь при выполнении следующих обязательных условий:

- Зависимость между признаками носит *линейный* характер;
- Оба признака распределены *нормально*.

На рис. 7.1 приведены данные о систолическом давлении крови у людей разного возраста. Рассчитаем коэффициенты линейного регрессионного уравнения, описывающего связь между этими двумя биологическими параметрами.

Data: Pressure data.sta (2v ...

	1 Возраст, лет	2 Давление, мм.рт.ст.
1	30	108
2	30	110
3	40	125
4	40	120
5	40	118
6	50	132
7	50	137
8	50	134
9	60	148
10	60	151
11	60	146
12	60	147
13	70	162
14	70	156
15	70	164
16	70	159

Рисунок 7.1. Пример оформления данных для выполнения регрессионного анализа.

Расчет коэффициентов регрессионных уравнений можно выполнить в нескольких модулях программы STATISTICA. Мы воспользуемся модулем *Multiple Regression Analysis* (Анализ множественной регрессии). Для выполнения регрессионного анализа необходимо:

- Запустить соответствующий модуль из меню *Statistics* или нажать на кнопку на дополнительной панели инструментов.
- В появившемся окне нажать на кнопку *Variables* и указать, какая из анализируемых переменных является зависимой (*Dependent variable*), а какая – независимой (*Independent variable*) (в нашем примере систолическое давление зависит от возраста). Нажать кнопку *OK*. В итоге появится окно (рис. 7.2), которое уже на данном этапе анализа содержит некоторые важные его результаты:

  а) *Dependent*: имя зависимой переменной;

  б) *No. of cases*: число наблюдений;

Рукопись

в) *Intercept*: значение свободного члена регрессионного уравнения;
г) *Std. error*: стандартная ошибка свободного члена регрессионного уравнения;
д) *Multiple R*: коэффициент множественной корреляции;
е) R^2: коэффициент детерминации. Это очень важный показатель в регрессионном анализе. Он изменяется от 0 до 1 и отражает «качество» рассчитанной регрессии, показывая долю (%) общего разброса выборочных точек, которая «объясняется» построенной регрессией (например, при R^2 = 0,85, следует вывод о том, что 85% дисперсии зависимой переменной y объясняется вариацией независимой переменной x);
ж) *Adjusted* R^2: скорректированный на число степеней свободы коэффициент детерминации (*Adjusted R-square* = 1 - (1 - *R-square*)×[n/(n - p)], где n – число наблюдений, p – число независимых переменных плюс 1);
з) *Standard error of estimate*: параметр, отражающий степень разброса выборочных значений относительно линии регрессии;
и) *F, df* и *p*: *F*-критерий, число степеней свободы, принятое при его расчете, и вероятность ошибки для нулевой гипотезы *F*-теста. *F*-тест в регрессионном анализе применяется для оценки статистической значимости модели (см., например, книгу Гланц 1999). При $P < 0{,}05$ можно заключить, что рассчитанная регрессия удовлетворительно описывает связь между исследуемыми признаками;
к) *t(df)* и *p*: критерий Стьюдента *t* используется для проверки нулевой гипотезы о равенстве 0 свободного члена регрессионного уравнения. *P* – вероятность ошибки для этой нулевой гипотезы;
л) *beta*: стандартизованный коэффициент регрессии – это коэффициент регрессии, который мы получили бы в случае предварительной стандартизации обеих переменных (т.е. при таком преобразовании, когда их средние значения стали бы равны 0, а стандартные отклонения -1). Расчет *beta* позволяет оценить, в какой степени значения зависимой переменной определяются значениями независимой переменной. *Beta* может оказаться особенно полезным показателем при включении в анализ нескольких

независимых переменных, выражающихся в разных единицах измерения – в таком случае коэффициент отражал бы удельный вклад каждой из этих переменных в вариацию зависимой переменной. При наличии одной независимой переменной коэффициент *beta* идентичен *Multiple R*.

