**КАЗАНСКИЙ (ПРИВОЛЖСКИЙ)**
**ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**

**Факультет географии и экологии**

# СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ЭКОЛОГИИ И ПРИРОДОПОЛЬЗОВАНИИ

## (с использованием программы STATGRAPHICS Plus)

**Учебно-методическое пособие**

**КАЗАНЬ – 2011**

Составители:
кандидат географических наук, старший преподаватель
К.А. Мальцев,
старший преподаватель С.С. Мухарамова

Данное учебно-методическое пособие разработано для студентов естественных факультетов и может быть использовано при проведении курсов, посвященных изучению методов статистического описания и анализа данных. В пособии кратко приводятся: классификация типов данных, методы описательной статистики для числовых данных, корреляционного и регрессионного анализа, анализа временных рядов. Применение изучаемых статистических методов рассматривается на примере их реализации в программе STATGRAPHICS Plus. Для выполнения практических заданий используется информация базы данных «Геоэкология речных бассейнов Республики Татарстан» и данные о среднемесячных концентрациях сульфатов в атмосферных осадках, полученные на метеостанции Мудьюг.

Печатается по решению учебно-методической комиссии факультета географии и экологии.

ОГЛАВЛЕНИЕ

# Классификация типов данных

Изучаемые величины (признаки, переменные), требующие статистического анализа, принято относить к одному из трех типов данных – номинальному (категориальному, качественному), ординальному (порядковому, ранговому) или скалярному (количественному, числовому). Их значения могут быть получены (измерены, наблюдаемы) с использованием соответствующих шкал измерения.

*Номинальные* переменные могут принимать значения, измеренные на некоторой номинальной шкале, состоящей из наименований категорий, которые никак естественным образом не упорядочиваются. Например, номинальная переменная «лесные формации», которая может иметь категории «дубравы», «липняки», «осинники», «березняки», «сосняки», «ельники». Если в номинальных шкалах используются числа, то они служат только для различения отдельных возможностей, заменяя названия и имена. Никаких соотношений, кроме равенства или неравенства, между такими значениями нет. Эти данные не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

*Ординальные* переменные измерены на ординальной шкале, имеющей упорядоченные категории. Например, ординальная переменная «степень присутствия вида» может иметь категории «вид отсутствует», «вид присутствует», «вид содоминирует», «вид доминирует». К таким переменным можно отнести различные балльные или экспертные оценки с очевидным упорядочением значений. Балльная оценка успеваемости («неудовлетворительно», «удовлетворительно», «хорошо»,

«отлично») является типичным примером порядковой величины. Измерения в ординальной шкале содержат информацию только о порядке следования величин, но не позволяют количественно выразить, насколько или во сколько раз одно значение больше или меньше другого. Для таких данных применимы только операции сравнения и ранжирования: «равно», «не равно», «больше», «меньше»; арифметические действия не могут быть произведены.

*Скалярные* переменные определяют числовые величины, измеряемые на некоторой или интервальной, или относительной, или абсолютной шкале. Они подразделяются на *дискретные* (в качестве значений которых выступают отдельные числа, обычно целые) и *непрерывные* (значениями которых могут служить любые действительные числа из какого-либо отрезка числовой оси). В качестве примера непрерывных переменных можно привести температуру, рост, массу тела, а дискретных – численность населения, число особей в популяции. Скалярные величины, измеренные на интервальной шкале, могут сравниваться, упорядочиваться, складываться, вычитаться. Примеры таких величин – время, высота местности, температура по Цельсию – это величины, которые по физической природе либо не имеют абсолютного нуля, либо допускают свободу выбора в установлении начала отсчета. Для числовых величин, измеренных на относительной или абсолютной шкале, помимо операций сравнения и упорядочивания, применимы любые арифметические действия. Примеры величин, измеренных в шкале отношений - вес, длина, электрическое сопротивление, температура по Кельвину, деньги; в абсолютной шкале – количество предметов. Эти шкалы имеют абсолютную нулевую

точку, которая характеризует полное отсутствие измеряемого свойства, а абсолютная шкала – еще и абсолютную безразмерную единицу.

Шкалы могут приводиться одна к другой: количественная – к ординальной или номинальной, ординальная шкала – к номинальной. Приведение одной шкалы к другой называют понижением шкалы; оно ведет к потере некоторой части информации об изучаемых признаках. Обратные операции считаются некорректными. Понижение шкал применяется при анализе переменных, измеренных в разных шкалах.

Тип данных определяет, какими статистическими методами эти данные могут обрабатываться и анализироваться. Поэтому первым шагом любого статистического анализа является определение типа исследуемых данных и отнесение их к той или иной шкале измерения - номинальной, порядковой или одной из количественных.

## Данные для практических заданий

Для выполнения практических заданий в рамках настоящей работы в качестве исходных данных используется информация базы данных «Геоэкология речных бассейнов Республики Татарстан», структура которой приведена в таблице 1. Здесь поле ID содержит идентификаторы водосборных бассейнов; остальные поля содержат наблюденные значения различных переменных (признаков), характеризующих бассейны по условиям протекания эрозионных процессов; их статистическому анализу посвящены практические задания.

Таблица 1.

| **Имя поля** | **Содержание** | **Тип данных** |
|---|---|---|
| ID | номер водосборного бассейна | номинальные |
| W05 | залесенность | скалярные |
| W06 | средний уклон | скалярные |
| W07 | глубина эрозионного расчленения | скалярные |
| W0402 | модуль годового стока | скалярные |
| W13 | гранулометрический состав почв | ординальные |
| W17 | распаханность | скалярные |
| W20 | содержание гумуса | скалярные |
| W080101 | показатель бассейновой эрозии | скалярные |

Для выполнения практического задания по анализу временных рядов используются данные о среднемесячных концентрациях сульфатов в атмосферных осадках, полученные на метеостанции Мудьюг за период с 1958 г. по 2007 г.

## Методы описательной статистики

В задачах экологии и природопользования мы часто имеем совокупность наблюдений, на основе которых нужно сделать какие-либо выводы. Часто подобных наблюдений много, и возникает задача их компактного описания с использованием различных показателей и графиков. Эта совокупность показателей и графиков относятся к средствам описательной статистики. Здесь мы очень коротко остановимся на том, что включает в себя описательная одномерная статистика. Подробную информацию об этом предмете можно найти в многочисленных книгах по математической статистике.

Описательная одномерная статистика обеспечивает простой путь для организации и систематизации выборочных данных. Эта организация данных использует *гистограмму*, которая есть график *таблицы частот*, то есть таблицы, регистрирующей, как часто наблюдаемые данные попадают в определенные классы (интервалы значений). Таблица частот и гистограмма дают представление о законе распределения изучаемой величины. Статистическое описание данных в общем случае начинается с анализа гистограммы, по которой, в первую очередь, определяются и, если признаются ошибочными, корректируются выделяющиеся данные.

Для числовых данных важнейшие особенности большинства распределений могут быть представлены с помощью нескольких показателей описательной статистики. Эти показатели можно разбить на три группы:

1) *характеристики положения* – описывают положение данных на числовой оси - выборочные среднее, медиана, мода, минимум, максимум, квартили, квантили;

2) *характеристики рассеяния* - описывают степень разброса данных относительно своего центра – выборочные дисперсия, среднеквадратическое отклонение, размах выборки;

3) *характеристики формы* – выборочные коэффициент асимметрии, коэффициент эксцесса, положение выборочной медианы относительно выборочного среднего и относительно выборочных квартилей.

Статистическое описание выборочных данных должно предварять любой статистический анализ. Посмотрим, как описательные статистики могут быть получены с помощью программы STATGRAPHICS.

## *Загрузка данных*

Первое что необходимо сделать для работы в любом статистическом пакете – это загрузить в него данные. В STATGRAPHICS загрузка данных осуществляется через пункт меню **«File»->«Open»**. В зависимости от того, что необходимо загрузить (файл с данными или рабочий набор), можно воспользоваться подпунктами **«Open DataFile»**, **«Open StatFolio»**. Если данные загружаются первый раз, то необходимо выбрать пункт меню **«Open DataFile»**. После этого появится диалог (рис. 1), который регламентирует то, как читать текстовый файл в кодировке ASCII.

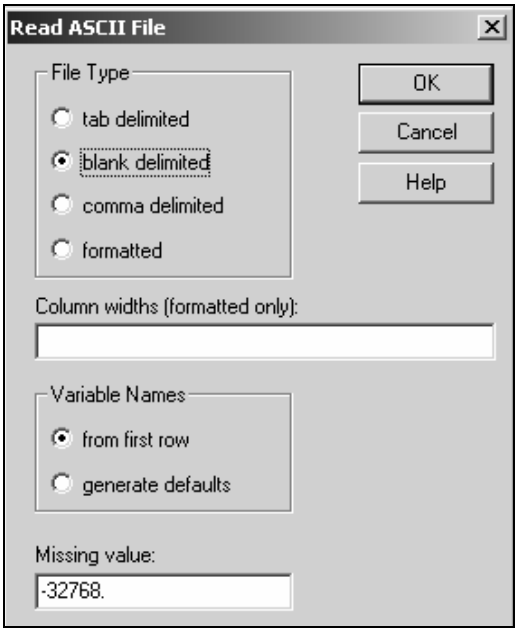



Рис.1. Диалог параметров чтения текстовых файлов в кодировке ASCII

Здесь в блоке «File Type» можно выбрать то, какой разделитель используется в строках исходного текстового файла.

