ВВЕДЕНИЕ

Актуальной проблемой современной экологии является решение задач, связанных с исследованием и эксплуатацией природных популяций и экосистем в целом, что подразумевает наличие больших массивов информации. Для обработки и анализа этой информации с целью извлечения из данных продуктивных знаний и принятия оптимального решения необходимо умение пользоваться специальными компьютерными системами.

Данный практикум разработан для студентов, обучающихся по специальности 013100 «Экология» и предназначен для ознакомления их на практике с такими методами анализа данных как дескриптивная обработка, регрессионный, корреляционный и дисперсионный анализы, реализованные с помощью пакета статистических программ Statistica. Для каждого вида анализа приводится содержательный пример, разработанный на основе экологических данных.

Краткое описание основных характеристик пакета Statistica

Пакет Statistica представляет собой интегрированную систему статистического анализа и обработки данных в среде Windows и по многочисленным рейтингам является мировым лидером на рынке статистического программного обеспечения. Система производится фирмой StatSoft Inc.(США), основанной в 1984 г. Последняя версия Statistica 6.0 – это высокотехнологичный продукт, позволяющий проводить исчерпывающий, всесторонний анализ данных, представлять результаты анализа в виде таблиц и графиков, автоматически создавать отчеты о проделанной работе, а также благодаря встроенным языкам программирования SCL и Statistica-BASIC автоматизировать рутинные процессы в системе в соответствии с потребностями пользователя. С помощью удобной системы подсказок можно обучаться не только работе с самим пакетом, но и современным методам статистического анализа.

Статистические процедуры системы Statistica сгруппированы в нескольких специализированных статистических модулях, в каждом из которых выполняется определенный способ обработки информации. Основные модули с использованием примеров будут описаны ниже.

Тема 1. РАБОТА С ДАННЫМИ И ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ

Теоретическое введение

Введем основные понятия описательной (дескриптивной) статистики.

Средняя арифметическая – сумма значений варьирующего признака, деленная на их число.

$$\bar{x} = (\sum_{i=1}^{n} x_i) / n .$$

Медианой – это значение признака, который делит ранжированный ряд наблюдений на две равные по объему группы (или значение признака, приходящееся на середину ранжированного ряда наблюдений).

Модой называется такое значение признака, которое наблюдалось наибольшее число раз.

Вариационный размах является показателем вариации, равным разности между экстремальными значениями вариационного ряда, т.е.

$$R_b = x_{\max} - x_{\min} .$$

Вариационный размах показывает широту рассеяния.

Дисперсией называется средняя арифметическая квадратов отклонений от их средней арифметической:

$$s^2 = \left(\sum_x (x - \bar{x})^2 \right) / n .$$

Мера рассеивания должна выражаться в тех же единицах, что и значения признака, поэтому вместо дисперсии в качестве показателя вариации чаще используется корень квадратный из дисперсии.

Арифметическое значение корня квадратного из дисперсии называется *средним квадратическим отклонением*:

$$\sigma = \sqrt{\left(\sum_x (x - \bar{x})^2 \right) / n} .$$

Центральным моментом порядка q называется средняя арифметическая q-х степеней отклонений вариантов от их средней арифметической, т.е.

$$\tilde{\mu}_q = \overline{(x - \bar{x})^q} .$$

Коэффициентом асимметрии называется отношение центрального момента третьего порядка к кубу среднего квадратического отклонения:

$$\tilde{A} = \frac{\tilde{\mu}_3}{s^3} = \frac{\tilde{\mu}_3}{(\sqrt{\tilde{\mu}_2})^3}.$$

Если полигон вариационного ряда скошен, т.е. одна из его ветвей, начиная от вершины, зримо короче другой, то такой ряд называется *асимметричным*. Различают левостороннюю и правостороннюю асимметрию.

Эксцессом или коэффициентом крутости называется уменьшенное на 3 единицы отношение центрального момента четвертого порядка к четвертой степени среднего квадратического отклонения:

$$\tilde{E} = \frac{\tilde{\mu}_4}{s^4} - 3 = \frac{\tilde{\mu}_4}{\tilde{\mu}_2^2} - 3.$$

За стандартное значение эксцесса принимают $\tilde{E}$ =0.

Стандартная ошибка среднего – эта величина, характеризующая стандартное отклонение выборочного среднего, рассчитанное по выборке размера n из генеральной совокупности:

$$\sigma_o = \sqrt{s^2 / n}.$$

Практическая часть

Откройте пакет **Statistica** модуль **Basic Statistics and Tables (Основные статистики и таблицы)**. Перед вами на экране появится окно системы, вид которого аналогичен стандартным программам, работающим в среде Windows. В верхнем левом углу окна высвечивается название запущенного модуля (в данном случае это Basic Statistics and Tables).

Для проведения любого анализа необходимо наличие данных в системе. Напомним, что входные данные в Statistica организованы в виде электронной таблицы Spreadsheet.

1. Создание таблицы данных

Чтобы создать файл данных нужно из меню **File (Файл)** выбрать **New Data (Новые данные)**. В появившемся диалоговом окне **New Data: Specify File Name (Новые данные: Определить имя файла)** введите имя файла: Zagr.sta. Нажмите ОК. Автоматически откроется пустая таб-

лица размером 10×10. Столбцы таблицы называются **Variables (Переменные)** и по умолчанию обозначаются VAR1, VAR2,…, VAR10. Строки именуются **Cases (Случаи).**

Создаваемая таблица будет хранить информацию по выпуску сточных вод на восточном побережье Амурского залива. Размер таблицы – 3×12, т.е. 3 переменные и 12 случаев.

Произведем настройку размеров имеющейся на экране таблицы. Нужно удалить из нее 7 переменных (10–3=7) и добавить 2 случая (12–10=2), тогда получим таблицу, размером 3×12.

Для работы с переменными на панели инструментов нажмите кнопку VARS **Variables (Переменные)**. Из ниспадающего меню выберите команду **Delete (Удалить)**. В появившемся диалоговом окне в поле **From variable (С какой переменной)** укажите VAR4, в поле **To variable (По какую переменную)** – VAR10. Нажмите ОК. С таблицы удалятся переменные с 4-ой по 10-ую.

Теперь нужно добавить два случая. Для этого нажмите кнопку CASES **Cases (Случаи)**, выберите команду **Add (Добавить)**. В диалоговом окне **Add Cases (Добавить случаи)** в поле **Number of Cases to Add (Число добавляемых случаев)** впишите цифру 2, в поле **Insert after Case (После какого случая вставить)** – 10. Нажмите ОК. На экране появится таблица, состоящая из 3-х переменных и 12-ти случаев.

Введем заголовок таблицы. Дважды щелкните левой клавишей мыши на белом поле под словами: Data: Zagr.sta 3v×12c. В появившемся окне **Data File Header (Заголовок файла данных)** можно задать в верхнем поле **One-Line Data File Header** заголовок таблицы, в поле **Data File Information/Notes** – дополнительную информацию о данных. Введем заголовок таблицы: Данные по выпуску сточных вод. Нажмите ОК.

Зададим имена случаям. Нажмите кнопку **Cases (Случаи)**, из ниспадающего меню выберите команду **Names (Имена).** В диалоговом окне **Case Name Manager (Менеджер имен случаев)** задайте ширину столбца: 15; нажмите Yes и введите имена случаев (табл. 1).

Зададим имена переменных. Дважды щелкните на имени переменной VAR1 в электронной таблице. На экране появится окно спецификации переменной **Variable 1**. В поле **Name (Имя)** впишите: расход. В поле **Display Format (Формат отображения)** в опции **Column Width (Ширина столбца)** оставляем 8, в опции **Decimals (Десятичные знаки)** – 0, т.к. эта переменная – целое число.

То же самое проделайте с переменными VAR2 и VAR3: присвойте им имена Zn и Pb соответственно и в поле **Display Format (Формат отображения)** поставьте одну цифру после запятой.

Теперь с помощью клавиатуры введите в таблицу данные (табл. 1). Для сохранения данных из меню **File (Файл)** выберите **Save (Сохранить).**

Таблица 1

NUMERIC VALUES	Расход	Zn	Pb
Де-Фриз	16900	20.0	3.2
Трудовое	2150	16.0	3.2
Фанзавод	2200	16.0	0.0
Черная р.	6000	16.0	0.0
Океанская	1500	14.0	3.2
Седанка	2500	12.0	2.4
ул. Кирова	3500	10.0	0.0
Вторая р.	50000	8.0	3.6
м. Чумака	50000	8.0	4.0
оз. Чан	2215	8.0	6.0
Первая р.	30000	8.0	6.0
б. Спортивная	4000	12.0	2.8

Единицы измерения: расход – $м^3$/сут.; Zn, Pb – мкг/л.

Процедуру создания файла данных можно выполнить другим способом. В пакете Statistica существует модуль **Data Management (Управление данными)**, в котором доступны все операции для работы с данными. Для создания нового файла с данными используется команда **Create new data file**, где устанавливаются все выше перечисленные параметры переменной. Так как файл данных у нас имеется, рассмотрим операцию – сортировка случаев.

Откройте модуль **Data Management (Управление данными)**. Из списка процедур выделите команду **Sort Cases (Сортировка случаев)**. В появившемся диалоговом окне нужно выбрать переменную, по значениям которой будут отсортированы имеющиеся данные. Сортировку можно проводить по возрастанию или по убыванию, по текстовым или по числовым значениям переменной.

Будем проводить процедуру сортировки по возрастанию случаев переменной «расход». В рамке **Key 1 (Ключ 1)** введите переменную «расход» и нажмите ОК. В исходной таблице произойдет сортировка случаев всех переменных.

2. Вычисление описательных статистик

В модуле **Basic Statistics and Tables (Основные статистики и таблицы)** содержатся процедуры для первичной обработки данных,

выяснения их структуры и определения зависимости между данными, а также их группировки.

В стартовой панели модуля выделите строку **Descriptive statistics (Описательные статистики)** и нажмите ОК. В появившемся диалоговом окне нажмите кнопку **Variables (Переменные)** и выберите переменную «расход» для анализа. Щелкните на кнопку **Detailed Descriptive statistics (Детальные описательные статистики)**. На экране появится таблица, в которой вычислены следующие статистики для выбранной переменной: **Valid** – количество случаев, **Mean** – среднее значение, **Minimum** – минимум, **Maximum** – максимум, **Std.dev.** – стандартное отклонение. В пакете Statistica таблицы результатов называются Scrollsheets.

Для задания дополнительных статистик нажмите на кнопку **More statistics (Больше статистик)**. Появится диалоговое окно, в котором можно выбрать описательные статистики: **Sum** – суммарное значение переменной, **Mediana** – медиана, **Variance** – дисперсия, **Std. error of mean** – стандартная ошибка среднего, **95% confidence limits of mean** – 95%-ый доверительный интервал, **Lower & upper quartiles** – нижний и верхний квартили, **Range** – размах, **Skewness** – коэффициент асимметрии, **Kurtosis** – коэффициент эксцесса, **Std. error of skewness** – стандартная ошибка коэффициента асимметрии, **Std. error of kurtosis** – стандартная ошибка эксцесса. Выбрав нужные статистики, щелкните ОК. На экране появится таблица результатов.

Выделите переменные “Zn” и “Pb” в таблице данных Zagr.sta, вычислите описательные статистики для каждой переменной и сравните их.

Тема 2. СТАТИСТИЧЕСКИЕ МОДУЛИ И ОСОБЕННОСТИ ИХ РАБОТЫ

Теоретическое введение

Статистические процедуры по соответствующим разделам статистического анализа сгруппированы в пакете Statistica в нескольких специализированных модулях: Основные статистики и таблицы, Непараметрическая статистика, Множественная регрессия, Нелинейное оценивание, Дисперсионный анализ, Кластерный анализ, Факторный анализ, Анализ временных рядов и прогнозирование и др. В каждом модуле можно выполнить определенный способ статистической обработки, не обращаясь к процедурам из других модулей.