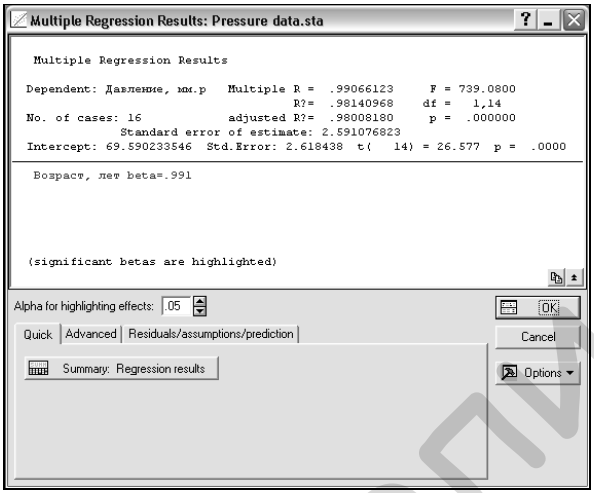


Рисунок 7.2. Окно предварительных результатов регрессионного анализа.

- Нажать кнопку *Summary: Regression results* (Результаты регрессионного анализа). Появится таблица со следующими результатами анализа (рис. 7.3):

  а) *Beta*: стандартизованный коэффициент регрессии;

  б) *Std. err. of beta*: стандартная ошибка стандартизованного коэффициента регрессии;

  в) *B*: один из самых важных столбцов в этой таблице, поскольку именно он содержит искомые значения свободного члена регрессионного уравнения (в строке *Intercept*) и коэффициента регрессии (нижняя строка таблицы);

  г) *Std. err. of B*: стандартные ошибки коэффициентов уравнения;

  д) *t(df)*: значения *t*-критерия Стьюдента, который используется для проверки гипотезы о равенстве обоих коэффициентов уравнения 0;

  е) *p-level*: вероятность ошибки для нулевой гипотезы о равенстве коэффициентов уравнения нулю.

РУКОПИСЬ

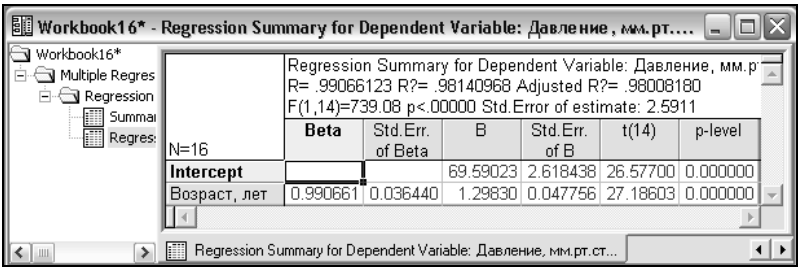


Рисунок 7.3. Результаты регрессионного анализа.

Из рис. 7.3 видно, что оба коэффициента регрессии статистически значимо отличаются от 0 ($P << 0{,}001$) и что в целом построенная регрессионная модель отлично описывает связь между возрастом людей и систолическим давлением крови ($R^2 = 98\%$). Само же рассчитанное уравнение мы можем записать следующим образом:

$$H = 1{,}298 \times A + 69{,}590,$$

где H – давление, A – возраст человека.

Важной частью регрессионного анализа является т.н. *анализ остатков* (остатки представляют собой разности между наблюдаемыми значениями зависимой переменной и теми ее значениями, которые предсказываются регрессионной моделью). Он запускается путем нажатия кнопки *Perform residual analysis* (Выполнить анализ остатков) на закладке *Residuals / Assumptions / Predictions* (Остатки / Условия / Предсказания) (рис. 7.2, 7.4).

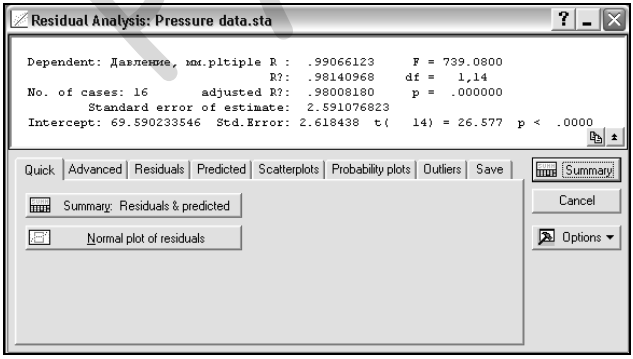


Рисунок 7.4. Подмодуль анализа остатков модуля множественного регрессионного анализа.