Если выбран пункт «formatted», то, используя строку «Column width», можно задать ширину столбца при чтении текстового файла и обойтись без разделителя. В блоке «Variable Names» указывается, использовать ли в качестве названия столбцов значения из первой строки текстового файла, или (если этих названий нет в текстовом файле) задать их по умолчанию. Последний пункт «Missing value» предназначен для определения кода, заменяющего отсуствующие значения (обозначение отсутствия данных). Задав необходимые параметры чтения текстового файла, нажимаем на кнопку «ОК». В результате этих действий появится окно с таблицей загруженных данных. По умолчанию окно таблицы свернуто и находится в левом нижнем углу рабочего окна программы.

Итак, данные загружены, и мы можем приступать непосредственно к статистическому анализу.

## *Статистическое описание одномерной выборки*

Для проведения статистического описания числовых величин нужно выбрать пункт меню **«Describe»** подпункт **«Numeric Data» -> «One-Variable Analysis»**, после чего появляется стандартный диалог программы, предназначенный для выбора анализируемых данных (рис. 2). Поскольку этот диалог (с небольшими модификациями) появляется в программе при проведении большинства видов статистического анализа, остановимся на его описании подробнее.

Диалог выбора данных состоит из одного списка, двух строк ввода и пяти кнопок. В списке левой части диалога перечислены поля (столбцы) таблицы данных, доступные для статистического анализа. Справа вверху находится строка ввода

«Data», в которую помещается имя поля, данные из которого будут проанализированы. Имя поля можно задать вручную, или выбрать из списка указателем курсора и нажать кнопку  . Строка «Select», расположенная ниже, позволяет произвести отбор данных по определенному условию. Например, условие W0402>0 говорит о том, что будут анализироваться те данные из поля, указанного в строке «Data», у которых в поле W0402 стоит значение, большее нуля. Из пяти кнопок особого внимания заслуживает кнопка **«Transform»**. Здесь дается возможность задать функцию от значений полей и далее анализировать уже не сами значения того или иного поля, а функцию от них. Например, запись «LOG10 (W0402)» в строке «Data» означает, что будет анализироваться десятичный логарифм от значений, расположенных в поле W0402. Логарифмирование может быть полезным для приведения выборки к виду нормального распределения, если изначально распределение имело вид логнормального.

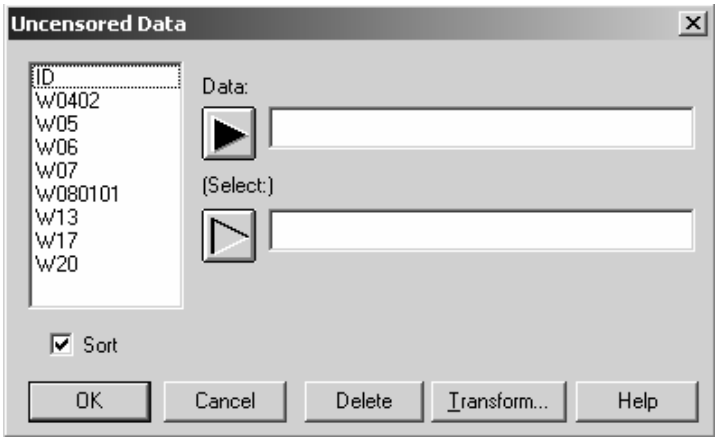



Рис.2. Диалоговое окно для определения анализируемой переменной

После нажатия на кнопку «ОК» появляется окно результата *«One-Variable Analysis»*. Первоначально в этом окне присутствует информация только об объеме выборки, а также о минимальном и максимальном выборочных значениях.

Для оценки других статистических показателей нужно нажать на кнопку  - **«Tabular Options»**. В появившемся диалоге, который показан на рис. 3, присутствует семь различных переключателей. При этом изначально всегда отмечен первый - **«Analysis Summary»**, что и обеспечивает вычисление объема выборки, минимума и максимума.

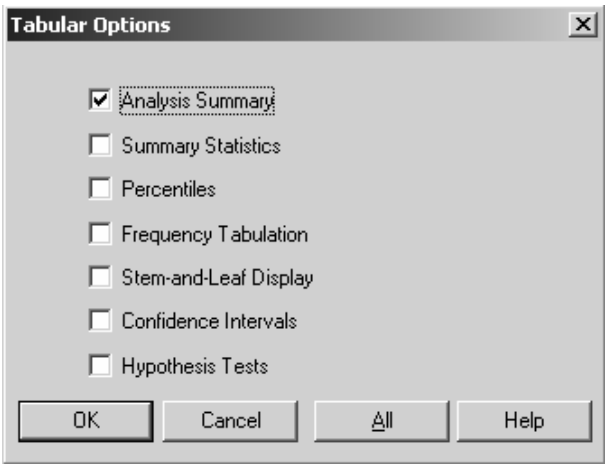



Рис.3. Диалоговое окно для задания рассчитываемых статистических характеристик

При выборе пункта **«Summary Statistics»** будут рассчитываться статистические показатели, приведенные в таблице 2. Однако этот список отражает не полный набор показателей, характеризующих выборку. Если в окне результата *«One-Variable Analysis»* в разделе *«Summary Statistics»* нажать на правую клавишу мыши, а во всплывающем меню выбрать пункт

**«Pane Options»**, то появится окно, в котором можно задать расчет показателей, приведенных в таблице 3.

Таблица 2.

Статистические показатели (основной список)

| **Summary Statistics** | **Описательные статистики** |
|---|---|
| Count | Объем выборки |
| Average | Выборочное среднее |
| Variance | Выборочная дисперсия |
| Standard deviation | Среднеквадратическое отклонение |
| Minimum | Минимум |
| Maximum | Максимум |
| Stnd. skewness | Нормированный коэффициент асимметрии |
| Stnd. kurtosis | Нормированный коэффициент эксцесса |
| Sum | Сумма |

Таблица 3.

Статистические показатели (дополнительный список)

| **Summary Statistics** | **Описательные статистики** |
|---|---|
| Median | Медиана |
| Mode | Мода |
| Geometric mean | Геометрическое среднее |
| Std.Error | Стандартная ошибка |
| Range | Размах |
| Lower quartile | Нижний квартиль |
| Upper quartile | Верхний квартиль |
| Interquar. Range | Межквартильный размах |
| Skewness | Коэффициент асимметрии |
| Kurtosis | Коэффициент эксцесса |
| Coeff. of var | Коэффициент вариации |

Кроме показателей, приведенных в таблицах 2 и 3, можно получить еще ряд статистических характеристик. Для этого в диалоговом окне, показанном на рис. 3, нужно выбрать другие пункты. При выборе пункта **«Percentiles»** вычисляются выборочные процентили (квантили) для указанных пользователем процентов. По умолчанию рассчитываются процентили для 1; 5; 10; 25; 50; 75; 90.0; 95.0; 99%.

При выборе пункта **«Frequency Tabulation»** строится таблица частот и накопленных частот после группировки данных в заданное число классов.

При выборе пункта **«Confidence Intervals»** будут рассчитаны доверительные интервалы для среднего и среднеквадратического отклонения с заданной доверительной вероятностью.

При выборе пункта **«Hypothesis tests»** проверяются гипотезы (на заданном уровне значимости) о равенстве среднего различным значениям, которые может задать исследователь.

Кроме рассмотренных числовых показателей характер распределения выборки хорошо представляют различные графики. Если в окне *«One-Variable Analysis»* нажать на кнопку  - **«Graphical options»**, то появится диалог, представленный на рис. 4. Здесь также присутствует семь переключателей, каждый из которых отвечает за определенный вид графика. Мы в рамках данного методического пособия рассмотрим лишь построение гистограмм частот, которое реализуется при выборе пункта **«Frequency Histogram»**.

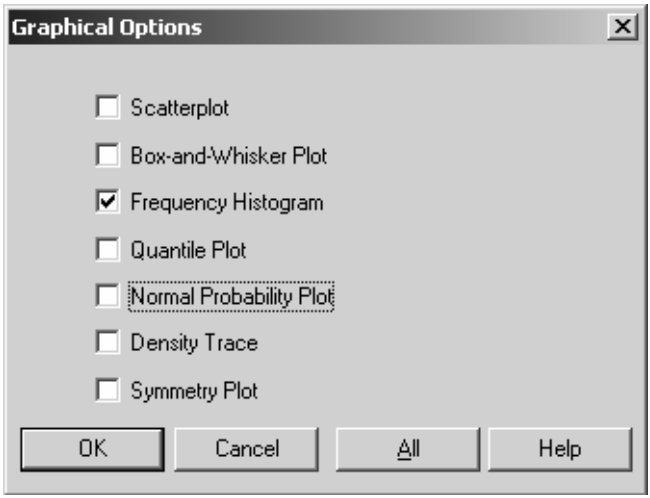



Рис.4. Диалоговое окно выбора графиков

Вид гистограммы частот, построенной для одних и тех же данных, может отличаться в зависимости от того, на сколько классов разбит интервал, в котором изменяются значения выборки.