Модули запускаются из **Переключателя модулей (Statistica Module Switcher)**. Чтобы запустить модуль, нужно из меню **Analysis (Анализ)** выбрать **Other Statistics (Другие статистики)**, либо нажать на кнопку **Statistica Module Switcher (Переключатель модулей)** на панели инструментов (рис.1). Запустить можно сразу несколько модулей и переключаться между ними с помощью панели задач Windows внизу экрана монитора.

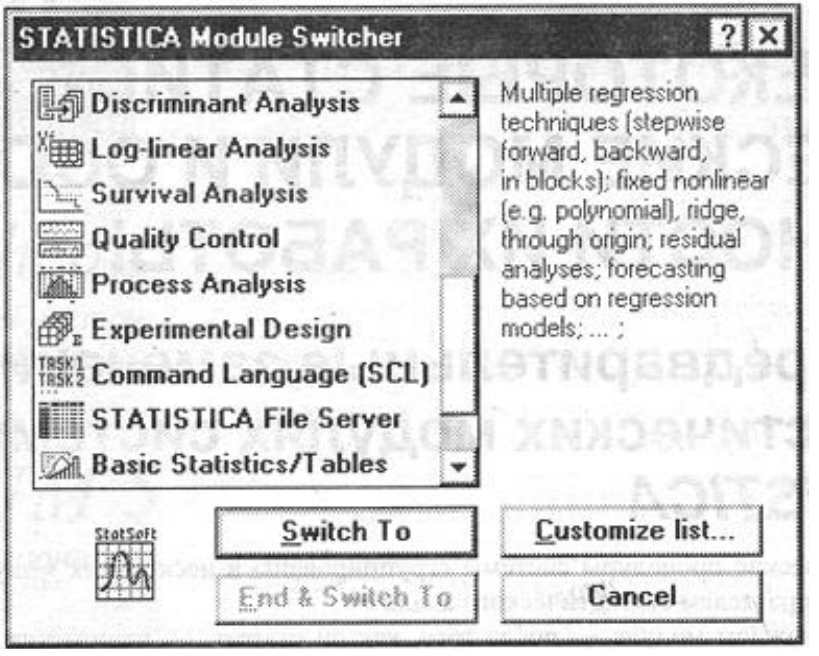


Рис. 1. Окно Переключателя модулей

Для просмотра полного списка установленных модулей в диалоговом окне **Statistica Module Switcher (Переключатель модулей)** нажмите на кнопку **Customizing list (Настройка списка)**. Откроется диалоговое окно **Customize List of Modules (Настройка списка модулей)**, в котором с помощью следующих кнопок можно настроить список модулей в Переключателе модулей:

Append (Добавить) – перемещает выделенные модули в конец прокручиваемого списка в Переключателе модулей;

Replace (Заменить) – замещает текущий список модулей в Переключателе модулей на выделенный;

Add/Remove (Установка/Удаление) – запускает программу Setup для изменения конфигурации системы Statistica.

Всего в системе Statistica имеется 31 модуль для статистической обработки, анализа и графического представления данных. На практических примерах рассмотрим принципы работы следующих модулей: **Basic Statistics/Tables** – Основные статистики/Таблицы; **Multiple Regression** – Множественная регрессия; **Nonlinear Estimation** – Нелинейное оценивание; **ANOVA/MANOVA** – Дисперсионный анализ.

Меню стартовой панели модуля **Statistics/Tables (Основные статистики/Таблицы)** включает следующие методы: **Descriptive statistics** – описательные статистики; **Correlation matrices** –корреляционные матрицы; **t-test for independent samples** – t –критерий для независимых выборок; **t-test for dependent samples** – t –критерий для зависимых выборок; **Breakdown & one-way ANOVA** – классификация и однофакторный дисперсионный анализ; **Frequency tables** – таблицы частот; **Tables and banners** – таблицы и банеры; **Probability calculator** – вероятностный калькулятор; **Other significance test** – другие критерии значимости.

Практическая часть

1. Создание автоотчета

В пакете Statistica имеется возможность сохранять результаты анализа в специальном файле, который называется *отчет*. Отчет, в который автоматически выводятся все результаты анализа (таблицы scrollsheet и графики), называется *автоотчетом*.

Отчет обычно используется для проведения и визуального представления статистического исследования. Он может быть сохранен в файле с расширением *.rtf (по умолчанию) либо в текстовом файле в формате ASCII, а также выведен на принтер.

Для создания автоотчета из меню **File (Файл)** выберите команду **Page/Output Setup (Параметры страницы/вывода)**. В появившемся диалоговом окне для определения параметров вывода числовой и текстовой

информации щелкните на переключателе **Text/Scrollsheets/Spreadsheets (Текст и электронные таблицы)** (рис. 2).

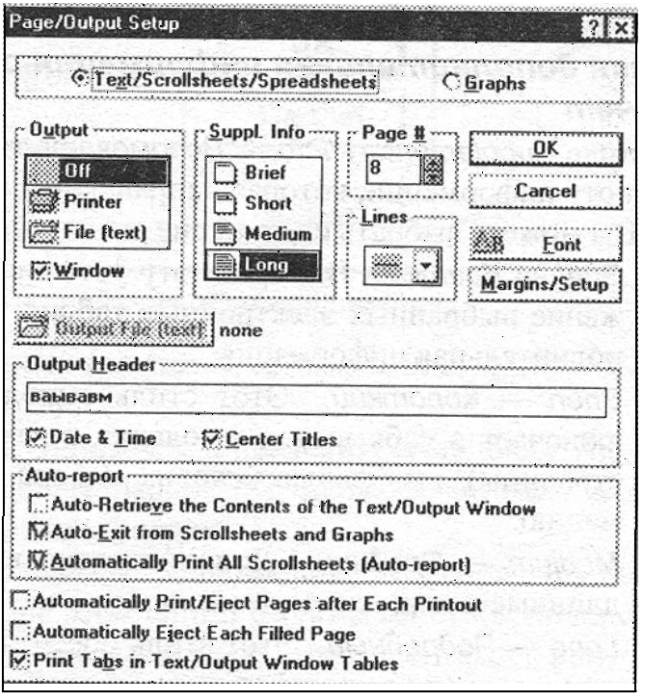


Рис. 2. Диалоговое окно задания параметров вывода таблиц

В рамке **Output (Вывод)** определим варианты вывода информации: **Off (Отключить вывод)** – информация не будет выводится ни на принтер ни в файл; **Printer (Принтер)** – посылается на принтер; **File (Text)** – Файл (текст) – выводится в текстовый файл; **Window (Окно с отчетом)** – отправляется в окно отчета на экран монитора. Выберите опцию **Off (Отключить вывод)** и поставьте галочку возле **Window (Окно с отчетом).**

В рамке **Suppl. Info (Вспомогательная информация)** устанавливают стиль информации: **Brief** – краткий; **Short** – короткий; **Medium** – средний; **Long** – подробный. Выберите стиль – **Short (Короткий),** включающий в себя помимо содержания таблиц имя файла и другую вспомогательную информацию.

В рамке **Output Header (Заголовок вывода)** – можно ввести заголовок; **Date and Time** – дату и время; **Center Titles** – центровать заголовки. Информация будет печататься на каждой новой странице.

В рамке **Auto-report (Автоматический отчет)** имеются следующий установки для автоотчета: **Auto-Retrieve the Contents of the Text/Output Window** – автоматически дополнять содержание Окна текста/вывода; **Auto-Exit from Scrollsheets and Graphs** – автоматический выход из графиков и таблиц Scrollsheets (выделите); **Automatically Print All Scrollsheets (Auto-report)** – автоматическая печать всех таблиц Scrollsheets (автоотчет) (выделите); **Automatically Print/Eject Pages after Each Printout** – автоматическая печать/выдача страницы после каждой операции вывода; **Automatically Eject Each Filled Page** – автоматическая выдача каждой заполненной страницы (выделите); **Print Tabs in Text Output Tables** – печатать символ табуляции при выводе таблиц.

Для задания параметров вывода графиков в диалоговом окне **Page/Output Setup (Параметры страницы/вывода)** выберите переключатель **Graphs (Графики)** (рис. 3). В рамке **Output (Вывод)** выделите **Off (Отключить вывод)** и **Window (Окно с отчетом)** для вывода всей информации в окно с отчет.

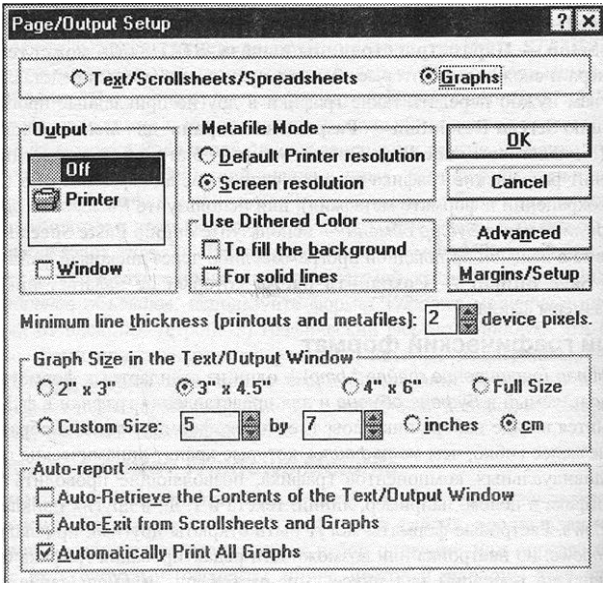


Рис. 3. Диалоговое окно задания параметров вывода графиков

В рамке **Metafile Mode (Режим метафайла)** задают тип разрешения и представления файлов в формате метафайла: **Default Printer Resolution** – разрешение принтера; **Screen Resolution** – разрешение монитора. Выделите вторую опцию, чтобы вся графическая информация

выводилась на экран с разрешением монитора. Для задания дополнительных опций при установки уровня разрешения нажмите кнопку **Advanced Option (Дополнительные опции).**

Если информация выводится на цветной принтер, то в рамке **Use Dithered Color (Использовать текстуру для вывода цветных элементов графики)** выберите одну из опций: **To fill the background** – использовать текстуру при печати подложки; **For solid lines** – использовать текстуру при печати сплошных линий.

В поле **Minimum line thickness (printouts and metafiles)** – минимальная толщина линии при печати и для метафайла, задайте толщину линии на графике, равную 2.

Задать размер графика можно в рамке **Graph Size in the Text/Output Window (Размер графика в окне текста/вывода)**. Выберите «3×4.5 cm».

В рамке **Auto-report (Автоматический отчет)** задаются установки создания автоотчета: **Auto-Retrieve the Contents of the Text/Output Window** – автоматически дополнять содержание окна текста/вывода; **Auto-Exit from Scrollsheets and Graphs** – автоматический выход из графиков и электронных таблиц Scrollsheets; **Automatically Print All Graphs** – автоматическая печать всех графиков. Выберите первую и третью опции.

Сделав нужные установки в открытом диалоговом окне, нажмите ОК. На экране появится пустое окно автоотчета, куда автоматически будет выводиться вся текстовая, числовая и графическая информация. Просмотр и редактирование автоотчета проводят с использованием панели инструментов встроенного текстового редактора, меню которого высвечивается в верхней части окна пакета Statistica.

2. Вычисление корреляционной матрицы и построение графиков

Напомним, что коэффициент корреляции является мерой зависимости рассматриваемых величин и изменяется в пределах от –1 до +1. Чем ближе модуль значения коэффициента к 1, тем сильнее связь между переменными. Значение, превышающее 0.5, говорит о существенности связи между переменными; значение, равное нулю, означает отсутствие корреляции.

Для вычисления корреляционной матрицы создайте новый файл с данными по глубине, температуре и освещенности в Амурском заливе под названием Osvech.sta размером 3×10 (табл. 2).

Заполнив таблицу, сохраните файл данных.

Теперь вычислим и проанализируем коэффициенты корреляции по переменным. Для этого откройте модуль **Basic Statistics/Tables (Oc-**

новные статистики/Таблицы) и выберите строку **Correlation matrices (Корреляционные матрицы)**. Откроется диалоговое окно **Pearson Product-Moment Correlation (Корреляция Пирсона)**. Щелкните на кнопке **One variable list (square matrix) – Один список переменных (квадратная матрица).** В открывшемся окне нажмите на **Select All (Выбрать все)** для выбора всех переменных из таблицы данных. Нажмите дважды ОК. На экране появится таблица корреляции для выбранных переменных.