Рукописе

Первое, что нужно проверить в отношении остатков – это *нормальность* их распределения. Для этого на закладке *Quick* подмодуля анализа остатков (рис. 7.4) можно нажать кнопку *Normal plot of residuals*, чтобы построить график нормальных вероятностей (см. разд. 3.4). Если точки на этом графике достаточно тесно укладываются вдоль теоретически ожидаемой прямой, то можно заключить, что остатки распределяются нормально (рис. 7.5). Иначе линейная регрессионная модель для анализируемых переменных будет неприменима (в ряде случаев, однако, помогает трансформация данных, см. ниже).

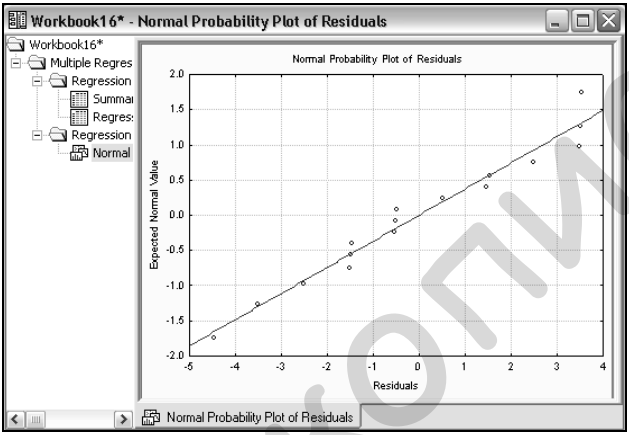


Рисунок 7.5. Результат проверки нормальности распределения остатков с помощью графика нормальных вероятностей.

Второе условие в отношении остатков состоит в том, что их *дисперсия должна оставаться неизменной* во всем диапазоне значений анализируемых переменных. Для проверки этого условия на закладке *Scatterplots* (Диаграммы рассеяния) (рис. 7.4) можно нажать кнопку *Predicted vs. residuals*, чтобы построить график зависимости значений остатков от предсказываемых моделью значений зависимой переменной. Если проверяемое условие выполняется, то точки на этом графике будут располагаться хаотично, не проявляя никакой закономерности (рис. 7.6). Если же в расположении точек имеется тенденция (разброс увеличивается слева направо, точки тесно укладываются вдоль прямой, и т.п.), линейный регрессионный анализ также неприменим (однако и в этом случае может помочь трансформация исходных данных, см. ниже).

РУКОПИСЬ

В рассмотренном примере оба условия в отношении остатков выполняются, что еще раз подтверждает адекватность рассчитанной регрессионной модели для описания связи между артериальным давлением и возрастом людей.

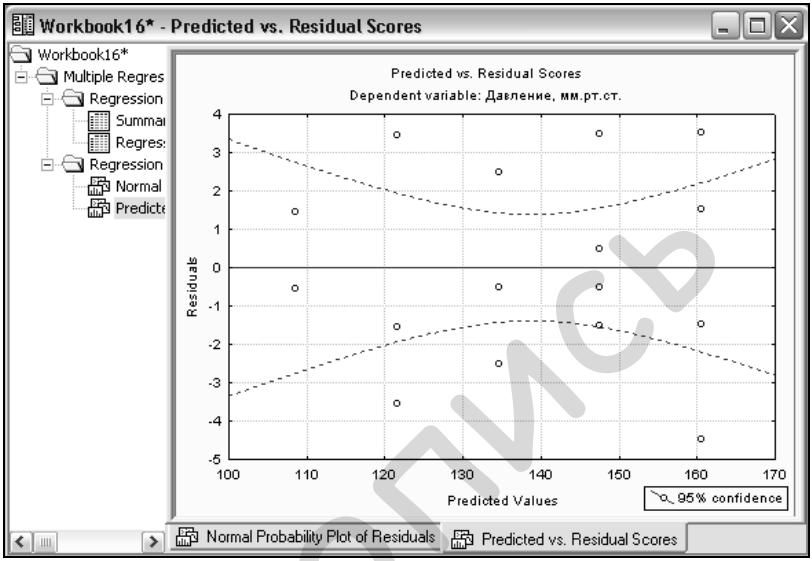


Рисунок 7.6. Результат проверки однородности дисперсии остатков.