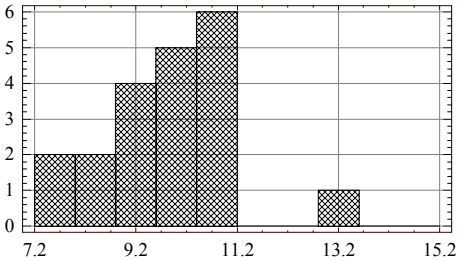



Рис.5. Гистограмма частот с 10 классами

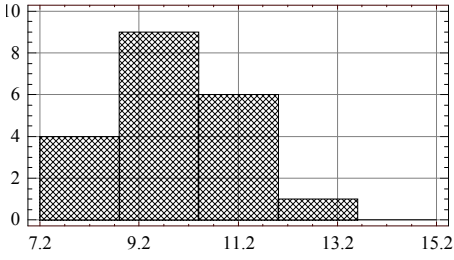



Рис.6. Гистограмма частот с 5 классами

Количество классов можно задать в окне, появляющемся при выборе пункта **«Pane Options»** во всплывающем меню после нажатии правой клавиши мыши на графике гистограммы.

## *Задание 1*

1. Провести статистическое описание следующих выборочных переменных: залесенности (W05), среднего уклона (W06), глубины эрозионного расчленения (W07), модуля годового стока (W0402), содержания гумуса (W20), распаханности (W17), показателя бассейновой эрозии (W080101). В качестве статистик, описывающих распределение выборочной переменной, использовать следующие показатели: среднее, медиану, минимум, максимум, 1%-й квантиль, 99%-й квантиль, стандартное отклонение, коэффициент асимметрии. Полученные данные представить в виде таблицы, где по строкам будут расположены переменные, а по столбцам статистические показатели.

2. Построить гистограммы частот для каждой выборочной переменной с использованием 5 и 10 классов. Графики представить так, как показано на рис. 5 и 6.

3. Проанализировать полученные результаты и дать ответы на следующие вопросы:

- Можно ли визуально (исходя из гистограмм частот) отнести распределение к нормальному или логнормальному?
- Какая из выборочных переменных имеет наибольшую изменчивость?
- Для какой из выборочных переменных наблюдаются наибольшие отличия между минимумом и 1%-м квантилем, а также максимумом и 99%-м квантилем?

## *Проверка на нормальность, критерии согласия*

Применению большинства методов статистического анализа числовых данных непрерывного типа предшествует проверка выборочных данных на согласие с нормальным распределением, поскольку эти методы (в том числе многие методы корреляционного, регрессионного анализа и др.) исходят из предположения нормальности распределения изучаемых данных.

Существует несколько тестов (критериев согласия), которые позволяют проверить гипотезу о нормальности распределения. К ним относятся критерии хи-квадрат, Колмогорова-Смирнова, критерии асимметрии и эксцесса и др. Одной из главных особенностей этих методов является требование достаточно больших объемов (сотни или тысячи) анализируемых данных для получения надежных выводов. При небольшом объеме выборки эти методы способны отвергнуть гипотезу о нормальности распределения только при грубом отклонении от нормального распределения.

Для проверки гипотезы о нормальности распределения нужно выбрать пункт меню **«Describe» -> «Distribution Fitting» -> «Uncensored Data»**, и в появившемся стандартном диалоге выбрать анализируемые данные. В окне результата *«Uncensored Data»* первоначально будут приведены лишь минимум, максимум, среднее и среднеквадратическое отклонение данных.

В STATGRAPHICS реализованы как специальные критерии согласия, предназначенные для проверки именно нормальности распределения – критерии нормальности, так и общие критерии согласия, применимые к гипотезе о согласии выборочных данных с любым априорно предполагаемым распределением

вероятностей. Это предполагаемое теоретическое распределение можно задать в окне, появляющемся при выборе пункта **«Analysis Options»** во всплывающем меню после нажатии правой клавиши мыши в окне результата. По умолчанию задано нормальное распределение.

Для вычисления критериев согласия необходимо в окне *«Uncensored Data»* нажать на кнопку  - **«Tabular Options»**, а в появившемся диалоге выбрать пункты **«Tests for normality»** и **«Goodness-of-Fit Tests»**. В окне результата *«Uncensored Data»* появятся соответствующие разделы, где будут приведены наблюденные (рассчитанные по выборке) значения критериев нормальности: специальная версия классического критерия хи-квадрат, критерий Шапиро-Уилка, критерии асимметрии и эксцесса (рис. 7), а также общих критериев согласия: хи-квадрат, Колмогорова-Смирнова и др. (рис. 8), и их достигаемые уровни значимости (p-value).

Уровень значимости - это допустимая для данной задачи вероятность ошибки 1-ого рода при проверке гипотезы по статистическому критерию, т.е. вероятность отклонить нулевую гипотезу (в нашем случае отвергнуть гипотезу о нормальности распределения), когда на самом деле она верна. Стандартные значения уровня значимости: 0.005; 0.01; 0.05; 0.1.

Достигаемый уровень значимости (p-value) - это значение функции распределения критерия для его наблюденного значения. Чем меньшее значение p-value мы получаем, тем сильнее совокупность данных свидетельствует против нулевой гипотезы.

Достигаемый уровень значимости (p-value) сравнивается с заданным уровнем значимости, который определяется исследователем. Если p-value ниже заданного уровня значимости, то гипотеза отвергается. Если выше, то принимается.

```
Tests for Normality for RAND1

Computed Chi-Square goodness-of-fit statistic = 50.24
P-Value = 0.000337771

Shapiro-Wilks W statistic = 0.861899
P-Value = 1.73972E-13

Z score for skewness = 3.89454
P-Value = 0.0000984231

Z score for kurtosis = 4.01603
P-Value = 0.0000592146



The StatAdvisor
---------------
   This pane shows the results of several tests run to determine
whether RAND1 can be adequately modeled by a normal distribution.  The
chi-square test divides the range of RAND1 into 24 equally probable
classes and compares the number of observations in each class to the
number expected.  The Shapiro-Wilks test is based upon comparing the
quantiles of the fitted normal distribution to the quantiles of the
data.  The standardized skewness test looks for lack of symmetry in
the data.  The standardized kurtosis test looks for distributional
shape which is either flatter or more peaked than the normal
distribution.
   The lowest P-value amongst the tests performed equals 1.73972E-13.
Because the P-value for this test is less than 0.01, we can reject the
idea that RAND1 comes from a normal distribution with 99% confidence.
```

Рис. 7. Значения критериев нормальности и достигаемых уровней значимости (p-value) при проверке гипотезы о нормальности распределения

Например, если окажется, что достигаемый уровень значимости одного из статистических критериев (хи-квадрат, Шапиро-Уилка и т.д.) меньше заданного уровня значимости равного 0.01, то это означает, что гипотеза о нормальном распределении выборки отвергается с доверительной вероятностью 99%.

Goodness-of-Fit Tests for RAND1

Chi-Square Test

| Lower Limit | Upper Limit | Observed Frequency | Expected Frequency | Chi-Square |
|---|---|---|---|---|
| at or below | 0.40217 | 5 | 12.50 | 4.50 |
| 0.40217 | 0.789942 | 14 | 12.50 | 0.18 |
| 0.789942 | 1.07992 | 27 | 12.50 | 16.82 |
| 1.07992 | 1.33958 | 13 | 12.50 | 0.02 |
| 1.33958 | 1.59923 | 17 | 12.50 | 1.62 |
| 1.59923 | 1.88921 | 5 | 12.50 | 4.50 |
| 1.88921 | 2.27698 | 9 | 12.50 | 0.98 |
| above 2.27698 | | 10 | 12.50 | 0.50 |

Chi-Square = 29.1198 with 5 d.f.  P-Value = 0.0000219661

Estimated Kolmogorov statistic DPLUS = 0.166193
Estimated Kolmogorov statistic DMINUS = 0.0899093
Estimated overall statistic DN = 0.166193
Approximate P-Value = 0.00797931

| EDF Statistic | Value | Modified Form | P-Value |
|---|---|---|---|
| Kolmogorov-Smirnov D | 0.166193 | 1.6744 | <0.01* |
| Anderson-Darling A^2 | 3.48187 | 3.50876 | 0.0000* |

*Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the currently selected distribution. Other P-values are based on general tables and may be very conservative.

The StatAdvisor

This pane shows the results of tests run to determine whether RAND1 can be adequately modeled by a normal distribution. The chi-square test divides the range of RAND1 into nonoverlapping intervals and compares the number of observations in each class to the number expected based on the fitted distribution. The Kolmogorov-Smirnov test computes the maximum distance between the cumulative distribution of RAND1 and the CDF of the fitted normal distribution. In this case, the maximum distance is 0.166193. The other EDF statistics compare the empirical distribution function to the fitted CDF in different ways.

Since the smallest P-value amongst the tests performed is less than 0.01, we can reject the idea that RAND1 comes from a normal distribution with 99% confidence.