Таблица 2

ГЛУБИНА	ТЕМПЕР	ОСВЕЩ
5	12,0	90
6	11,5	80
7	11,0	70
8	10,0	67
9	9,5	60
10	9,0	51
12	8,7	40
15	8,5	30
17	8,3	22
20	8,0	17

Единицы измерения: «глубина» – м, «темпер» – °С, «освещ» – кал/см^2.

Рассмотрим полученную таблицу. Коэффициенты корреляции между переменными имеют следующие значения: между переменными «глубина» и «темпер» – (-0.90), между «глубина» и «освещ» – (-0.98), между «темпер» и «освещ» – 0.96. Получаем, что между рассматриваемыми переменными связь существенна, т.к. коэффициенты корреляции по модулю близки к 1. В двух случаях коэффициенты отрицательны, это указывает на существование обратной связи между переменными, т.е. значение одной переменной увеличивается с уменьшением другой (в нашем случае, с увеличением глубины температура и освещенность падают).

Заметим, что в системе Statistica автоматически красным цветом выделяются значимые коэффициенты на уровне значимости $p<0.05$.

Определим графически зависимость между переменными. В диалоговом окне **Pearson Product-Moment Correlation (Корреляция Пирсона)** нажмите кнопку **2D scatterplot (2D диаграмма рассеяния)**.

В следующем окне выберите переменные для построения диаграммы рассеяния: в первом поле выделите переменную «глубина», во втором – «темпер»; нажмите ОК. На экране появится график, представляющий собой диаграмму рассеяния для переменных «глубина» и «темпер» с подгонкой линейной регрессией и 95%-ой доверительной полосой (рис. 4). В верхней части окна графика записывается уравнение прямой и значение коэффициента корреляции (оно с некоторой погрешностью совпадает со значением из таблицы корреляции).

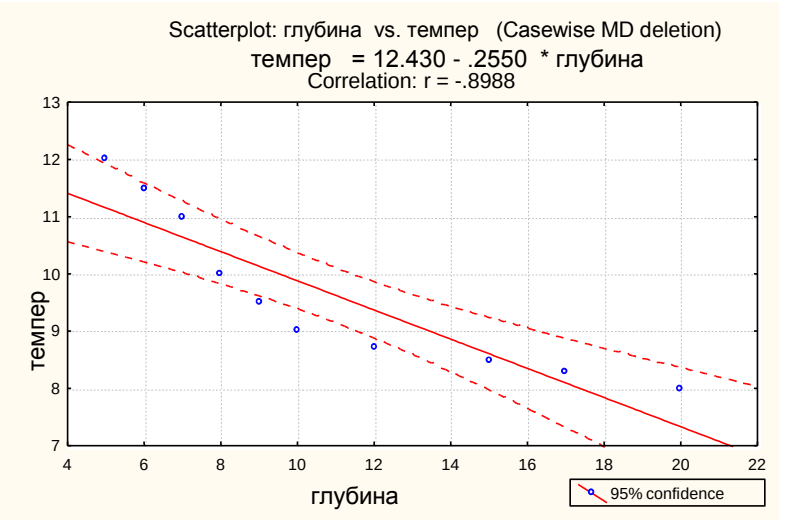


Рис. 4. Диаграмма рассеяния для переменных «глубина» и «темпер»

Из графика видно, что прямая линейной регрессии не очень хорошо «ложится» на данные, и для определения вида зависимости нужно углубить исследования.

Постройте самостоятельно и проанализируйте диаграммы рассеяния для переменных «глубина» и «освещ», «темпер» и «освещ».

Для построения трехмерного графика в диалоговом окне **Pearson Product-Moment Correlation (Корреляция Пирсона)** нажмите кнопку **3D scatterplot (3D диаграмма рассеяния)**. Выберите переменные: в первом поле – «глубина», во втором поле – «темпер», в третьем – «освещ». Нажмите ОК. На экране появится график, показывающий, что с увеличением глубины температура и освещенность падают (рис. 5).

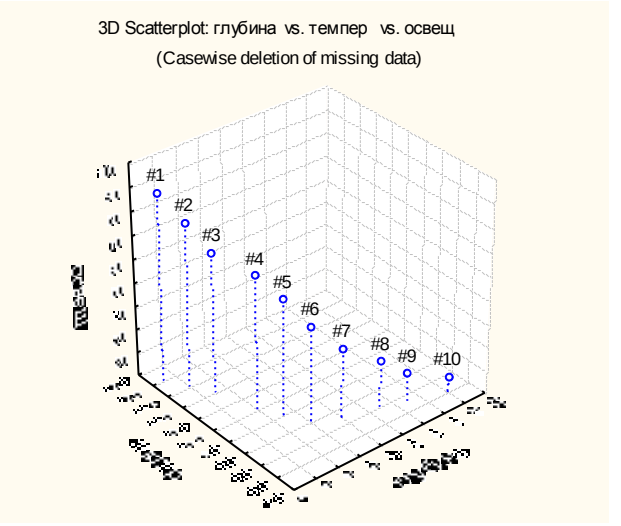


Рис. 5. Трехмерная диаграмма рассеяния

3. Построение таблицы частот

Таблицы частот или одновходовые таблицы представляют собой простейший метод анализа категориальных (номинальных) переменных. Часто их используют как одну из процедур разведочного анализа, чтобы просмотреть, каким образом различные группы данных распределены в выборке.

В окне модуля **Basic Statistics/Tables (Основные статистики/Таблицы)** выберите строку **Freguency tables (Таблицы частот)**. В появившемся диалоговом окне нажмите кнопку **Variables (Переменные)** и выберите все переменные с помощью кнопки **Select All (Выбрать все)**. Нажмите ОК. Щелкните на кнопке **Freguency tables (Таблицы частот)**. На экране последовательно появятся три таблицы частот для каждой выбранной переменной. В первом столбце **Category (Группы)** каждой таблицы значения переменной разбиты на 10 равных интервалов (в нашем случае переменные имеют всего 10 значений, которые и выводятся в таблицу). Во втором столбце **Count (Счет)** показано количество значений переменной, попавших в соответствующий интервал; в третьем столбце **Cumulative Count (Совокупный Счет)** даны накопленные частоты; в четвертом столбце **Percent (Процент)** – выраженные в процентах частоты; в последнем столбце **Cumulative Percent (Совокупный Процент)** – процент накопленных частот от общего числа.

Если вам нужно определенным образом сгруппировать переменные по их значениям, в диалоговом окне **Freguency tables (Таблицы частот) в поле Categorization method for tables & graphs (Категоризированный метод для таблиц и графиков)** выберите опцию **User-specified categories (Группы пользователя)** и нажмите на кнопку. В окне **Define Categories (Определить группы)** определите на какие группы нужно разбить значения выделенной переменной.

Например, разобьем значения переменной «освещ» на три группы. Для этого в окне **Define Categories (Определить группы)** в поле **Category 1 (Группа 1)** запишите неравенство: освещ<30; в поле **Category 2 (Группа 2)** – освещ<=60; в поле **Category 3 (Группа 3)** – освещ<=90. Нажмите два раза ОК. На экране появится таблица **Frequency Table for User-Defined Categories (Таблица частот определяемых пользователем групп)**. В первом столбце таблицы даны три группы, т.е. область значений переменной «освещ», разбитая на три интервала. Во втором столбце показано количество значений переменной, попавших в соответствующий интервал (2; 4; 4); в третьем столбце – накопленные (просуммированные) частоты (2; 6; 10); в четвертом столбце – частоты, выраженные в процентах (20; 40; 40); в последнем столбце – процент накопленных частот от общего числа (20; 60; 100).

4. Вероятностный калькулятор

Вероятностный калькулятор помогает решать разнообразные вероятностные задачи, заменяя таблицы распределений.

В стартовом окне модуля **Basic Statistics/Tables (Основные статистики/Таблицы)** высветите строку **Probability calculator (Вероятностный калькулятор)**. Перед вами появится окно **Probability Distribution Calculator (Калькулятор вероятностных распределений)**. В левой части окна в поле **Distribution (Распределение)** имеется список стандартных вероятностных распределений: Бета, Коши, хи-квадрат, экспоненциальное, нормальное, логнормальное и др. (всего 15 распределений). Для выбора фиксированной шкалы ниже списка указывается опция **Fixed Scaling (Фиксированная шкала)**. В центре окна задаются параметры выбранного распределения, например, для хи-квадрат в строке **df** указывают число степеней свободы, в строке **p** – вероятность. После нажатия кнопки **Compute (Вычислить)** в строке **Chi I** появится квантиль хи-квадрат распределения с указанными степенями свободы, а в нижней части окна соответствующие графики плотности и функции распределения. Можно сделать наоборот, в зависимости от имеющихся данных – по заданному значению Chi I вычислить вероятность p.

В верхней части окна имеются следующие опции: **Inverse (Обратная функция распределения), Two-tailed (Двухсторонний), 1-Cumulative (1-p), Print (Печать), Create graph (Создать график).**

Перед использованием вероятностного калькулятора рассмотрим вкратце основные вероятностные распределения.

Нормальное распределение – наиболее часто встречающийся вид распределения. С ним приходится сталкиваться при анализе и прогнозировании различных явлений в экологии, биологии, медицине и других областях знаний. Главная особенность нормального закона состоит в том, что он является предельным законом, к которому приближаются другие законы распределения.

Нормальному закону распределения подчиняются только непрерывные случайные величины. Поэтому распределение нормальной совокупности может быть задано в виде плотности распределения:

$$f(x) = 1/[(2*\pi)^{1/2}*\sigma] * e^{**}\{-1/2*[(x-\mu)/\sigma]^2\},$$

$$-\infty < x < +\infty,$$

где $-\infty < \mu < +\infty$ – среднее;

$\sigma > 0$ – стандартное отклонение;

e – число Эйлера (2.71...);

π – число Пи (3.14...).

Нормальный закон распределения обозначается $N(\mu,\sigma^2)$. Преобразованная величина z, определяемая соотношением $z = (x-\mu)/\sigma$, обладающая средним $\mu=0$ и стандартным отклонением $\sigma=1$, имеет распределение N(0,1), называемое стандартным нормальным распределением (рис. 6).

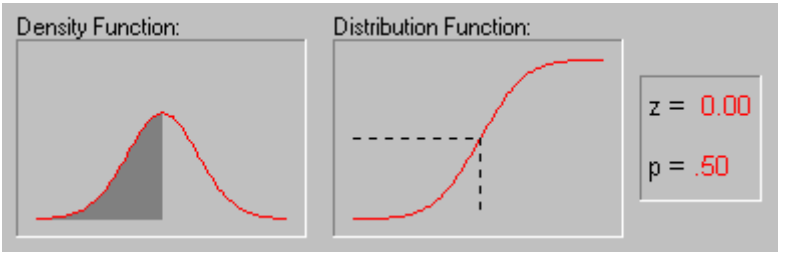


Рис. 6. График плотности и функции нормального распределения

Пример. Известно, что концентрация цинка в заливе Угловой приближенно имеет нормальное распределение со средним 18 мкг/л и стандартным отклонением 6 мкг/л. Произвольным образом берем пробы воды. Какова вероятность того, что концентрация цинка в пробе будет больше 12 и меньше 24 мкг/л.

Решение. Откройте вероятностный калькулятор (**Probability calculator**). В поле **Distribution (Распределение)** из списка распределений

выберите **Z(Normal)** – нормальное распределение. В поле **mean (среднее)** задайте 18; в поле **st.dev.(стандартное отклонение)** – 6; в поле **X(квантиль)** – 24. Нажмите кнопку **Compute (Вычислить)**. В поле **p(вероятность)** появится значение, равное p_1=0.841345. Затем, в поле **X(квантиль)** введите число 12 и нажмите **Compute (Вычислить).** Появится новое значение вероятности: p_2=0.158655. На следующем шаге вычисляем: p_1-p_2=0.841345-0.158655=0.68269. Получаем, что с вероятностью 0.68269 во взятой нами пробе концентрация цинка будет находиться в пределах от12 до 24 мкг/л.