7.2. Трансформация нелинейно связанных признаков

Многие биологические признаки проявляют нелинейный характер связи. Например, размеры тела и интенсивность метаболических процессов имеют степенную или экспоненциальную зависимость. Как же быть, если встает задача по расчету уравнения зависимости между подобными переменными? Иногда в таких случаях помогает определенная трансформация исходных данных, которая позволяет перевести их в другую шкалу измерения и тем самым «выровнять» нелинейную зависимость между признаками.

На рис. 7.7 приведены данные об интенсивности дыхания рачков *Daphnia magna* разных размеров. Необходимо рассчитать уравнение, описывающее связь между этими двумя признаками.

Рукопись

Data: Daphnia data.sta* (4v by 10c)

	1 Длина рачка, мм	2 R, мгO_2/особь*ч	3 Var3	4 Var4
1	0.556	0.018		
2	0.600	0.021		
3	0.694	0.035		
4	0.779	0.055		
5	0.849	0.130		
6	0.954	0.280		
7	1.099	0.480		
8	1.102	0.540		
9	1.204	0.720		
10	1.205	0.731		

Рисунок 7.7. Пример двух нелинейно связанных биологических признаков.

Характер связи между двумя переменными можно проверить еще до запуска модуля регрессионного анализа. Для этого достаточно построить диаграмму рассеяния (*Graphs > Scatterplots*), подобную приведенной на рис. 7.8. Из этого рисунка видно, что связь между размером дафний и интенсивностью потребления ими кислорода далека от линейной и больше напоминает степенную зависимость вида $y = ax^b$. Линейный регрессионный анализ здесь неприменим. Однако степенные и экспоненциальные зависимости обычно легко можно привести к линейным путем логарифмирования значений одного или (чаще) обоих анализируемых признаков. Такую трансформацию в программе STATISTICA выполнить очень просто. Еще в главе 2 говорилось о том, что при подготовке данных к анализу в окне настройки свойств переменных последним можно присваивать т.н. *длинные имена* (*Long name*) в виде формул (разд. 2.1).

Рукопись

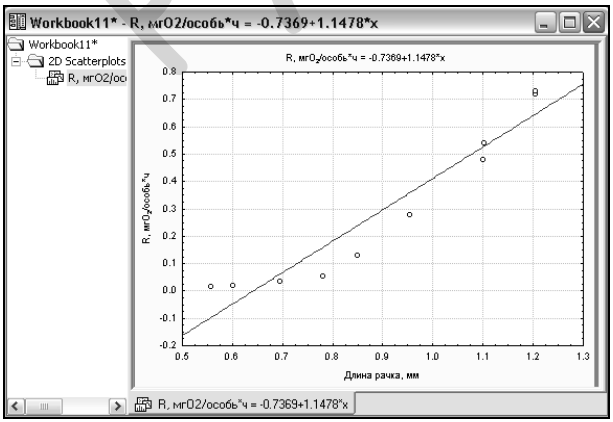


Рисунок 7.8. Визуальная оценка характера связи между двумя признаками при помощи диаграммы рассеяния.

Прологарифмируем столбец 1, содержащий значения длины тела дафний (рис. 7.7). Для этого воспользуемся столбцом 3, который следует за данными по интенсивности дыхания. Кликнув два раза по его заголовку, мы попадем в окно настроек свойств переменной. В поле *Long name* необходимо ввести формулу =*log*10(*v*1), где *v*1 – это столбец с данными о длине рачков. В поле *Name* введем короткое имя переменной, например, «*Log L*» (рис. 7.9).

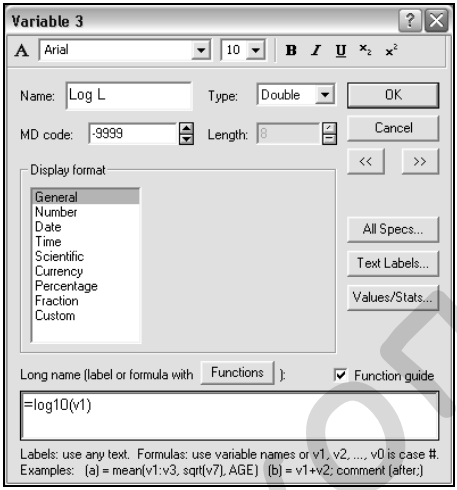


Рисунок 7.9. Введение формулы =*log*10(*v*1) в поле длинного имени переменной.