Рис. 8. Значения критериев согласия и достигаемых уровней значимости (p-value) при проверке гипотезы о нормальности распределения

Как уже говорилось, тип и распределение данных определяет выбор методов статистического анализа. Одним из необходимых условий применения параметрических методов является нормальное распределение. Если исходные данные не подчиняются закону нормального распределения, их можно трансформировать таким образом, что распределение приблизится к нормальному. При правосторонней асимметрии («хвост» вправо) чаще всего применяют следующие виды «нормализующей трансформации»: извлечение квадратного корня $\sqrt{x}$, логарифмическое преобразование $Ln(x)$ или $Log_{10}(x)$, гармоническое преобразование $(-1/x)$ (знак «минус» используется для сохранения направленности отношений; в противном случае наименьшие и наибольшие значения поменяются местами, что может затруднить интерпретацию результатов). Если переменная может принимать нулевое значение, то для проведения логарифмического или гармонического преобразования следует добавлять к переменной некоторое малое число, например 0,001. При левосторонней асимметрии имеет смысл преобразовывать данные путем их возведения в степень (обычно во вторую или третью). Выбор наиболее подходящего вида трансформации для имеющихся данных определяется методом проб и ошибок, а об успешности преобразования смотрят по графикам, коэффициентам асимметрии и эксцесса и результатам проверки распределения с помощью статистических критериев. В STATGRAPHICS подбор подходящего вида трансформации на основе степенного преобразования (Box-Cox Transformation) может быть выполнен с помощью выбора пункта меню **«Describe»** подпункта **«Numeric Data» -> «Power Transformations»**.

### *Задание 2*

Проверить гипотезу о нормальности распределения следующих переменных: залесенности (W05), среднего уклона (W06), глубины эрозионного расчленения (W07), модуля годового стока (W0402), содержания гумуса (W20), распаханности (W17), показателя бассейновой эрозии (W080101). Полученные результаты представить в виде таблицы, где по строкам будут расположены переменные, по столбцам – примененные критерии согласия, в ячейках - знак «+», если для переменной соответствующий критерий показывает согласие с нормальным распределением, или знак «-», если тестирование дает отрицательный результат.

## Исследование связи признаков

### *Корреляционный анализ*

Для обнаружения связи между переменными, исследования ее силы, направленности служит совокупность методов, называемая корреляционным анализом, в рамках которого оцениваются и анализируются различные показатели связи и их значимость. Цель показателя (меры) связи состоит в том, чтобы дать простой численный ответ на вопрос о степени корреляционной зависимости между двумя переменными. В зависимости от того, к какому типу относятся данные (номинальному, ординальному или скалярному), существуют различные приемы исследования связи признаков.

Для изучения связи между двумя признаками номинального типа применяются таблицы сопряженности, статистика Фишера-

Пирсона хи-квадрат, различные меры связи признаков (коэффициенты Крамера, Юла, Чупрова и др.).

Для признаков, измеренных в порядковой шкале, при исследовании связи применяются коэффициенты корреляции рангов, например Спирмена или Кендала.

При исследовании связи двух количественных переменных применяются коэффициент корреляции, корреляционное отношение, коэффициенты корреляции рангов. Коэффициент корреляции нашел широкое применение в практике, но важно помнить, что он не является универсальным показателем корреляционной зависимости, так как способен характеризовать только линейную форму связи (в отличии от корреляционного отношения). Для количественных признаков, которые не показывают нормальности своего распределения, при исследовании связи корректным является использование лишь коэффициентов корреляции рангов.

В STATGRAPHICS расчет коэффициентов корреляции для скалярных и ординальных величин реализован в пункте главного меню **«Discribe» -> «Numeric Data» -> «Multi-Variable Analysis»**, при выборе которого появляется стандартный диалог для определения анализируемых данных. Здесь в строке «Data» нужно задать две или более переменные, между которыми необходимо оценить показатели связи. В окне результата *«Multi-Variable Analysis»* первоначально выводится список анализируемых переменных и объем многомерной выборки. Для расчета коэффициентов корреляции нужно нажать на кнопку  - **«Tabular Options»**, а в появившемся окне (рис. 9) выбрать пункты **«Correlations»** и/или **«Rank Correlations»**. При этом

нужно помнить о типе анализируемых данных и корректно выбирать оцениваемые показатели связи, а также интерпретировать получаемые результаты (например, не вычислять коэффициент корреляции, если данные относятся к ординальному типу).

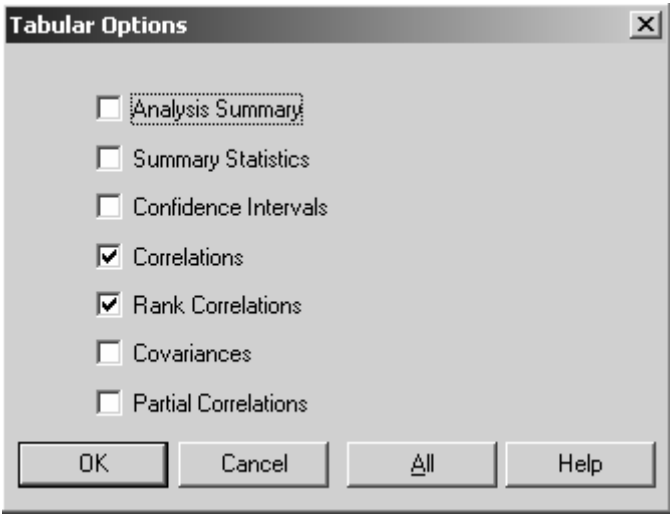
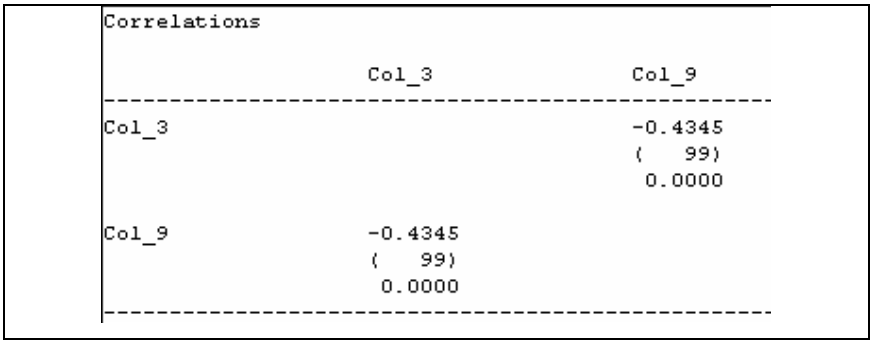



Рис. 9. Диалоговое окно для задания рассчитываемых статистических показателей.

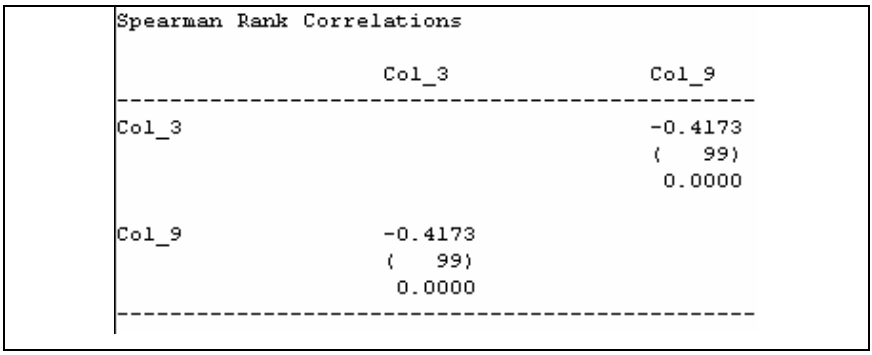
Результаты оценки коэффициентов корреляции и коэффициентов ранговой корреляции будут представлены в виде корреляционных матриц в окне результата *«Multi-Variable Analysis»* в разделах *«Correlations»* и *«Spearman Rank Correlations»*, соответственно (рис. 10, 11).



Correlations

| | Col_3 | Col_9 |
|---|---|---|
| Col_3 | | -0.4345<br>( 99)<br>0.0000 |
| Col_9 | -0.4345<br>( 99)<br>0.0000 | |

Рис. 10. Результаты расчета коэффициента корреляции



Spearman Rank Correlations

| | Col_3 | Col_9 |
|---|---|---|
| Col_3 | | -0.4173<br>( 99)<br>0.0000 |
| Col_9 | -0.4173<br>( 99)<br>0.0000 | |

Рис. 11. Результаты расчетов рангового коэффициента корреляции Спирмена

Для каждой пары переменных выводятся три значения. Первое – это оценка коэффициента корреляции (рис.10) или коэффициента корреляции рангов Спирмена (рис.11). Второе значение – объем выборки. Третье значение - это достигаемый уровень значимости (p-value) полученной оценки коэффициента корреляции. Если p-value меньше 0.05, то это говорит о наличии статистической значимой ненулевой корреляции между этими двумя переменными с 95%-й надежностью.

В приведенном примере на рис. 10, 11 достигаемый уровень значимости равный 0.0000 означает, что с вероятностью не

меньшей 99 % приведенные оценки коэффициентов корреляции и корреляции рангов значимы, т.е. переменные достоверно связаны корреляционной зависимостью.

### *Задание 3*

Оценить корреляционную связь следующих пар переменных: W05/W080101, W06/W080101, W07/W080101, W0402/W080101, W20/W080101, W17/W080101, W13/W080101. При этом необходимо учесть, что переменная W13 (гранулометрический состав почв) является ординальной. Полученные результаты свести в таблицу и дать ее краткий (3-4 предложения) анализ, в котором отразить достоверность статистических оценок, направленность и силу корреляционных связей (между какими признаками связь максимальна, между какими - минимальна).

## *Линейный регрессионный анализ*

Регрессионная модель связывает числовую зависимую переменную (отклик) с одним (простая регрессия) или несколькими (множественная регрессия) независимыми переменными (предикторами, регрессорами). Регрессионным анализом называется поиск такой модели, т.е. поиск некоторой функции, описывающей эту зависимость.