Распределение хи-квадрат (χ^2). Случайная величина X, определяемая как сумма квадратов v независимых стандартных нормальных величин, обладает распределением хи-квадрат с параметром v и плотность ее распределения имеет следующий вид:

$$f(x) = \{1/[2^{V/2} * \Gamma(v/2)]\} * [x^{(v/2)-1} * e^{-x/2}],$$

$$v= 1, 2,...; x >0,$$

где v – число степеней свободы;
e – число Эйлера (2.71...);
Γ – гамма-функция.

Закон распределения хи-квадрат обозначается χ^2 (v) и используется при исследовании оценки дисперсии нормальной выборки, при проверке зависимостей в таблицах сопряженности и в критериях согласия (рис. 7). При достаточно больших значениях v распределение χ^2 переходит в нормальное.

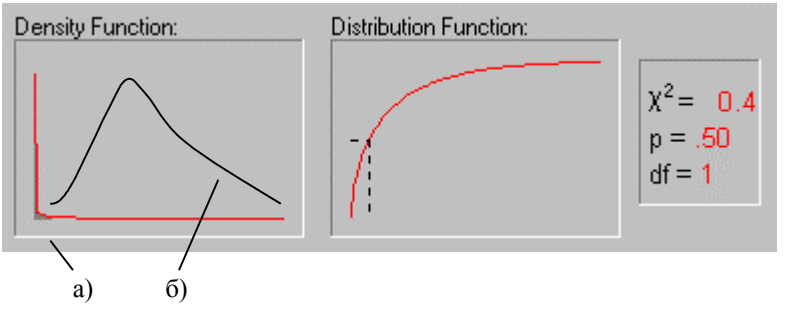


Рис. 7. Изменение формы плотности распределения хи-квадрат при увеличении числа степеней свободы: а) df=1; б) df= 25

В списке распределений вероятностного калькулятора выберите **Chi I (распределение хи-квадрат)**. В строке **df (число степеней свободы)** введите 10, а в поле **p(вероятность)** – 0.95. Нажмите кнопку **Com-**

pute (Вычислить), и в строке **Chi I (хи-квадрат)** появится значение 18.307038, являющееся 0.95-квантилем хи-квадрат распределения с 10 степенями свободы.

Для вывода графика плотности и функции распределения выберите в окне **Probability Distribution Calculator (Калькулятор вероятностных распределений)** опцию **Create graph (Создать график)** и нажмите **Compute (Вычислить).**

t-распределение Стьюдента. Если случайная величина Z имеет распределение N(0,1), а U – распределение $\chi^2(v)$ и величины Z и U независимы, то случайная величина X, определяемая неравенством $X = \frac{Z}{\sqrt{U/Z}}$, имеет распределение Стьюдента с параметром v. Функция плотности распределение имеет следующий вид:

$$f(x) = \Gamma[(v+1)/2] / \Gamma(v/2) * (v*\pi)^{-1/2} * [1 + (x^2/v)^{-(v+1)/2}],$$

где v – число степеней свободы;

Γ – гамма-функция;

π – число Пи (3.1415...).

Закон распределения Стьюдента обозначается t(v) и применяется в регрессионном анализе и анализе временных рядов (рис. 8).

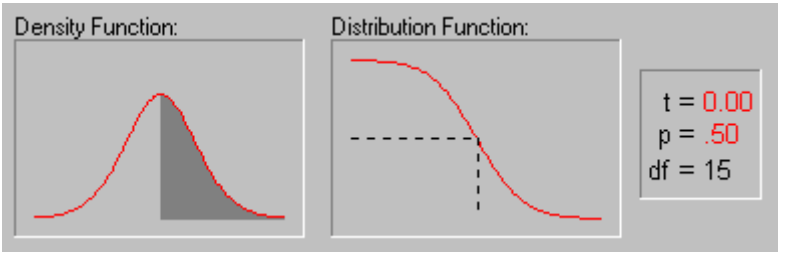


Рис. 8. График плотности и функции распределения Стьюдента при 15 степенях свободы

При $v \to \infty$ t-распределения быстрее, чем χ^2 стремится к нормальному распределению.

В поле **Distribution (Распределение)** вероятностного калькулятора выберите **t (Student) – t-распределение Стьюдента**. В строке **df (число степеней свободы)** введите 7, а в поле **p (вероятность)** – 0.6. Нажмите кнопку **Compute (Вычислить)**, и в строке **t (t-распределение)** появится значение 0.26316. Затем, увеличивая число степеней свободы, обратите внимание на график плотности распределения. При значениях больших

30 распределение Стьюдента практически совпадает со стандартным нормальным распределением.

F-распределение Фишера. Если случайная величина V имеет распределение χ^2 (v), а W – распределение χ^2 (w) и эти величины независимы, то случайная величина $X = \frac{V/v}{W/w}$ обладает F-распределением с параметрами v и w. Параметры v и w называются числами степеней свободы числителя и знаменателя соответственно. F-распределение Фишера (для x > 0) имеет следующую функцию плотности (для v= 1, 2,...; w= 1, 2,...):

$$f(x) = \{\Gamma[(v+w)/2]\}/[\Gamma(v/2) * \Gamma(w/2)] * (v/w)^{v/2} * x^{(v/2)-1} * \{1+[(v/w)*x]\}^{-(v+w)/2}$$

$$0 \le x < \infty .$$

Закон F-распределения величины X обозначается F(v,w) и возникает в регрессионном, дисперсионном и во многих других многомерных анализах данных (рис. 9).

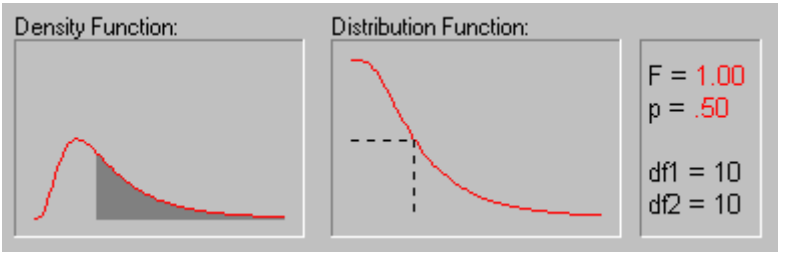


Рис. 9. График плотности и функции распределения Фишера

Использование вероятностного калькулятора для подсчета F-распределения и построения графиков аналогично предыдущим случаям.

5. t-критерий для зависимых выборок

В пакете Statistica имеются два t-критерия для зависимых и независимых выборок, которые позволяют сравнивать средние в двух группах. В зависимости от того, насколько различны значения между групповыми средними, определяют более сильную или более слабую степень зависимости между рассматриваемыми переменными.

Рассмотрим применение t-критерия для зависимых выборок при решении следующей задачи.

Пример. При проверке работы одного из предприятий представители комитета по охране окружающей среды обнаружили следующие нарушения: при взятии проб концентрация вредных веществ в отходах предприятия, прошедших определенную очистку, превышает ПДК. Сток отходов, поступая в прибрежную зону моря, загрязнет ее. На предприятие был наложен штраф при условии, что в течение следующего месяца работа очистных сооружений будет улучшена.

Через месяц повторно были взяты пробы. Результаты представлены в виде таблицы. Улучшилась ли работа очистных сооружений после принятия административных мер?

Решение. Создайте файл данных под названием Proverka.sta. Введите данные и сохраните файл (табл. 3).

Таблица 3

DO	POSLE
105.842	90.744
89.655	82.497
98.710	71.305
110.969	95.532
91.850	86.715
112.346	77.659
107.175	99.510
103.978	92.047
91.172	87.978
105.694	79.960
100.822	100.777
95.506	83.219
111.507	67.832
80.734	87.831
104.223	97.105

Предполагается, что результаты проверки до и после принятых административных мер (переменные DO и POSLE) зависимы. Для подтверждения этого воспользуемся t-критерием.

Из модуля **Basic Statistics/Tables (Основные статистики/Таблицы)** выберите строку **t-test for dependent samples (t-критерий для зависимых выборок)**. В появившемся диалоговом окне нажмите на кнопку **Variables (Переменные)**. В первом столбце окна **Select one or**

two variables list (Выбрать один или два списка переменных) выделите переменную DO, во втором столбце – POSLE. Нажмите OK.

Для визуализации данных нажмите на кнопку **Box & whisker plots (Графики «ящики с усами»)**. Появится диалоговое окно для определения типа графика. Выберите **Mean/SE/SD (Среднее/Стандартная ошибка/Стандартное отклонение)**, нажмите OK. График «ящик с усами» высветится на экране монитора (рис. 10).

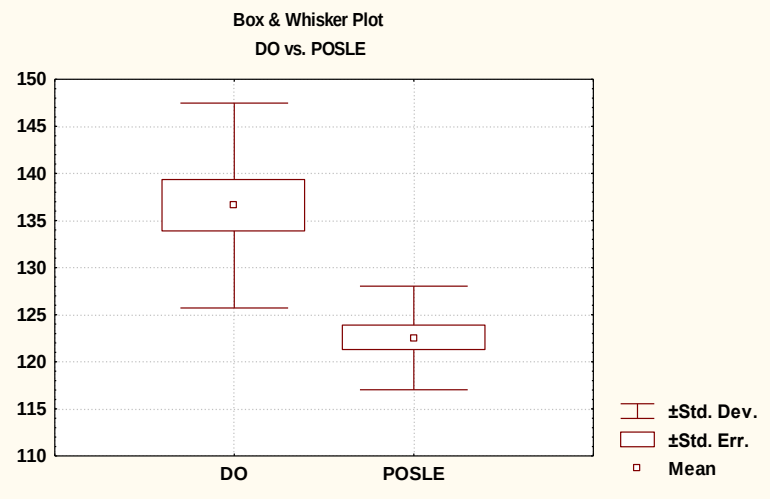


Рис. 10. График «ящик с усами» для переменных DO и POSLE

Рассмотрим график. Левый «ящик с усами» соответствует переменной DO, правый – переменной POSLE. В нижнем правом углу графика имеется пояснение: точки в центре ящиков соответствуют средним значениям переменных (Mean); прямоугольники – стандартным ошибкам средних (± Std.Err.); линии – стандартным отклонениям от средних (± Std.Dev.). Получаем, что среднее значение концентрации загрязняющих веществ в отходах после административного взыскания значительно уменьшилась, интервалы средних ошибок не пересекаются, т.е. антропогенная нагрузка на прибрежную зону моря уменьшилась.

Перейдем к численным значениям. В поле **Display (Показать)** диалогового окна **t-test for dependent samples (t-критерий для зависимых выборок)** выделите опцию Detailed table of results (Подробная таблица результатов) и нажмите кнопку **T-test (T-критерий)**. Появится таблица, в столбцах которой даны следующие результаты: средние значения переменных, стандартные отклонения, число наблюдений, разность между

средними значениями переменных, значение статистики t-критерия, число степеней свободы и уровень значимости (табл. 4).

Таблица 4

T-test for Dependent Samples (Proverka.sta)								
BASIC STATS Variable	Marked differences are significant at p <.05000							
	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
DO	136.5933	10.87768						
POSLE	122.5333	5.50151	15	14.06001	11.09589	4.907600	14	.000231

При отсутствии табличных значений статистики t-критерия, обратим внимание на уровень значимости. Так как $p=0.000231<0.05$, следовательно различие в средних значениях концентраций загрязняющих веществ в пробах до и после административного взыскания высокозначимо. Можно с уверенностью заключить, что работа очистных сооружений значительно улучшилась.

Тема 3. РЕГРЕССИОННЫЙ АНАЛИЗ

Теоретическое введение

Регрессионный анализ предназначен для изучения связи между одной зависимой переменной и несколькими (или одной) независимыми переменными. Эта связь представляется с помощью математической модели, т.е. уравнения, которое связывает зависимую переменную с независимой с учетом множества соответствующих предположений. Если функция линейна относительно независимых параметров, то говорят о линейной модели регрессии. В противном случае модель называется нелинейной.