После нажатия на кнопку *OK* появится небольшая панель с надписью *«Expression OK. Recalculate the variable now?»* (Формула введена правильно. Пересчитать значения переменной?). Нажимаем *Yes*. В результате в третьем столбце таблицы с данными появятся пересчитанные значения первого столбца.

Аналогичную операцию логарифмирования (рис. 7.10) следует выполнить и для столбца с данными по интенсивности дыхания. Для этого можно воспользоваться 4-м столбцом таблицы и в качестве его длинного имени ввести формулу =*log*10(*v*2). Если теперь мы построим диаграмму рассеяния для прологарифмированных значений анализируемых признаков, то увидим, что точки гораздо теснее укладываются вдоль прямой линии и мы можем применить обычный регрессионный анализ для нахождения зависимости между признаками (рис. 7.11). Регрессионное уравнение для прологарифмированых переменных запишется в виде $log\ y = b \times log\ x + log\ a$.

РУКОПИСЬ

Data: Daphnia data.sta* (4v by 10c)

	1 Длина рачка, мм	2 R, мгO_2/особь*ч	3 **Log L**	4 Log R
1	0.556	0.018	-0.25493	-1.74473
2	0.600	0.021	-0.22185	-1.67778
3	0.694	0.035	-0.15864	-1.45593
4	0.779	0.055	-0.10846	-1.25964
5	0.849	0.130	-0.07109	-0.88606
6	0.954	0.280	-0.02045	-0.55284
7	1.099	0.480	0.040998	-0.31876
8	1.102	0.540	0.042182	-0.26761
9	1.204	0.720	0.080626	-0.14267
10	1.205	0.731	0.080987	-0.13608

Рисунок 7.10. Результат логарифмирования первого и второго столбцов в таблице с данными.

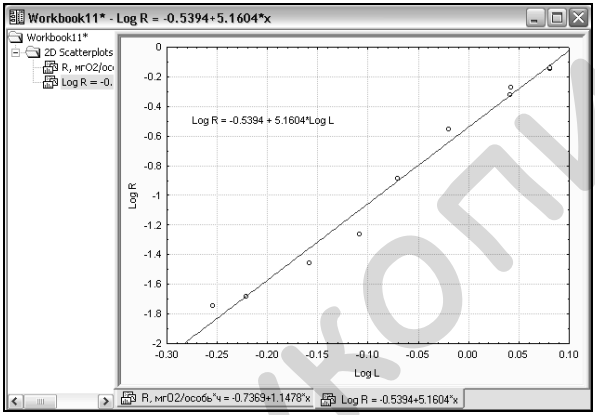


Рисунок 7.11. Визуальная оценка характера связи между двумя признаками после их логарифмирования.

РУКОПИСЬ

Необходимо отметить, что процедура трансформации исходных данных часто применяется не только для «выравнивания» нелинейных связей между признаками. Часто логарифмирование или иные распространенные способы трансформации позволяют привести асимметрично распределенные данные к нормальному распределению, а также добиться однородности дисперсии в группах, подлежащих анализу с использованием параметрических методов. Подробнее о способах трансформации можно узнать в книгах Zar (1999) и Sokal and Rohlf (2002).

7.3. Оценка коэффициентов уравнения нелинейной зависимости

Трансформация данных не всегда позволяет привести зависимость между двумя признаками к линейной форме. Кроме того, научный интерес может представлять уравнение именно нелинейной связи. В программе STATISTICA можно рассчитать уравнение зависимости практически любого функционального вида. Для этого служит специальный модуль *Nonlinear estimations* (Нелинейные оценивания) (*Statistics* > *Advanced Linear/Nonlinear models* > *Nonlinear estimations*, либо кнопка на дополнительной панели инструментов). Применим этот анализ к приведенным выше данным о связи между интенсивностью дыхания дафний и их размерами (рис. 7.7).

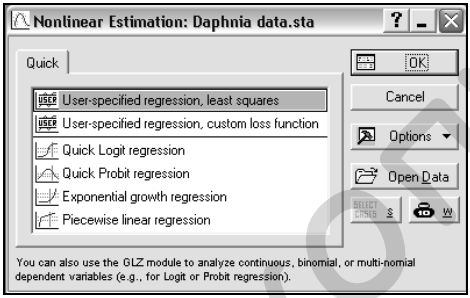


Рисунок 7.12. Окно модуля *Nonlinear estimations*.