Различают линейные и нелинейные регрессионные модели. Линейная регрессионная модель является линейной функцией от параметров. Линейный регрессионный анализ представляет собой достаточно сложный статистический метод, и здесь мы ограничимся рассмотрением случая одной зависимой переменной $Y$ и одной независимой переменной $X$. Это задача простой

линейной регрессии. Рассмотрим последовательность действий при решении задачи простой линейной регрессии.

Подбор вида модели. Первым шагом является предположение о возможном виде зависимости $Y$ от $X$. Помимо линейной зависимости $Y$ от $X$ ($Y = a + bX$), примерами таких предположений могут быть зависимости: $Y = exp(a + bX)$; $Y = 1/(a + bX)$ и некоторые другие, которые сводятся к линейной модели (например, путем замены переменных), а также полиномиальная зависимость $Y = a + bX + cX^2$. Здесь $a$, $b$, $c$ – неизвестные параметры, которые надо определить по исходным данным. Для подбора вида модельной зависимости $Y$ от $X$ полезно построить и изучить график (так называемый scatterplot), на котором выборочные данные представлены как точки (по одной из осей откладываются значения первой переменной - $X$, а по другой - соответствующие значения второй переменной - $Y$).

Оценка параметров модели. После выбора предполагаемого вида модели, по исходным данным проводится оценка параметров модели (например, коэффициентов $a$, $b$ или $a$, $b$, $c$ в приведенных выше примерах). Значения параметров в случае линейной регрессии находят с помощью метода наименьших квадратов. Использование этого метода обосновано предположением о нормальном распределении случайной переменной $Y$.

Анализ адекватности модели. После построения регрессионной модели необходимо выяснить, насколько хорошо полученная модель описывает имеющиеся данные. С этой целью рассматриваются такие показатели как коэффициент детерминации, значимость оценок параметров, а также целый ряд показателей, основанных на анализе остатков. Остатками

(residuals) называются разности между фактическими значениями зависимой переменной *Y* и модельными, т.е. рассчитанными по подобранной регрессионной функции. Анализ остатков позволяет сказать насколько хорошо подобрана модель и насколько правильно выбран метод оценки параметров. Если построенная регрессионная модель хорошо описывает истинную зависимость, то остатки должны быть независимыми нормально распределенными случайными величинами с нулевым средним. Поэтому необходимо провести проверку выполнения этих условий.

Начинать анализ остатков следует с построения их графиков, на которых можно выявить особенности не учтенные при построении регрессионной модели.

График, на котором по одной из осей откладываются значения независимой переменной *X*, а по другой - соответствующие значения остатков, позволяет увидеть возможные проблемы: неслучайный вид рисунка указывает, что модель неадекватно описывает наблюденные данные. Это говорит о необходимости пересмотра модели (преобразовании или вводе новых переменных, перехода к другому виду модели). Если модель корректна и все предположения выполняются, остатки должны показывать отсутствие какой бы то ни было структуры. Значения остатков должны располагаться случайным образом вокруг нуля; они не должны смещаться ни в положительную, ни в отрицательную сторону.

График, на котором по одной из осей откладываются модельные значения *Y*, а по другой - значения остатков, позволяет судить о постоянстве (гомоскедастичности) или непостоянстве (гетероскедастичности) дисперсии ошибки

(остатков). Если точки нанесены на график неупорядоченно, то дисперсия ошибки - величина постоянная. В противном случае изменчивость остатков меняется с изменением зависимой переменной, что говорит о невыполнении необходимых предположений и о неадекватности модели, о том, что необходимо преобразование переменных.

График, где значения остатков показаны напротив номера строки выборки, в случае, если его форма отлична от случайной, может указывать на наличие автокорреляции или на зависимость данных от времени. Этот график позволяет визуально проверить допущение, что остатки не коррелированны.

Если графики остатков показывают резко отклоняющиеся от модели наблюдения (выбросы), то подобным наблюдениям надо уделять особенно пристальное внимание, так как их присутствие может грубо искажать значение оценок параметров (особенно если используется метод наименьших квадратов) и привести к ошибочным выводам. Устранение эффекта выбросов можно проводить либо путем удаления этих данных из анализируемой выборки (метод цензурирования), либо с помощью применения методов оценивания параметров, устойчивых к подобным грубым отклонениям.

В STATGRAPHICS классические методы регрессионного анализа реализованы в пункте главного меню **«Relate»**. В частности в подпункте **«Simple Regression»** (простая регрессия) реализовано построение модели простой линейной регрессии - методом наименьших квадратов оцениваются параметры линейной и ряда нелинейных (сводящихся к линейной) моделей.

После выбора этого пункта появляется стандартный диалог выбора анализируемых данных (рис.12). В нем в полях Y и X

нужно задать имена колонок, где находятся значения зависимой и независимой переменных. После нажатия на кнопку «ОК» появляется окно результата *«Simple Regression»* (рис.13), содержащее результаты регрессионного анализа.

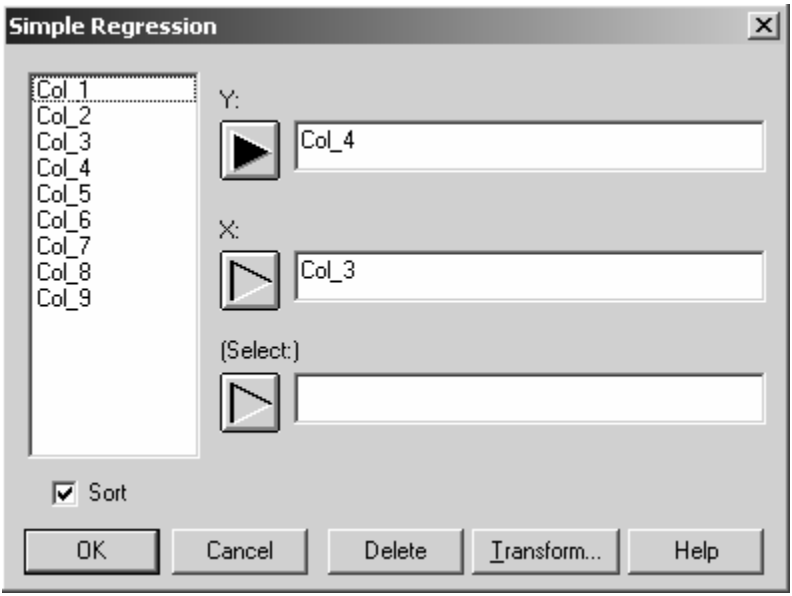



Рис.12. Диалоговое окно для задания зависимой и независимой переменных регрессионного анализа

Regression Analysis - Linear model: Y = a + b*X

Dependent variable: Col_4
Independent variable: Col_3

| Parameter | Estimate | Standard Error | T Statistic | P-Value |
|---|---|---|---|---|
| Intercept | 56.9776 | 4.44831 | 12.8088 | 0.0000 |
| Slope | -0.156698 | 0.26558 | -0.590021 | 0.5565 |

Analysis of Variance

| Source | Sum of Squares | Df | Mean Square | F-Ratio | P-Value |
|---|---|---|---|---|---|
| Model | 355.496 | 1 | 355.496 | 0.35 | 0.5565 |
| Residual | 99053.9 | 97 | 1021.17 | | |
| Total (Corr.) | 99409.4 | 98 | | | |

Correlation Coefficient = -0.0598003
R-squared = 0.357608 percent
Standard Error of Est. = 31.9558

Рис.13. Результаты регрессионного анализа

Дадим пояснение приведенным здесь значениям на примере линейной модели. В первой строке есть запись «Linear Model», и приведен вид модели зависимости «$Y=a+bX$». Вид модели можно изменить, если в окне результата *«Simple Regression»* нажать на правую клавишу мыши, во всплывающем меню выбрать пункт **«Analysis options»,** а в появившемся списке выбрать нужный тип. В следующих строках указывается имя зависимой переменной («Dependent variable») и независимой переменной («Independent variable»).

Далее идут две таблицы. В первой таблице для параметров модели $a$ (Intercept) и $b$ (Slope) даны их оценки (в поле «Estimate»), стандартные ошибки оценок (в поле «Standard Error»), значения статистик для проверки гипотез о равенстве этих параметров нулю и их достигаемые уровни значимости

p-value (в полях «T-Statistic» и «p-value», соответственно). В нашем примере значения p-value говорят о том, что коэффициент *a* значимо отличается от нуля, в то время как коэффициент *b* отличается от нуля незначимо, поскольку для него значение p-value=0.5565 больше 0.1.

Вторая таблица, которая называется «Analysis of Variance», служит для оценки адекватности построенной модели. В случае регрессионного анализа общая вариация отклика *Y* относительно его среднего значения распадается на вариацию, обусловленную моделью («Model») и остаточную вариацию («Residual»), приписываемую случайным ошибкам. Для проверки гипотезы о равенстве коэффициента *b* нулю используется отношение Фишера (F-Ratio) – отношение дисперсии, обусловленной моделью, к дисперсии ошибок (остатков). В таблице приводится полученное значение F-Ratio и достигаемый уровень значимости p-value. В нашем примере значение p-value говорит о том, что коэффициент *b* незначимо отличается от нуля, что указывает на неадекватность модели.