Рассмотрим линейную зависимость Y от X, которая имеет следующий вид:

$$y_i = \beta_1 * x_i + \beta_0 + e_i,$$

$$0<i<=n,$$

где β_1, β_0 – коэффициенты регрессии (определяются по методу наименьших квадратов);

e_i – ошибки наблюдений (предполагается, что ошибки имеют нормальное распределение).

Задача регрессионного анализа состоит в том, чтобы по имеющимся наблюдениям (x_1,y_1), (x_2,y_2),…, (x_n,y_n) оценить наилучшим образом параметры модели β_1 и β_0, построить для них доверительные интервалы, проверить гипотезу о значимости регрессии и оценить степень адекватности модели.

Иногда, при проведении анализа линейной модели, исследователь получает данные о ее неадекватности. В этом случае, его по-прежнему интересует зависимость между переменными, но для уточнения модели в уравнение добавляются некоторые нелинейные члены. Самым удобным способом оценивания параметров полученной регрессии является нелинейное оценивание.

В пакете Statistica имеется модуль **Нелинейное оценивание (Nonlinear Estimation)**, в котором собраны процедуры, позволяющие оценить нелинейные зависимости между данными. Стартовая панель модуля содержит следующие методы обработки: логистическая регрессия, пробит-регрессия, кусочно-линейная регрессия, регрессия экспоненциального роста и определяемая пользователем регрессия.

Практическая часть

1. Линейная регрессия

Используя данные табл. 2, установим, существует ли линейная зависимость между освещенностью и глубиной. Для этого, из **Переключателя модулей (Statistica Module Switcher)** выберите модуль **Multiple Regression (Множественная регрессия)**. В появившемся диалоговом окне нажмите кнопку **Open Data (Открыть данные)** и откройте файл данных Osvech.sta. Затем для выбора переменных нажмите **Variables (Переменные)**. В окне **Select dependent and independent variable list (Выбрать списки зависимых и независимых переменных)** в левом поле выберите зависимую переменную – «освещ», в правом поле независимую переменную – «глубина». Дважды нажмите ОК. Программа произведет необходимые вычисления и откроет следующее диалоговое окно: **Multiple Regression Results (Результаты множественной регрессии**). Рассмотрим его подробно.

Окно результатов состоит из двух частей: информационной и функциональной. Информационная часть содержит краткую информацию о результатах анализа.

Dep. Var. – имя зависимой переменной (освещ.).

No. of cases – число случаев, по которым построена регрессия. В нашем примере – 10.

Multiple R – коэффициент множественной корреляции. Измеряет степень линейных связей между переменными. В примере коэффициент R=0.97595970 близок к единице, что указывает на сильную линейную зависимость между выбранными переменными.

R^2 – квадрат коэффициента множественной корреляции (или коэффициент детерминации). В нашем случае R^2 =0.95249733 – хорошее значение, показывающее, что построенная регрессия объясняет более 95% разброса значений зависимой переменной относительно среднего.

Adjusted R^2 – скорректированный коэффициент детерминации (Adj. R^2=0.94655949).

Std. Error of estimate – стандартная ошибка оценки, являющаяся мерой рассеяния наблюдаемых значений относительно регрессионной прямой (Std. Error of estimate=5.761344098).

Intercept – оценка свободного члена регрессии (значение коэффициента β_0 в уравнении регрессии). Intercept =105.73632726.

Std. Error – стандартная ошибка оценки свободного члена (стандартная ошибка коэффициента). Std. Error=4.566676.

t(df) and p-value – значение t-критерия и уровень значимости p (используются для проверки гипотезы о равенстве нулю свободного члена уравнения регрессии). Так как в нашем примере $p<0.05$ ($t(8)=23.154$, $p<0.000$), то гипотезу отвергаем и, следовательно, $\beta_0 \neq 0$.

F – значение F-критерия (используется для проверки гипотезы о значимости регрессии).

df – число степеней свободы F-критерия.

p – уровень значимости.

Рассмотрим результаты F-критерия. Проверяем гипотезу, согласно которой между зависимой переменной «освещ» и независимой переменной «глубина» нет линейной связи, т.е. $\beta_1=0$. В примере значение F-критерия равно 160.4116 и уровень значимости p=0.000001. Для того, чтобы принять или отвергнуть гипотезу, нужно полученное значение F-критерия сравнить с табличным (если полученное значение больше табличного, то гипотезу отвергаем) или сравнить полученное p-значение со стандартным уровнем значимости α (обычно он равен либо 0.01 либо 0.05). В нашем случае при α=0.01 табличное значение F-критерия, равное 11.26, намного меньше полученного, и $p<\alpha$, следовательно, можно говорить о высокой значимости построенной регрессии.

Теперь перейдем в функциональную часть окна. Щелкните на кнопку **Regression summary (Итоговый результат регрессии)**. На экране появится таблица с результатами регрессионного анализа (табл. 5).

Таблица 5

Regression Summary for Dependent Variable: освещ						
R=.97595970 RI=.95249733 Adjusted RI=.94655949 F(1,8)=160.41 p<.00000 Std.Error of estimate: 5.7613						
MULTIPLE REGRESS. N=10	BETA	St. Err. of BETA	B	St. Err. of B	t(8)	p-level
Intercpt			105.7363	4.566676	23.15389	1.29E-08
глубина	-0.97596	0.077057	-4.86572	0.384175	-12.6654	1.42E-06

В первом и во втором столбцах таблицы (BETA и St. Err. of BETA) даны значения стандартизованного коэффициента регрессионного уравнения и его стандартная ошибка; в третьем столбце – оценки неизвестных параметров модели: свободный член β_0=105.7363 и коэффициент при независимой переменной β_1= -4.86572. В следующих столбцах имеются стандартные ошибки для β_0 и β_1, значения статистик t-критерия и уровень значимости.

На основе табличных данных можно построить искомую модель зависимости освещенности от глубины:

освещ = 105.7363 – 4.86572 *глубина.

Оценим адекватность полученной модели с помощью анализа остатков. Напомним, что под адекватностью модели простой линейной регрессии подразумевают, что никакая другая модель не даст значимого улучшения в предсказании значений зависимой переменной.

Остатками называется разность между исходными (наблюдаемыми) значениями и предсказанными (модельными), то есть значениями, предсказанными с помощью модели.

Для всестороннего анализа остатков в окне **Multiple Regression Results (Результаты множественной регрессии)** в нижнем правом углу нажмите кнопку **Residual Analysis (Анализ остатков)**. В появившемся диалоговом окне с помощью функциональных кнопок можно посмотреть остатки модели, как в графическом виде, так и в виде электронных таблиц.

Вначале воспользуемся визуальными методами. Для этого построим график остатков на вероятностной бумаге. Из правого блока **Probability Plots (Вероятностные графики)** инициируйте кнопку **Normal plot of resides (M) (График остатков на нормальной вероятностной бумаге)**. На экране появится график, из которого видно, что остатки довольно хорошо ложатся на прямую, соответствующую нормальному закону. Следовательно, предположение о нормальном распределении ошибок выполнено (рис. 11).

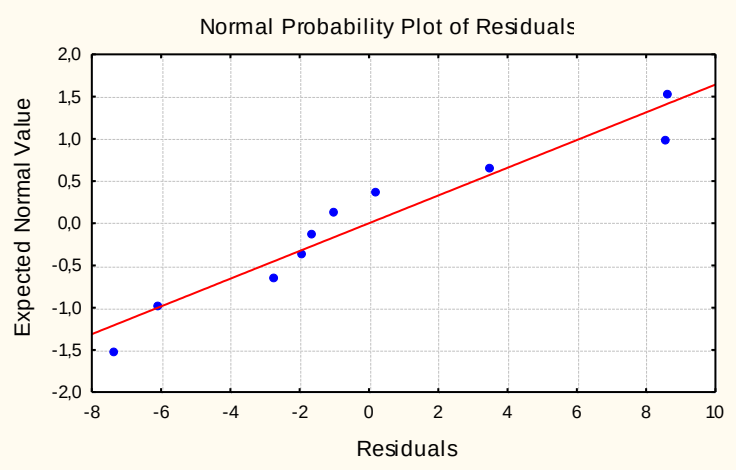
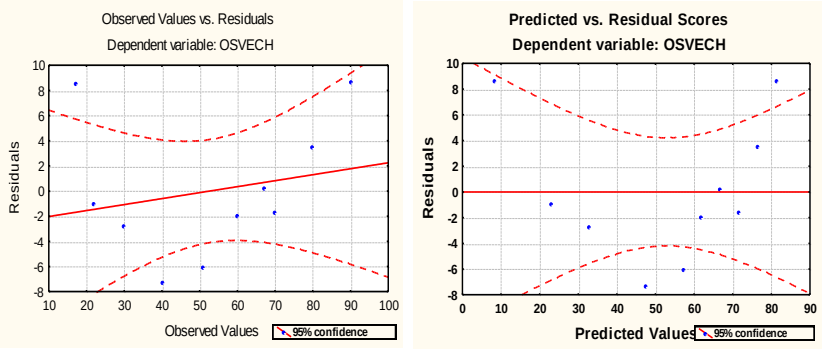


Рис. 11. График остатков на нормальной вероятностной бумаге

Из графиков предсказанные значения-остатки (кнопка **Pred.&residuals (D)**) и обследуемые значения-остатки (**Obs.&residuals (G)**) можно заключить, что модель достаточно адекватно описывает данные, так как остатки расположены хаотично относительно прямой. В их поведении нет закономерностей и нет резко выделяющихся остатков (рис. 12).



а) б)

Рис. 12. Графики остатков в линейной модели:
а) обследуемые значения-остатки; б) предсказанные значения-остатки

Таким образом, построенная линейная модель является приемлемой для описания линейной зависимости освещенности от глубины моря.

Самостоятельно с помощью регрессионного анализа проверьте следующую гипотезу: существует ли линейная зависимость между глубиной моря и температурой.

2. Нелинейная регрессия

Регрессия экспоненциального типа

Предполагаем, что между переменными Y и X существует функциональная зависимость экспоненциального типа, описываемая следующим уравнением:

$$Y = c + \exp(b_0 + b_1 * X),$$

где c, b_0, b_1 – неизвестные параметры, которые нужно оценить.

Для начала создайте таблицу данных значений переменных Y и X под названием ExpReg.sta (табл. 6).

Таблица 6

Y	X
7,476	0,100
8,220	0,200
9,182	0,300
10,177	0,400
11,852	0,500
13,878	0,600
15,835	0,700
18,635	0,800
21,974	0,900
26,140	1,000
30,772	1,100
36,679	1,200
44,000	1,300
53,075	1,400
64,107	1,500

Сохраните таблицу.

В модуле **Nonlinear Estimation (Нелинейное оценивание)** дважды щелкните левой клавишей мыши на строку **Exponential growth regression (Регрессия экспоненциального типа)**. В появившемся диалоговом окне нажмите на кнопку **Variables (Переменные)** и выберите следующим образом переменные для анализа: Y – зависимая переменная (Dependent), X – независимая переменная (Independent). Нажмите дважды ОК. Перед вами откроется окно **Model Estimation (Оценивание модели)** для выбора процедуры оценивания и начальных установок. В строке **Estimation method (Метод оценивания)** установите метод **Розенброка (Rosenbrock patten search)**, нажмите ОК. На экране появится окно **Parameter Estimation (Оценивание параметров)** с прокручиваемыми результатами оценивания параметров модели на каждой итерации. В нижней строке имеется сообщение: **Parameter estimation process converged – Процесс оценки параметров сошелся**. В первом столбце даны номера итераций (у нас за 16 итераций метод сошелся); во втором столбце даны значения функции потерь на каждой итерации (последнее значение мало – 0.212214), в следующих трех столбцах показаны оцен-

ки параметров: c, b_0, b_1. Нажмите ОК. Откроется окно, в котором можно детально рассмотреть полученные результаты.