Проверим, имеется ли между полученными данными степенная зависимость вида $y = ax^b$. В списке опций модуля *Nonlinear estimations* выберем *User specified regression, least squares* (Заданная пользователем регрессия, метод наименьших квадратов) (рис. 7.12). В появившемся в результате этого окне необходимо нажать кнопку *Function to be estimated* (Оцениваемая функция) и затем ввести формулу *v*2=*a***v*1^*b* (рис. 7.13).

После последовательного нажатия на кнопки *OK* > *OK* откроется диалоговое окно, в котором можно выбрать алгоритм расчета коэффициентов уравнения, получить параметры описательной статистики для каждой из переменных (закладка *Review*), и т.п. Мы оставим все настройки без изменений и нажмем кнопку *OK*. Откроется новое окно (рис. 7.13), в котором можно выполнить анализ остатков (закладка *Residuals*), проследить последовательность того, как программа рассчитывала коэффициенты уравнения (кнопка *Iteration history*), получить параметры описательной статистики, и т.д.

Рукопись

Мы просто нажмем на кнопку *Summary: Parameters & standard errors* (Результат: Параметры и их стандартные ошибки), чтобы увидеть коэффициенты уравнения.

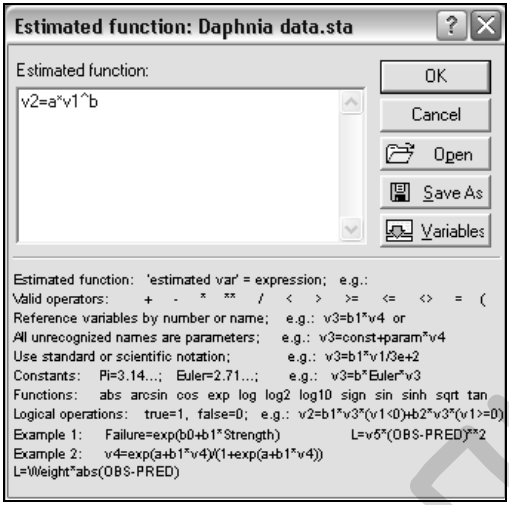


Рисунок 7.12. Окно модуля *Nonlinear estimations*.

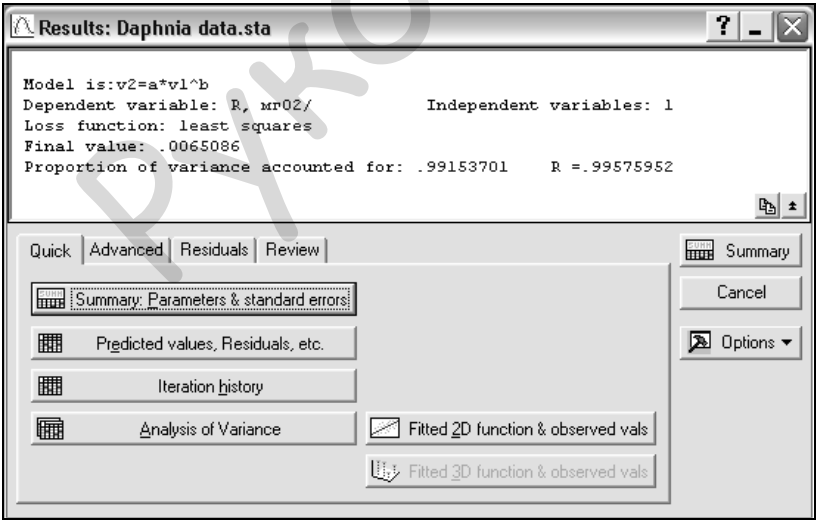
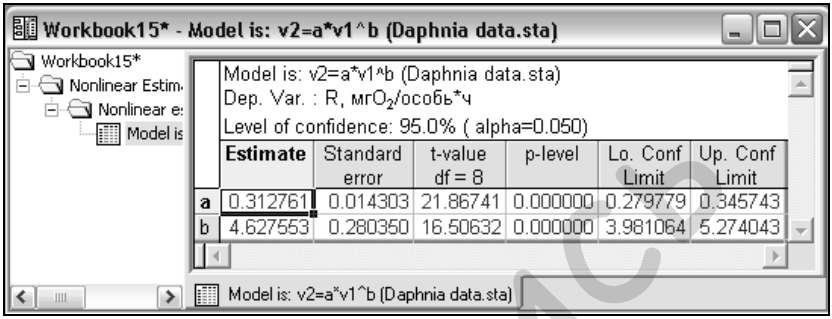


Рисунок 7.13. Окно выбора результатов нелинейного регрессионного анализа.