Далее (на рис.13) приведены еще три показателя качества модели. Correlation Coefficient - выборочный коэффициент корреляции зависимой и независимой переменных (чем ближе по модулю к единице, тем лучше модель описывает экспериментальные данные). R-squared - коэффициент детерминации $R^2$, который показывает долю изменчивости *Y*, объясняемую построенной моделью регрессии (чем ближе значение $R^2$ к ста процентам, тем лучше построенная модель описывает данные эксперимента). Standard Error of Estimation – стандартная ошибка оценки – среднеквадратическое отклонение регрессионных остатков.

Основываясь на показателе $R^2$ и коэффициенте корреляции в STATGRAPHICS можно провести сравнение качества, даваемого альтернативными видами регрессионных моделей: линейной и ряда нелинейных (сводящихся к линейной). Для этого нужно нажать на кнопку  - **«Tabular Options»**, а в появившемся окне отметить пункт **«Comparison of alternative model»**. В окне результата появится раздел, где будет представлена таблица сравнения подходящих альтернативных моделей (рис.14). В первом столбце таблицы указан вид модели; во втором - значение коэффициента корреляции; в третьем - значение коэффициента детерминации.

Comparison of Alternative Models

| Model | Correlation | R-Squared |
|---|---|---|
| Reciprocal-X | 0.3413 | 11.65% |
| Logarithmic-X | -0.2546 | 6.48% |
| Square root-X | -0.2039 | 4.16% |
| Square root-Y | -0.1566 | 2.45% |
| Linear | -0.1564 | 2.45% |
| Exponential | <no fit> | |
| Reciprocal-Y | <no fit> | |
| Double reciprocal | <no fit> | |
| Multiplicative | <no fit> | |
| S-curve | <no fit> | |
| Logistic | <no fit> | |
| Log probit | <no fit> | |

Рис.14. Таблица сравнения альтернативных моделей

Важным при анализе качества (адекватности) построенной модели является графическое представление полученных результатов. Для построения графика функции построенной модели зависимости нужно нажать на кнопку  - **«Graphical**

**options»**, а в появившемся диалоговом окне (рис. 15) выбрать **«Plot of Fitted Model»**.

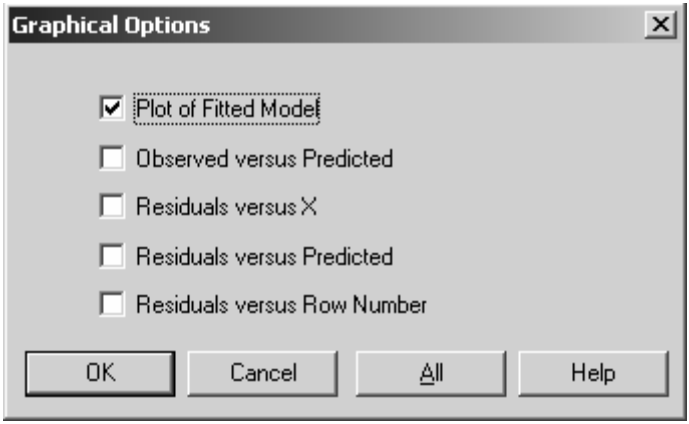



Рис.15. Диалоговое окно выбора графиков

Для того, что бы наглядно увидеть согласие фактических и модельных значений, полезно построить и изучить график **«Observed versus Predicted»**, на котором по одной из осей откладываются фактические значения *Y*, а по другой - соответствующие модельные значения. Чем ближе полученные точки расположены к диагональной прямой, тем выше качество модели. При анализе остатков необходимо рассмотреть графики «**Residuals versus X»**, **«Residuals versus Predicted»**, **«Residuals versus Row Number»**.

Для того, чтобы выполнить проверку остатков на нормальность, их необходимо предварительно вычислить и сохранить в новом столбце таблицы данных. Для этого, нажав на кнопку  - **«Save results»**, в появившемся окне (рис.16) отмечаем «Residuals» и нажимаем «OK». Значения остатков

будут помещены в колонку с именем «RESIDUALS» в таблице данных.

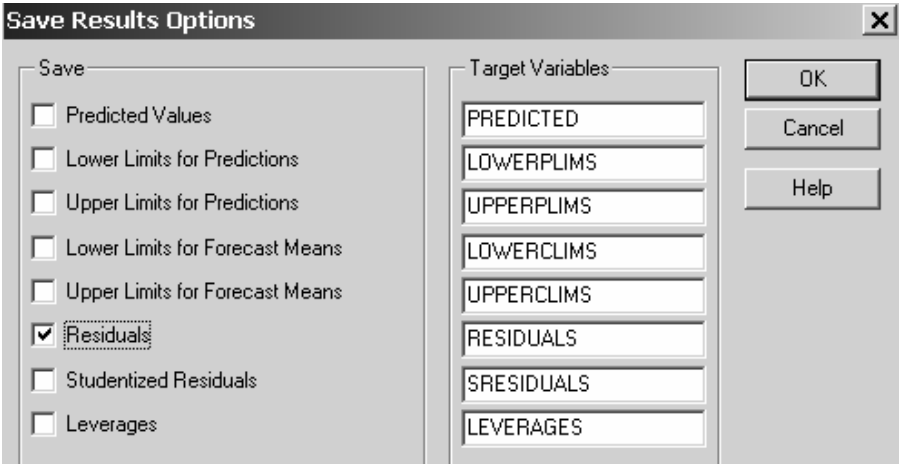



Рис.16. Диалоговое окно сохранения остатков

Кроме рассмотренных моделей в STATGRAPHICS есть возможность построения модели с использованием полиномиальных функций заданной степени ($Y = a + bX + cX^2$, $Y = a + bX + cX^2 + dX^3$ и т.д.). Такая возможность предоставлена в пункте главного меню **«Relate»** подпункте **«Polynomial Regression»**. Сменить степень полинома можно, если в окне результата регрессионного анализа нажать правую клавишу мыши, во всплывающем меню выбрать пункт **«Pane Options»**, и в появившейся строке ввода задать степень полинома.

Также возможно построение линейной регрессионной модели с использованием нескольких независимых переменных: пункт главного меню **«Relate»**, подпункт **«Multiple Regression»** (множественная регрессия).

Сравнение качества моделей, построенных разными методами, можно проводить на основе «скорректированного» значения коэффициента детерминации - R-squared (adjusted for d.f.). Предпочтение следует отдавать той модели, для которой значение R-squared (adjusted for d.f.) окажется максимальным.

## *Задание 4*

Построить простую регрессионную модель для показателя бассейновой эрозии (W080101), где в качестве независимой переменной последовательно рассмотреть распаханность (W17); содержание гумуса (W20); модуль годового стока (W0402); глубину эрозионного расчленения (W07); средний уклон (W06); залесенность (W05).

Для каждой полученной модели зависимости представить:

1. Вид модели (дать формулу), которая наилучшим образом описывает зависимость *Y* от *X* (использовать сравнение альтернативных моделей).

2. Модель с оцененными параметрами.

3. График модельной зависимости.

4. Коэффициент детерминации $R^2$, коэффициент корреляции *Y* и *X*.

5. Отношение Фишера (F-Ratio) и соответствующее значение p-value

6. График, отображающий согласие фактических и модельных значений *Y*.

7. Результаты проверки нормальности распределения остатков.

8. Заключение об адекватности модели (на основе значения коэффициента детерминации, графиков, анализа остатков).

# Анализ временных рядов

Временной ряд отличается от простой выборки данных тем, что в нем выборочные данные упорядочены по времени. Предполагается, что данные содержат регулярную составляющую (одну или несколько) и случайную составляющую, которая затрудняет обнаружение регулярных компонент. Большинство регулярных составляющих временных рядов принадлежит к двум типам: они являются либо *трендом*, либо *сезонной составляющей*. Тренд представляет собой, как правило, монотонную компоненту, изменяющуюся во времени и описывающую общую тенденцию в изменении анализируемого признака. Сезонная составляющая - это регулярная периодическая составляющая, описывающая периодически повторяющиеся колебания анализируемого признака. Оба эти вида регулярных компонент часто присутствуют во временном ряде одновременно.

Анализ временных рядов – это совокупность статистических методов, предназначенных для выявления структуры временных рядов и для их прогнозирования.

## *Проверка гипотезы о «белом шуме»*

Описательные методы анализа временного ряда включают процедуры корректировки и преобразования данных, позволяют обнаружить корреляцию данных, проверить гипотезу о случайности, построить различные графики, которые помогут выявить наличие трендов, периодичности, а также выбросы или ошибки в данных.

При анализе временного ряда в первую очередь проверяется гипотеза о том, что ряд представляет собой реализацию «белого

шума». В этом случае он не содержит в себе никаких регулярных составляющих. В качестве теста, проверяющего гипотезу о случайности элементов ряда, может воспользоваться статистикой Бокса-Пирса (Box-Pierce). В STATGRAPHICS она реализована в пункте главного меню **«Special»** -> **«Time Series Analysis»** -> **«Descriptive Methods»**. В окне результата *«Descriptive Methods»* нажимаем на кнопку  - **«Tabular Options»** и в появившемся диалоге выбираем **«Test for randomness»**. В окне результата в появившемся разделе *«Tests for Randomness»* приведены значения трех критериев, в том числе теста Бокса-Пирса, и их достигаемые уровни значимости (p-value). Если p-value ниже заданного исследователем уровня значимости, то гипотеза о случайности ряда отвергается; если выше, то принимается.