В функциональной части окна нажмите на кнопку **Parameter estimates (Параметры оценивания).** На основе результатов оценивания можно записать полученное уравнение модели:

$$Y=3.70420+\exp(1.097835+2.000715*X).$$

Теперь нужно определить, насколько данная модель подходит к данным. Щелкните мышкой на кнопке **Normal Probability Plot of Residuals (График остатков на нормальной вероятностной бумаге)**. Появится график, из которого можно заключить, что остатки достаточно хорошо ложатся на прямую, т.е. модель неплохо описывает наши данные (рис. 13).

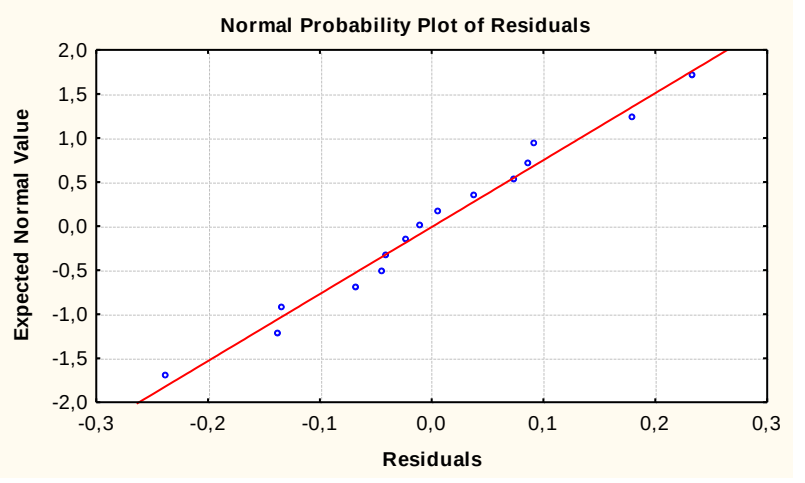


Рис. 13. График остатков на нормальной вероятностной бумаге

Далее в окне результатов нажмите на кнопку **Fitted 2D function & observed values (Подогнанная функция и наблюдаемые значения)**. Появится график, на котором подогнанная функция полностью легла на исходные данные (рис. 14).

Рассмотрим предсказанные значения и остатки. Нажмите на кнопки **Predicted values (Предсказанные значения)** и **Residual values (Значения остатков)**. Появятся две таблицы. Сравните предсказанные значения с исходными данными. Разница не существенная; значения остатков не велики. Следовательно, можно сделать вывод: между переменными Y и X существует функциональная зависимость экспонен-

циального типа, и построенная модель достаточно хорошо описывает эту связь.

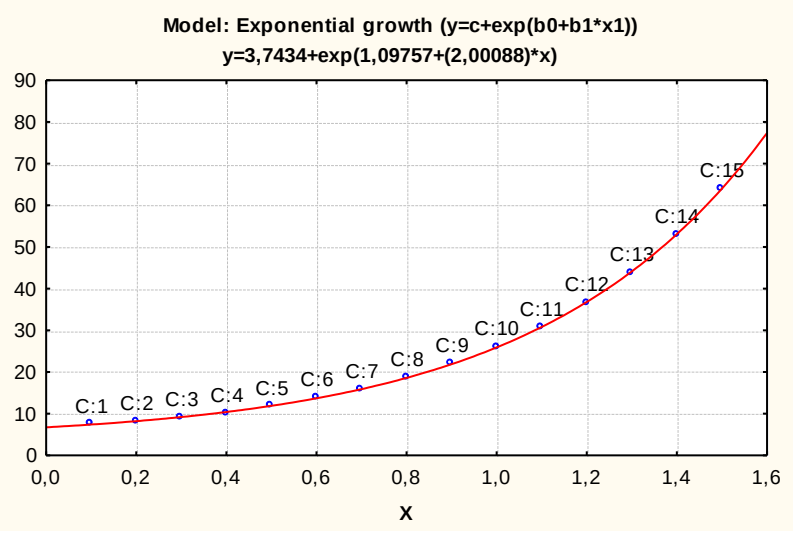


Рис. 14. График кривой, подогнанной к данным

Самостоятельно решите эту же задачу, используя квазиньютоновский метод оценивания параметров модели (Quasi-Newton). Сравните полученные результаты.

Определяемая пользователем регрессия

Допустим, нам известен вид функциональной зависимости между рассматриваемыми переменными с точностью до нескольких параметров. Следует определить функцию потерь, оценить неизвестные параметры и степень адекватности модели.

Создайте файл данных под названием User.sta (табл. 7). Сохраните его.

Из модуля **Nonlinear Estimation (Нелинейное оценивание)** выберите **User-specified regression (Определяемая пользователем регрессия)**, нажмите ОК. Появится окно для задания функциональной зависимости и функции потерь. В верхней части окна в поле **Estimated function (Оцениваемая функция)** задайте следующую функцию:

$$Y = b_0 * \cos(b_1 * X) + b_2 .$$

Таблица 7

Y	X
1,897	0,100
1,389	0,200
1,640	0,300
0,627	0,400
0,193	0,500
0,781	0,600
0,066	0,700
0,125	0,800
1,191	0,900
0,638	1,000
1,130	1,100
1,414	1,200
1,553	1,300
1,581	1,400
2,304	1,500
2,062	1,600
1,927	1,700
2,158	1,800
1,596	1,900
0,563	2,000

В поле **Loss Function (Функция потерь)** по умолчанию будет высвечиваться квадратическая функция потерь:

$$L = (OBS - PRED)**2,$$

воспользуемся ею. Дважды нажмите ОК. На экране появится диалоговое окно **Model Estimation (Оценивание модели)**. Выберите квазиньютоновский метод оценивания параметров. Для определения начальных значений неизвестных параметров, нажмите на кнопку **Start Values (Начальные значения)**. Измените только начальное значение для параметра b_1, введите b_1=3. Дважды нажмите ОК. Появится окно **Parameter Estimation (Оценивание параметров)**, в котором сообщается, что

процесс оценки параметров сошелся за 13 итераций; финальная функция потерь равна 1.57773; оценки параметров равны:

b0=0.977766; b1=4.032834; b2=1.104421.

Таким образом, можно записать следующее уравнение зависимости:

$$Y = 0.977766 * \cos(4.032834 * X) + 1.104421.$$

Теперь рассмотрим результаты графически. В окне **Results (Результаты)** нажмите на кнопку **Fitted 2D function & observed values (Подогнанная функция и наблюдаемые значения).** На графике, который появится на экране, точками изображены исходные данные, а кривая – это подогнанная функция. Видно, что функция достаточно хорошо подошла к данным (рис. 15).

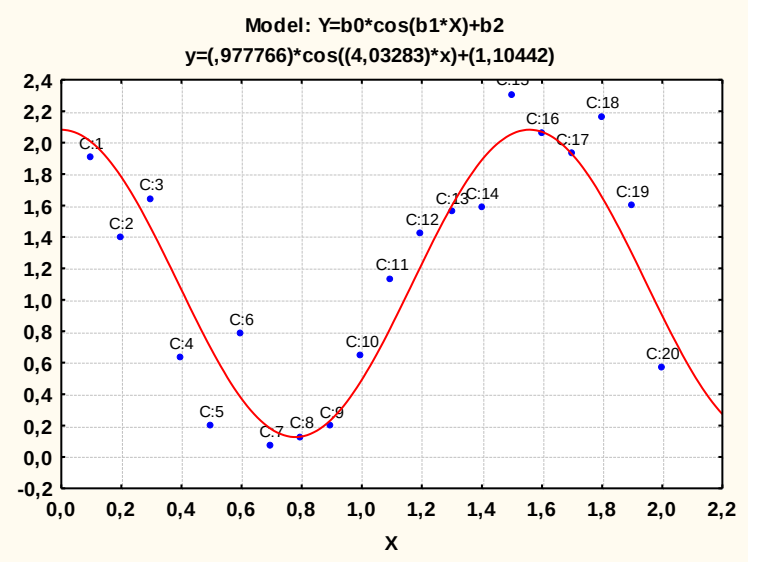


Рис. 15. График кривой, подогнанной к данным

Посмотрим распределение остатков на вероятностной бумаге. Нажмите на кнопку **Normal Probability Plot of Residuals (График остатков на нормальной вероятностной бумаге)**. Из графика видно, что остатки похожи на независимые нормальные величины (рис. 16).

Далее нажмите на кнопку **Predicted values (Прогнозируемые значения)** и сравните полученные данные с исходными. Отличия несущественны. Нажмите на кнопку **Residuals (Остатки)**. Остатки достаточно малы, ими можно пренебречь.

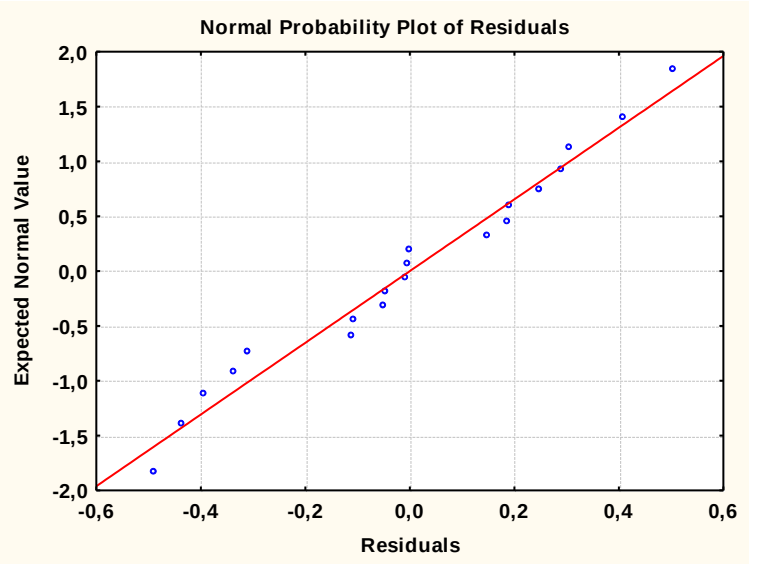


Рис. 16. График остатков на нормальной вероятностной бумаге

Следовательно, в результате проведенного анализа, с уверенностью можно говорить об адекватности построенной модели.

Пример для закрепления темы. Определите вид зависимости между переменными.

1) Создайте файл данных Analys.sta (табл. 8).

2) Подсчитайте по переменным POKAZ1 и POKAZ2 следующие статистики: среднее, минимум, максимум, стандартное отклонение.

3) Подсчитайте коэффициент корреляции между переменными POKAZ1 и POKAZ2.

4) С помощью теста “t-test for dependent samples” проверьте существование зависимости между переменными POKAZ1 и POKAZ2.

5) С помощью графиков определите вид зависимости между переменными POKAZ1 и POKAZ2.

6) Проверьте существование зависимости с помощью регрессионного анализа.

Таблица 8

Text Values	Показатели загрязнения воды		
	DATA	POKAZ1	POKAZ2
northern part	22/07/00	.0966	.1103
northern part	23/07/00	.0980	.1110
northern part	24/07/00	.0986	.1117
northern part	25/07/00	.0992	.1121
northern part	28/07/00	.0988	.1115
northern part	29/07/00	.0992	.1114
northern part	30/07/00	.0979	.1121
east part	22/07/00	.0981	.1149
east part	23/07/00	.0972	.1118
east part	24/07/00	.0974	.1118
east part	25/07/00	.0981	.1119
east part	28/07/00	.0990	.1127
east part	29/07/00	.0986	.1127
east part	30/07/00	.0987	.1129
Bosfor Vostochn	22/07/00	.0996	.1137
Bosfor Vostochn	23/07/00	.1000	.1145
Bosfor Vostochn	24/07/00	.1003	.1156
Bosfor Vostochn	25/07/00	.1016	.1163
Bosfor Vostochn	28/07/00	.1018	.1161
Bosfor Vostochn	29/07/00	.1039	.1176
Bosfor Vostochn	30/07/00	.1068	.1236

В первом столбце – имена случаев.

Комментарий к заголовку: Показатели загрязнения воды Амурского залива.