Рукопись

Как следует из рис. 7.14, оба искомых коэффициента (столбец *Estimate*) оказались высоко значимыми ($P < 0{,}001$; см. столбец *p-level*), т.е. мы не ошиблись с выбором именно степенного уравнения для описания связи между длиной тела дафний и интенсивностью их дыхания. Само же это уравнение можно записать в виде: $R = 0{,}3128 \times L^{4{,}6276}$.



	Estimate	Standard error	t-value df = 8	p-level	Lo. Conf Limit	Up. Conf Limit
a	0.312761	0.014303	21.86741	0.000000	0.279779	0.345743
b	4.627553	0.280350	16.50632	0.000000	3.981064	5.274043

Рисунок 7.14. Результат оценки коэффициентов уравнения степенной зависимости между двумя признаками.

Рукопись

Список использованных источников

Гланц С. Медико-биологическая статистика. Пер. с англ. под ред. Н. Е. Бузикашвили, Д. В. Самойлова – М.: Практика, 1999. – 460 с.

Боровиков В. П. Популярное введение в программу STATISTICA. – М.: Компьютер Пресс, 1998. – 267 с.

Боровиков В. П., Ивченко Г. И. Прогнозирование в системе STATISTICA в среде Windows. Основы теории и интенсивная практика на компьютере: Учебное пособие. – М.: Финансы и статистика, 2000. – 384 с.

Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. – Спб: Питер, 2003. – 688 с.

Реброва О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. – М.: МедиаСфера, 2003. – 312 с.

Леонов В. П. Ошибки статистического анализа биомедицинских данных. Международный журнал медицинской практики. 2007. – №2. – С. 19-13.

Sokal R. R., Rohlf F. J. Biometry: the principles and practice of statistics in biological research. 3rd edition. New York, W. H. Freeman, 2005 – 887 p.

Zar H. H. Biostatistical analysis. 4th edition. New York, Prentice Hall, 1999. – 663 p.

Рукопись

Предметный указатель

Анализ апостериорный 47
Ассоциации коэффициент 60
Гистограмма 15
Данных трансформация 68
Детерминации коэффициент 64
Диаграмма
  диапазонов 21
  размахов 23
  круговая 26
Дисперсионный анализ
  однофакторный 43
  двухфакторный 48
  Крускала-Уоллиса 52
  Фридмана 50
Кендалла коэффициент 51
Колмогорова-Смирнова тест 30
Коэффициент корреляции
  Пирсона 55
  Спирмена 59
Критерий χ^2 30, 62
Левена тест 45
Манна-Уитни тест 36
Нормальных вероятностей график 32
Ньюмена-Кейлса тест 47
Описательной статистики параметры 18
Остатков анализ 66
Переменная
  группирующая 34
  зависимая 34
Распределение нормальное 28
Распределения
  подгонка 28
  полигон 8
Регрессии коэффициент 62
  стандартизованный 64
Регрессия
  линейная 18
  нелинейная 72
Ряд вариационный 10
Свободный член уравнения регрессионного 62
Стьюдента *t*-тест для
  зависимых выборок 38
  независимых выборок 33
  одной выборки 41
Таблица четырехпольная 60
Тьюки тест 47
Уилкоксона тест 39
Фишера *F*-тест 36, 64
Шапиро-Уилка тест 31

РУКОПИСЬ

Сергей Эдуардович Мастицкий, канд. биол. наук

Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований

Редактор **Адамович Б. В.**
Технический редактор Довгалева И. Н.

Отпечатано с макета, предоставленного автором

Издатель
РУП «Институт рыбного хозяйства»
РУП «НПЦ НАН Беларуси по животноводству»
ЛИ № 02330.0133413 от 17.11.2009 г.
220024, г. Минск
ул. Стебенева, д. 22, к. 19 – 20

Тираж ____ экз.

РУКОПИСЬ