Эту же гипотезу можно проверить с использованием автокорреляционного анализа временного ряда. Автокорреляция (т.е. корреляции между самими членами ряда) вычисляется для определенного лага (смещения временного ряда относительно самого себя). Коэффициент автокорреляции на лаге k измеряет корреляцию между величинами, отстоящими друг от друга по времени на величину k, т.е. между каждым i-м элементом ряда и (i-k)-м элементом. Для вычисления, нажав на кнопку  - **«Tabular Options»** в появившемся диалоге выбираем **«Autocorrelations»** и/или **«Partial Autocorrelation»**, после чего в окне результата в соответствующих разделах появляются таблицы значений коэффициентов автокорреляции и/или частной автокорреляции, полученные на разных лагах. Их графики (коррелограммы) можно построить, нажав на кнопку  - **«Graphical options»** и выбрав в появившемся диалоге пункты

**«Autocorrelation Function»** и/или **«Partial Autocorrelation Function»**.

Если ряд является реализацией «белого шума», то значения его автокорреляционной и частной автокорреляционной функций ни при каком лаге не являются статистически значимыми, т.е. не выходят за пределы доверительных интервалов вокруг нулевого значения коэффициента корреляции. Границы этих доверительных интервалов (при заданной доверительной вероятности) также приведены в таблицах значений автокорреляции и показаны на графиках (рис.17, линии - границы доверительных интервалов).

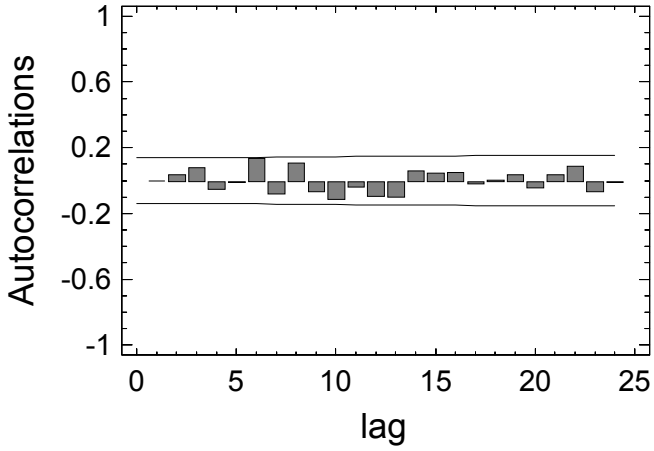



Рис.17. Пример графика значений коэффициентов автокорреляции на разных лагах для временного ряда, представляющего собой реализацию «белого шума»

## *Анализ и выделение тренда*

Выделение тренда может быть произведено двумя способами: с использованием сглаживания и за счет подгонки

функции. Оба метода могут применяться как отдельно друг от друга, так и последовательно: первым шагом является сглаживание, а затем подгонка функции. Цель сглаживания и подгонки функции – отделение длинновременных изменений от кратковременных.

Сглаживание всегда включает некоторый способ локального усреднения данных, при котором несистематические компоненты взаимно погашают друг друга. Сглаживание может осуществляться с использованием самых разных методов. Все эти методы отфильтровывают шум и преобразуют данные в относительно гладкую кривую.

Одним из самых простых методов является метод скользящего среднего - вычисление среднего значения внутри движущегося окна. Каждый член ряда заменяется простым или взвешенным средним n соседних членов, где n - ширина «окна». Главным параметром здесь является размер «окна». Он часто равен периоду гармонических колебаний, которые присутствуют во временном ряду и которые нужно убрать для выявления других менее заметных на первый взгляд компонент.

Сглаживание в STATGRAPHICS осуществляется в пункте главного меню **«Special»**-> **«Time Series Analysis»**-> **«Smoothing»**. Отобразить таблицу со сглаженными данными можно, если в окне результата «*Smoothing*» нажать на кнопку  - **«Tabular Options»** и в появившемся диалоге выбрать опцию **«Data Table»**. Для определения вида сглаживающего метода нужно в окне результата нажать правую клавишу мыши, во всплывающем меню выбрать пункт **«Pane Options»**, и сделать выбор в появившемся окне «Smothing Options» (рис.18). Здесь в

блоках «Smoother1» и «Smoother2» приведены алгоритмы сглаживания, а в блоке «Length of Moving Average» - размер скользящего окна.

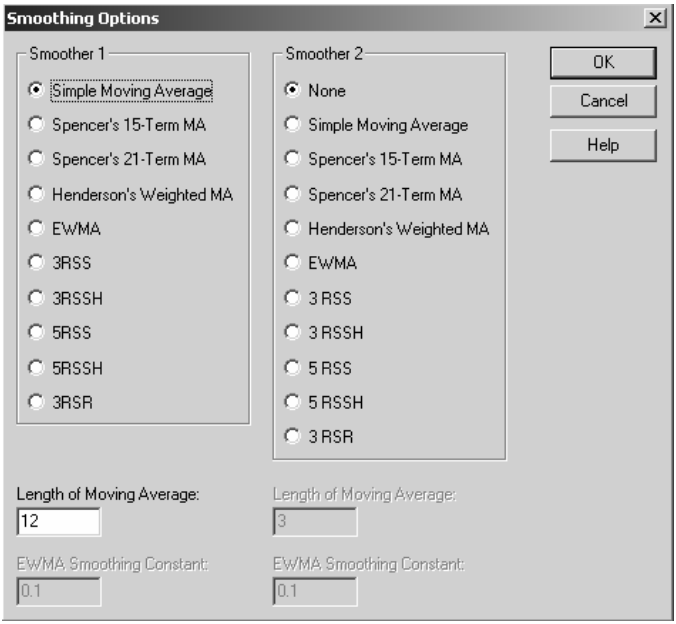



Рис.18. Диалоговое окно для выбора метода сглаживания

Наиболее информативным при анализе тренда является построение графика сглаженного ряда, по общему виду которого можно судить о модели тренда. График сглаженного ряда строится при нажатии на кнопку  - **«Graphical options»** и выборе в появившемся диалоге пункта **«Time Sequence Plot»**. По виду графика сглаженного ряда можно высказать предположения о виде функции, описывающей тренд. Для того, чтобы выполнить построение модели тренда, сглаженные данные необходимо предварительно сохранить в новом столбце таблицы данных. Для

этого, нажав на кнопку  - **«Save results»**, в появившемся окне отмечаем «Smooth» и нажимаем «OK». Сглаженные значения временного ряда будут помещены в колонку с именем «SMOOTH» в таблице данных.

Для примера на рис.19 приведен график сглаженного (усреднением за 5 лет) временного ряда значений среднемесячных концентрациях сульфатов в атмосферных осадках за 50 лет.

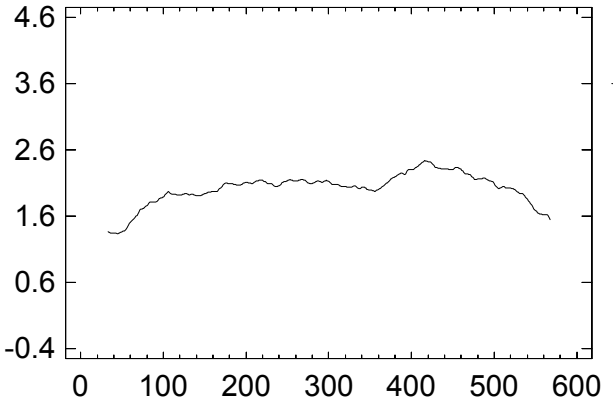



Рис.19. График сглаженного временного ряда

Подгонка функции. Подбор функции и построение модели тренда осуществляется методами регрессионного анализа (см.выше), где зависимой переменной будет сглаженный временной ряд, а независимой переменной – время. Многие монотонные временные ряды можно хорошо приблизить линейной функцией. Если же имеется явная нелинейная компонента, то данные могут потребовать предварительного преобразования с использованием логарифмического или других преобразований (также см.выше).

## *Выявление регулярной периодической составляющей и сезонная декомпозиция*

Кроме тренда, описывающего длинновременные изменения, во временном ряду, как правило, присутствует сезонная составляющая, влияние которой также необходимо оценить. В общем случае периодическая зависимость может быть формально определена как корреляционная зависимость порядка k между каждым i-м элементом ряда и (i-k)-м элементом, и измерена с помощью автокорреляции. Таким, образом, для выявления периодических зависимостей в данных (например, сезонных колебаний) и определения их периодов может помочь анализ коррелограмм - графиков автокорреляционной и частной автокорреляционной функций.

На рис.20 приведена коррелограмма временного ряда среднемесячных концентраций сульфатов в атмосферных осадках, наблюдаемых в течении 50 лет. Видно, что анализируемый ряд не является реализацией «белого шума», поскольку большинство значений автокорреляций выходят за пределы доверительных интервалов (черные линии), а содержит регулярную гармоническую составляющую с периодом равным 12 (период - число единиц времени, требующихся на полный цикл).