Тема 4. ДИСПЕРСИОННЫЙ АНАЛИЗ

Теоретическое введение

Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящий от разных, одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния.

Идея дисперсионного анализа заключается в разложении общей дисперсии случайной величины на независимые случайные слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение этих дисперсий позволяет оценить существенность влияния факторов на исследуемую величину.

Если исследовать влияние одного фактора на исследуемую величину, то речь идет об *однофакторном анализе*. Если изучается влияние двух факторов, то речь идет о *двухфакторной* анализе и т.д..

Практическая часть

Создайте новый файл под названием zaliv.sta (4v×15c), содержащий данные по глубине, температуре и освещенности по трем заливам: Амурский, Уссурийский и залив Восток (табл. 9).

Таблица 9

ДАННЫЕ ПО РАЙОНАМ			
РАЙОН	ГЛУБИНА	ТЕМПЕР	ОСВЕЩ
AMUR	500	12.0	90
AMUR	700	12.0	70
AMUR	1000	11.0	51
AMUR	1500	10.0	30
AMUR	2000	8.0	17
USSUR	500	12.0	90
USSUR	700	10.0	70
USSUR	1000	9.0	51
USSUR	1500	8.5	30
USSUR	2000	8.0	17
VOSTOK	500	12.1	90
VOSTOK	700	12.1	70
VOSTOK	800	12.0	69
VOSTOK	900	11.2	60
VOSTOK	1000	10.0	51

Единицы измерения: глубина – см; температура – °С; освещенность – кал/см^2.

Первую переменную «район» нужно вводить используя механизм «двойной связи». Для этого в окне переменной нажмите на кнопку **Text Values (Текстовые значения)**. В появившемся диалоговом окне **Text Values Manager** – район **(Менеджер текстовых значений)** в первой строчке в столбце **Text value (Текстовое значение)** введите «amur», в столбце **Numeric (Числовое)** – 1, в последнем столбце **Value label (Замечание меток)** напишите: «Амурский залив»; во второй строчке введите «ussur», 2, «Уссурийский залив», соответственно; в третьей строчке – «vostok», 3, «Залив Восток», соответственно. Нажмите ОК. Таким образом, с помощью механизма «двойной записи» переменной «район» присвоено текстовое значение и числовой эквивалент. Просмотреть установленное соответствие можно, используя кнопку **Text Values Manager (Менеджер текстовых значений)** на панели инструментов электронной таблицы, а для переключения между текстовым и числовым режимом используется кнопка **Display Numbers/Text Values (Показать числовые/текстовые значения)**.

После ввода данных в таблицу сохраните ее.

Для проведения однофакторного дисперсионного анализа в стартовом окне модуля **Basic Statistics and Tables (Основные статистики и таблицы**) щелкните дважды мышкой на строку **Breakdown&one-way ANOVA (Классификация и однофакторный дисперсионный анализ или однофакторная ANOVA)**. Появится диалоговое окно **Descriptive Statistics and Correlations by Groups (Описательные статистики и Корреляции группированных переменных)**, в котором с помощью кнопки **Variables (Переменные)** можно выбрать группирующие и независимые переменные. В нашем примере мы будем проверять гипотезу: значимо ли различие по освещенности в разных заливах. Следовательно, в первом столбце в качестве группирующей переменной нужно выбрать переменную «район», с помощью которой случаи будут разбиты на классы; во втором столбце в качестве зависимой переменной – «освещ», т.е. переменную, которую нужно исследовать. Нажмите дважды ОК. Система проведет необходимые вычисления и откроет диалоговое окно **Results (Результаты)**.

В верхней части окна результатов содержится следующая информация: зависимая переменная – «освещ», группирующая переменная – «район», разбитая на три класса: «amur», «ussur» и «vostok». В нижней части окна размещены кнопки для проведения детального анализа.

Нажмите на кнопку **Categorized histograms (Категоризованные гистограммы)**, и перед вами появятся гистограммы переменной «освещ» для трех классов. Из графиков видно, что для первых двух классов значения освещенности совпадают, а для третьего – существенно отличаются (рис. 17).

Далее щелкните на кнопку **Summary table of means (итоговая таблица средних)**, и увидите таблицу средних значений зависимой переменной для трех классов и общую среднюю. Третье значение больше предыдущих и общего среднего.

На сколько значимо отличие в средних значениях освещенности по трем заливам, можно узнать с помощью таблицы дисперсионного анализа.

Для этого нажмите на кнопку **Analysis of Variance (Дисперсионный анализ)**. Перед вами откроется таблица результатов анализа (табл. 10), где

SS – сумма квадратов внутригрупповая;

df – число степеней свободы;

MS – средние квадраты межгрупповые;

F – значения F-критерия;

p – уровень значимости.

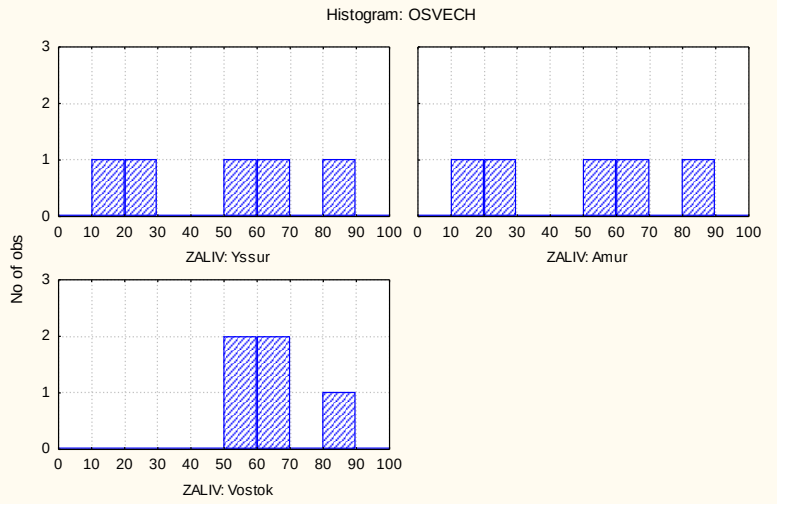


Рис. 17. Гистограмма распределения значений переменной «освещ» по трем классам

Таблица 10

STAT. BASIC STATS Variable	Analysis of Variance (zaliv.sta) Marked effects are significant at p <.05000							
	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
ОСВЕЩ	896.5333	2	448.2667	7796.400	12	649.7000	.689959	.520435

Так как полученный уровень значимости 0.520435>0.05, различия в средних не значимы. Если бы различия на данном уровне были значимы, то система автоматически выделила в таблице значение p красным цветом.

Из результатов проведенного анализа вытекает, что различия по освещенности в Амурском, Уссурийском заливах и заливе Восток в один и тот же период времени не существенны.

Самостоятельно с помощью дисперсионного анализа проверьте следующую гипотезу: значимо ли различие по температуре в разных заливах в один и тот же период времени.

Тема 5. ГРАФИЧЕСКИЕ ВОЗМОЖНОСТИ СИСТЕМЫ STATISTICA

Теоретическое введение

Пакет Statistica обладает широкими графическими возможностями. В нем содержатся сотни типов графиков, включая научные, деловые и специализированные статистические графики. Имеются уникальные графики такие, как лица Чернова, диаграммы Вороного, матричные, категоризованные графики, трассировочные, «ящики с усами» и др. В любом статистическом модуле и на любом этапе статистического анализа вы можете воспользоваться графическими средствами пакета. Также имеется большое количество инструментов настройки всех компонентов графиков, доступ к которым осуществляется при помощи контекстного меню или из панели инструментов графика.

Панель инструментов пакета включает следующие кнопки для вызова и построения графиков определенных типов:

 – Quick Basic Stats – Быстрые основные статистики;

 – Custom 2D Graphs – Пользовательские двумерные графики;

 – Custom 3D Sequential Graphs – Пользовательские трехмерные последовательные графики;

 – Custom 3D Graphs – Пользовательские трехмерные графики;

 – Custom Matrix Plots – Пользовательские матричные графики;

 – Custom Icons Plots – Пользовательские пиктографики;

 – Quick Stats Graphs – Быстрые статистические графики;

 – Graphs Gallery – Галерея графиков.

График является одним из типов документа пакета Statistica (другие типы документов – электронная таблица с данными Spreadsheet, таблица Scrollsheet, отчет) и хранится в файле с расширением *.stg.

Графические средства пакета могут быть использованы в следующих целях:

1. Визуализация численных и текстовых значений непосредственно из электронной таблицы с исходными данными или таблицы с результатами анализа. Для этого существуют следующие две основные группы графиков: статистические и пользовательские графики.

Статистические графики предназначены для визуализации всех значений переменных электронной таблицы и включают в себя Stats Graphs – Статистические графики и Quick Stats Graphs – Быстрые статистические графики.

Пользовательские графики предназначены для визуализации значений из предварительно выделенного блока в электронной таблице и содержат Custom Graphs – Пользовательские графики и Block Stats Graphs – Блоковые статистические графики.

2. Вывод результатов статистического анализа в виде последовательности графиков. В каждом статистическом модуле имеется возможность построения различных графиков для отображения результатов анализа или выбора последующего направления исследования.

Важной особенностью статистических и пользовательских графиков в пакете Statistica является возможность их обновления при изменении данных из электронных таблиц с исходными данными. Связь между данными и графиками может быть автоматической (обновление графика происходит автоматически при изменении данных) или ручной (для перерисовки графика необходимо выполнить дополнительные действия). Если график построен на основе таблицы Scrollsheet, то такая возможность отсутствует.

Практическая часть

Рассмотрим несколько примеров создания статистических и пользовательских графиков.

Откройте файл данных Osvech.sta. Для быстрого доступа к статистическим графикам существуют два способа: при помощи панели инструментов и команд меню.

Первый способ. Воспользуемся кнопкой **Graphs Gallery (Галерея графиков)** на панели инструментов (левая вертикальная панель). В появившемся диалоговом окне нужно выбрать: 1) категорию графиков – **Stats 2D Graphs (Статистические двумерные графики**); 2) тип графика – **Scatterplots (Диаграмма рассеяния)**; 3) вид точечного графика – **Regular (Регулярный).**

Второй способ. Из ниспадающего меню **Graphs (Графики)** выберите **Stats 2D Graphs (Статистические двумерные графики)**, затем опцию **Scatterplots (Диаграмма рассеяния).**

Построим график зависимости освещенности от глубины. В диалоговом окне **2D Scatterplots (Двумерная диаграмма рассеяния)** в первой колонке выделите переменную «глубина», во второй – «освещ». Нажмите ОК. Далее в поле **Graph Type (Тип графика)** выделите **Regular (Регулярный)**; в поле **Fit (Подгонка)** – **Linear (Линейный)**. Нажмите ОК. На экране появится график зависимости, где по оси OX отложены значения переменной «глубина», по оси OY – значения переменной «освещ». Сверху, под заголовком автоматически выводится уравнение прямой, изображенной на графике (рис. 18).

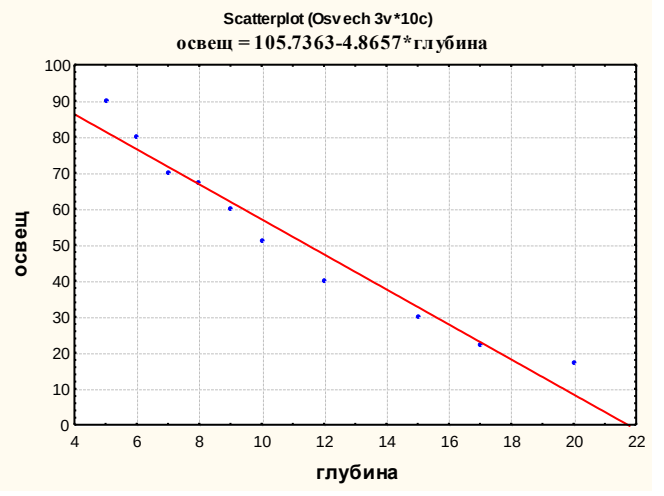


Рис. 18. Статистический график зависимости освещенности от глубины

Из графика видно, что при увеличении глубины освещенность падает, причем связь между величинами линейная.

В пакете Statistica имеется возможность настройки всех элементов графика. Для этого достаточно щелкнуть два раза мышкой на необходимом элементе и выполнить изменения.

Для построения пользовательского графика можно воспользоваться кнопкой **Custom 2D Graphs (Пользовательские двумерные графики)** на панели инструментов или из меню **Graphs (Графики)** выбрать **Custom 2D Graphs (Пользовательские графики).**

Построим график зависимости температуры от глубины. В диалоговом окне **Spreadsheet: Custom 2D Graphs (Пользовательские дву-**

мерные графики) в рамке **Plot1 (График 1)** выберите **Line Plot (Линейный график)**, в поле **X** – «глубина», в поле **Y** – «темпер». В рамке **Value (Значения)** можно задать номера случаев, которые будут отображены на графике. По умолчанию используются все случаи (в нашем примере десять случаев – **From:1 To:10**). Нажмите ОК. На экране появится график исследуемой зависимости, из которого видно, что с увеличением глубины, температура падает (рис. 19). Это соответствует действительности.

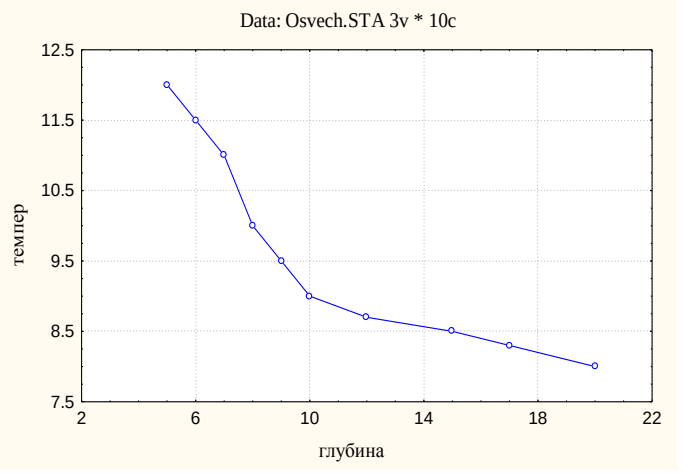


Рис. 19. Пользовательский график зависимости температуры от глубины

Самостоятельно постройте пользовательский трехмерный график зависимости температуры и освещенности от глубины (Graphs/Custom Graphs/3D XYZ Graphs). Проанализируйте его.

Для построения пользовательского матричного графика распределения глубины и температуры нажмите кнопку на панели инструментов **Custom Matrix Plots (Пользовательские матричные графики)**. В появившемся диалоговом окне в поле **Graph Type (Тип графика)** выберите **Scatterplot Matrix (Матрица диаграммы рассеяния)**; в поле **Plots (Графики)** введите: 1 – «глубина», 2 – «темпер»; ОК. На экране высветится матричный график, состоящий из набора кадров, где по главной диагонали выводятся гистограммы значений выбранных переменных, а в остальных кадрах – диаграммы рассеяния для всех комбинаций переменных из выделенного блока (рис. 20). В нашем случае матричный график показывает, что переменные «глубина» и «темпер» сильно коррелированны между собой.

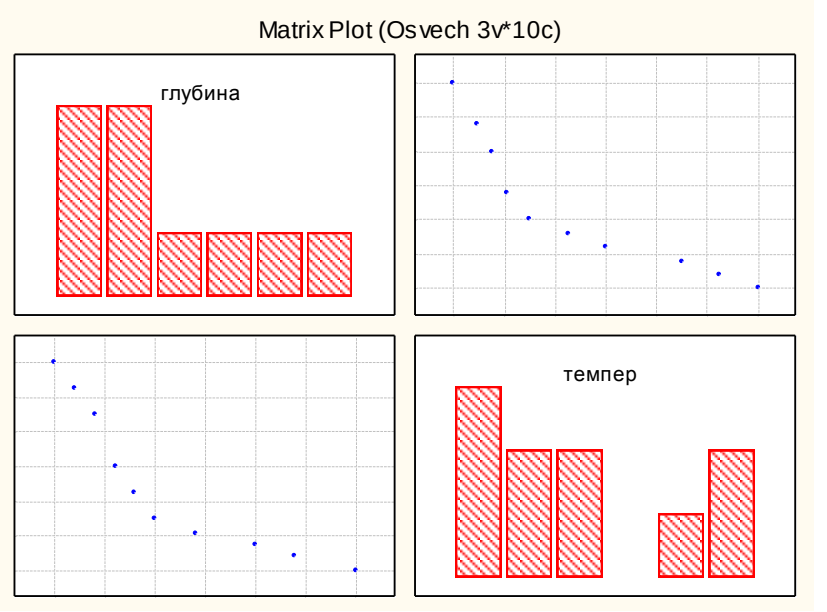


Рис. 20. Матричный график для выделенного блока переменных

Рассмотрим пример построения диаграммы «Лица Чернова». Вызовите диалоговое окно **Custom Icons Graphs (Пользовательские пиктографики)**. В поле **Graph Type (Тип графика)** выделите **Chernoff Faces (Лица Чернова)**, в поле **Plots (Графики)** введите последовательно «глубина», «темпер», «освещ». Нажмите ОК. Появится пользовательский пиктографик, позволяющий визуально анализировать изменение сразу трех переменных. Изменения отражаются формой и величиной различных частей «лиц» на графике (рис. 21).

В пакете Statistica существует удобный инструмент для интерактивного графического анализа данных, так называемая, Кисть, с помощью которой можно установить соответствие между точками на графике и их числовыми значениями. Можно также, выделив необходимые точки при помощи мыши, пометить их маркером, временно удалить или вывести их метки, перейти в режим просмотра координат этих точек.

Построим диаграмму рассеяния с доверительным интервалом на основе данных файла craps.sta из папки Examples (данные по крабам). Для этого в диалоговом окне **2D Scatterplots (Двумерная диаграмма рассеяния)** введите 4-ую и 7-ую переменные: width и catwidth; в поле **Graph Type (Тип графика)** выделите **Regular (Регулярный)**; в поле **Fit (Подгонка) – OF (Рассеяние);** в поле **Confidence bands (Доверительный интервал) – Prediction (Предсказанный доверительный интервал)**. Нажмите ОК.

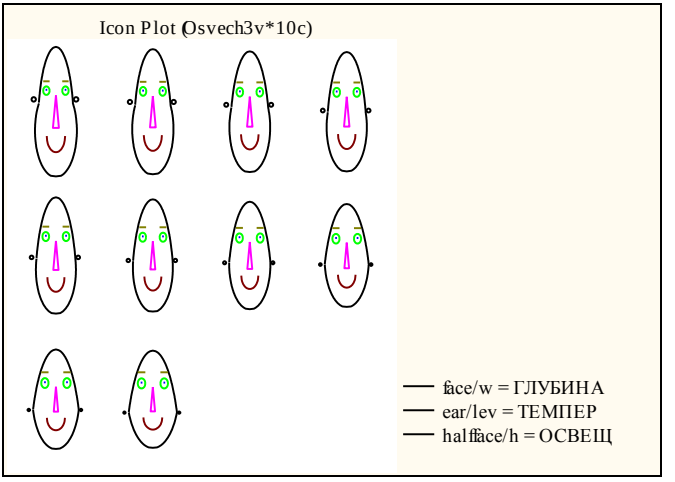


Рис. 21. Пользовательский пиктографик «Лица Чернова»

Анализируя полученный график, можно сделать вывод, что данные достаточно хорошо входят в доверительный интервал, за исключением нескольких точек (рис. 22). Для определения этих точек и последующего исключения их из анализа воспользуемся инструментом Кисть.

На панели инструментов нажмите на кнопку **Brushing (Кисть)**. В правой части окна появится панель управления Кистью. Рассмотрим ее подробнее.

Кнопка **Update (Обновить)** – предназначена для выполнения выбранной операции;

кнопка **Quit (Выход)** – для выхода из режима Кисть;

кнопка **De-select All (Отменить все)** – отмена всех ранее проведенных операций;

опция **Auto Update (Автоматическое обновление)** – автоматическое обновления графика после каждой операции;

рамка **Action (Операция)** содержит следующие возможные операции над выделенными точками:

Mark (Маркировка) – выделение точек;

Label (Метка) – вывод меток выделенным точкам;

Turn OFF (Отключить) – отключение выделенных точек, т.е. они не будут отображаться на графике и участвовать при построении аппроксимирующей кривой;

De-select (Отменить выделение) – отменяет выделение точек;

Reverse (Обратить операцию) – обращает все операции в противоположные;

в рамке **Brush (Кисть)** можно задать форму Кисти:

Point – точка;

Rectangle – прямоугольник;

Lasso – лассо;

опция **Animation (Анимация)** позволяет передвигать по графику точки, выделенные с помощью прямоугольника или лассо (имеет свое диалоговое окно);

кнопка **More (Больше)** – вызывает диалоговое окно с более специализированными условиями для выбора точек.

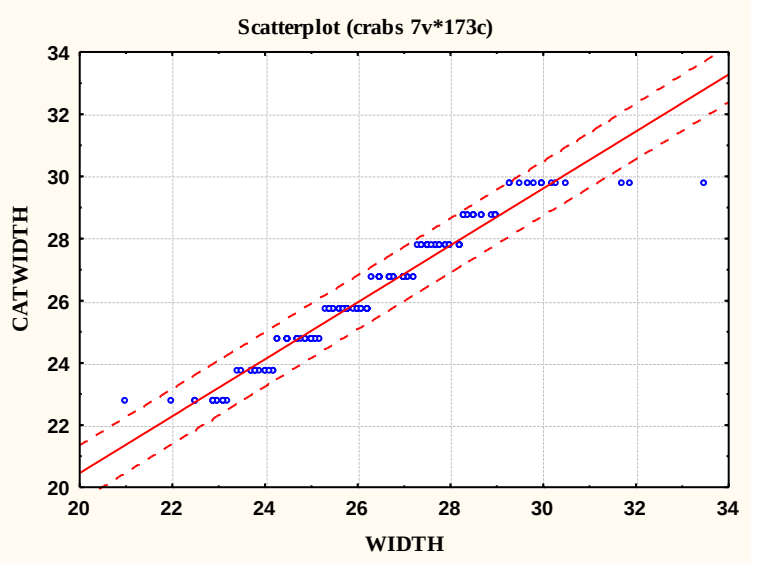


Рис. 22. Диаграмма рассеяния с доверительным интервалом

Вернемся к нашему графику (рис. 22). Имеются четыре точки, не вошедшие в доверительный интервал. Выделим их. Для этого в рамке **Action (Операция)** на панели управления Кистью выделите опцию **Mark (Маркировка)**, в рамке **Brush (Кисть)** выделите – **Point** (Точка). Затем, подведите «прицел» к нужным точкам и нажмите левой клавишей мыши. Точки выделятся черным цветом.

Для отображения меток выделенных точек в рамке **Action (Операция)** выберите **Label (Метка)** и нажмите кнопку **Update (Обновить)**. На графике появятся метки точек, соответствующие случаям исходной таблицы данных.

Для удаления этих точек выберите опцию **Turn OFF (Отключить)** и нажмите **Update (Обновить)**. Выделенные точки удалятся, и график автоматически будет перерисован.

СПИСОК РЕКОМЕНДОВАННОЙ ЛИТЕРАТУРЫ

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

2. Боровиков В.П., Боровиков И.П. Statistica: искусство анализа данных на компьютере. Для профессионалов. – СПб.: Питер, 2001. 656 с.

3. Бородин А.Н. Элементарный курс теории вероятностей и математической статистики: Учебное пособие для вузов. – СПб.: Лань, 2002. 254 с.

4. Гмурман В.Е. Теория вероятностей и математическая статистика: Учебное пособие. – М.: Высшая школа, 2002. 480 с.

5. Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel: Учебное пособие для вузов. – Ростов н/Д: Феникс, 2002. 400 с.

6. Елисеева И.И. Общая теория статистики: Учебник для вузов. – М.: Финансы и статистика, 2002. 479 с.