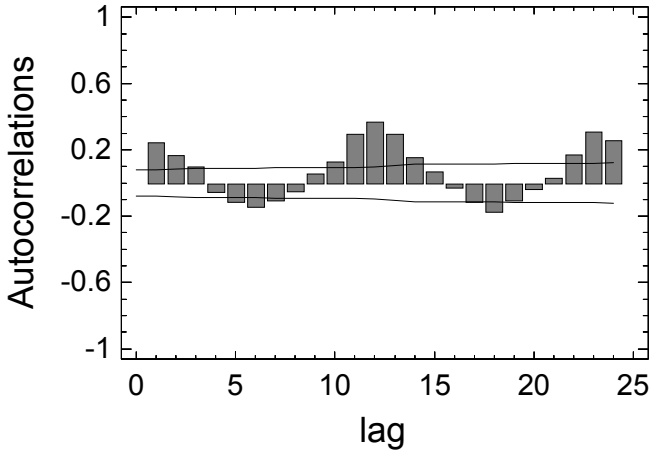



Рис.20. График значений коэффициентов автокорреляции (коррелограмма) временного ряда среднемесячных концентраций сульфатов в атмосферных осадках на метеостанции Мудьюг

Для выявления гармонических составляющих полезен также анализ периодограммы – оценки спектральной плотности мощности. График периодограммы можно построить, нажав на кнопку  - **«Graphical options»** и выбрав в появившемся диалоге пункт **«Periodogram»**. Он показывает квадраты амплитуд синусоид на различных частотах напротив этих частот. Частота - это число циклов в единицу времени (где каждое наблюдение составляет одну единицу времени). Например, частота 0.1 соответствует значению 10.

На рис.21 приведена периодограмма, построенная по тем же данным, на которой четко виден пик на частоте, соответствующей временному периоду 12. Из этого следует, что вклад сезонных колебаний в изменчивость данных много выше вклада других составляющих.

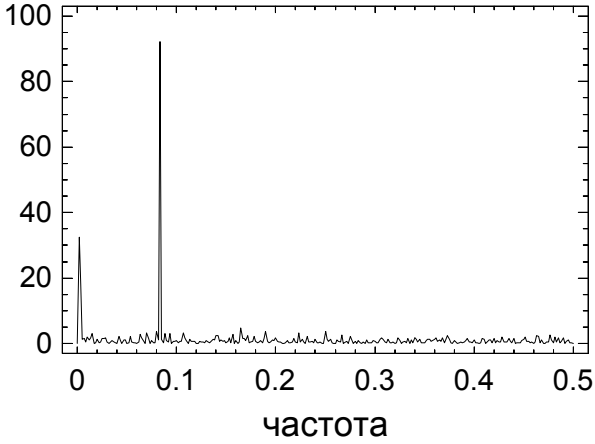



Рис.21. Периодограмма временного ряда среднемесячных концентраций сульфатов в атмосферных осадках на метеостанции Мудьюг

Для изучения и выделения сезонной составляющей существует целый ряд методов, основной из которых спектральный анализ, а наиболее простой - построение специальных показателей, которые называются сезонными индексами. Спектральный анализ представляет собой достаточно сложную для понимания студентами нематематических специальностей систему анализа временного ряда. Если говорить упрощенно, то это описание временного ряда с использованием набора гармонических функций. Построение сезонных индексов намного более простой подход при выделении сезонной компоненты и ее количественного описания. При этом точность прогнозов полученных с использованием сезонных индексов сопоставима с точностью получаемой при использовании спектрального анализа.

Совокупность сезонных индексов отражает сезонную волну. При этом в рассмотрение вводятся две модели взаимодействия компонент ряда: аддитивная и мультипликативная модели. В случае аддитивной модели составляющие временного ряда складываются, то есть имеет место следующая зависимость:

$$x(t) = tr(t) + s(t) + \varepsilon(t),$$

а в случае мультипликативной модели составляющие ряда перемножаются, то есть имеет место:

$$x(t) = tr(t) \cdot s(t) \cdot \varepsilon(t),$$

где $x(t)$ – значение временного ряда в момент времени $t$; $tr(t)$ – значение тренда в момент времени $t$; $s(t)$ – сезонная составляющая ряда в момент времени $t$; $\varepsilon(t)$ – случайная составляющая в момент времени $t$.

В зависимости от принятой модели сезонные индексы рассчитываются по-разному. Так для аддитивной модели сезонный индекс рассчитывается по следующей формуле:

$$s_i = \frac{1}{m+1} \sum_{l=0}^{m} (x_{i+lp} - tr_{i+lp}),$$

а при использовании мультипликативной модели по формуле:

$$s_i = \frac{1}{m+1} \sum_{l=0}^{m} \left(\frac{x_{i+lp}}{tr_{i+lp}} * 100\%\right),$$

где $p$ - период последовательности $s(t)$; $m$ - количество элементов ряда, принадлежащих одному сезону; $i$ - сезон ($1 < i < p$); $x_{i+lp}$, $tr_{i+lp}$ - значения временного ряда и его трендовой составляющей для $l$-ого элемента $i$-ого сезона.

Например, для нашего временного ряда, отражающего среднемесячные концентрации сульфатов в атмосферных осадках, в качестве периода может выступать количество месяцев в году - $p = 12$; в качестве сезонов $i$ - месяцы; $m$ – это количество лет во временном ряду.

В STATGRAPHICS сезонная декомпозиция производится в пункте главного меню **«Special»-> «Time Series Analysis»-> «Seasonal Decomposition»**. Цель декомпозиции – разложить временной ряд на три составляющие: тренд, сезонные колебания и случайную компоненту. При выборе данного пункта меню появляется диалог представленный на рис. 22. Здесь необходимо задать временной интервал, с которым измерялись данные (раз в год, в месяц, в день и пр.), в строке «Once every» задается частота, с которой нужно анализировать данные (по умолчанию стоит 1, то есть анализируются все данные). В этом же диалоге нужно задать период сезонных колебаний в строке «Seasonality». Далее после нажатия на «OK» появляется окно результата «*Seasonal Decomposition*», в котором нажав на кнопку  - **«Tabular Options»** и выбрав **«Seasonal Indices»** получаем раздел с сезонными индексами. При выборе пункта **«Data Table»** появится раздел с таблицей, в которой отражены шаги сезонной декомпозиции. Столбец под названием «Trend-cycle» показывает результаты сглаживания методом скользящего среднего; столбец «Seasonality» содержит разность между значениями ряда и тренда; «Irregular» - это случайная составляющая временного ряда.

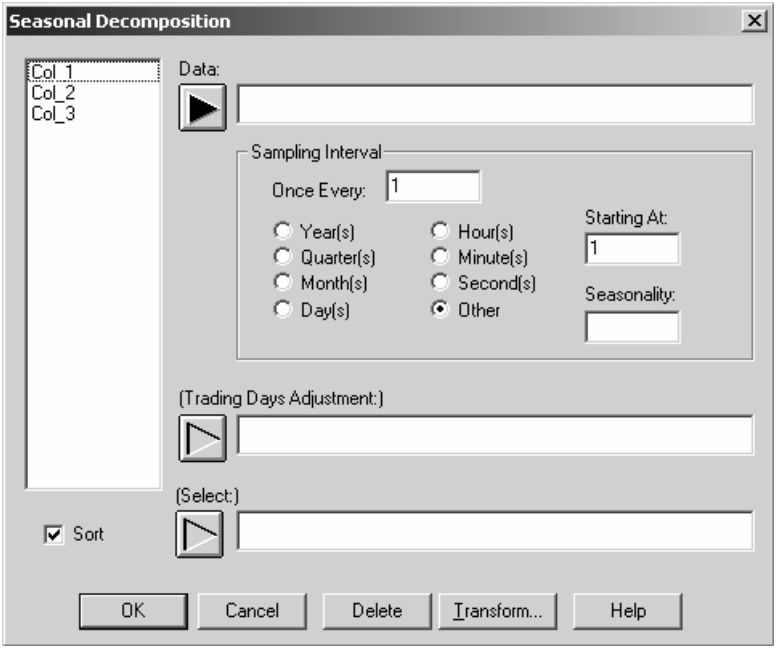



Рис.22. Диалоговое окно для задания параметров сезонной декомпозиции временного ряда

## *Задание 5*

Проанализировать временной ряд среднемесячных концентраций сульфатов в атмосферных осадках на метеостанции Мудьюг. Представить отчет по результатам проведенного анализа, в который должны быть включены:

1. График исходного временного ряда.
2. Рассчитанная статистика Бокса-Пирса.
3. График коррелограммы.
4. График периодограммы.
5. График сглаженной функции.

6. Уравнение, описывающее тренд данного временного ряда. Привести значение коэффициента детерминации.
7. Таблица сезонных индексов.
8. График прогноза поведения ряда на 5 лет.

В отчете отразить ответы на следующие вопросы: является ли ряд реализацией случайного процесса (аргументировать свой ответ); присутствуют ли в нем гармонические колебания (сколько их, с каким периодом, аргументировать ответ); почему было проведено сглаживание, окном какого размера оно было произведено и почему; какая функция описывает тренд данного временного ряда и почему.

# Список литературы


1. Гайдышев И. Анализ и обработка данных / Спб: Питер, 2001. - 750с.
2. Кендалл.М., Стюарт.А. Многомерный статистический анализ и временные ряды / М.: Наука, 1976. – 736 с.
3. Кобзарь А. И. Прикладная математическая статистика / М.: Физматлит, 2006. - 816 с.
4. Пузаченко Ю.Г. Математические методы в экологических и географических исследованиях / Изд-во: ИЦ Академия, 2004. - 416 с.
5. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере / Под ред. В. Э. Фигурнова. - М.: ИНФА-М, 1998. - 528 с.
6. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере / М.: ИНФА-М, 2003. - 544 с.