Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Н. Г. ГАЛИНОВСКИЙ, С. А. ЗЯТЬКОВ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Рекомендовано Учебно-методическим объединением по естественно-научному образованию в качестве пособия для студентов учреждений высшего образования, обучающихся по специальности 1-31 01 01 «Биология» (по направлениям), направление специальности 1-31 01 01-02 «Биология (научно-педагогическая деятельность)»

2-е издание, стереотипное

Гомель
ГГУ им. Ф. Скорины
2021

УДК 57.08:004
ББК 28с51
Г157

Рецензенты:

кандидат биологических наук В. С. Бирг;
кандидат биологических наук Ж. Е. Мелешко

Рекомендовано к изданию научно-методическим советом учреждения образования «Гомельский государственный университет имени Франциска Скорины»

Галиновский, Н. Г.

Г157 Информационные технологии в биологических исследованиях : пособие / Н. Г. Галиновский, С. А. Зятьков ; М-во образования Республики Беларусь, Гомельский гос. ун-т им. Ф. Скорины. – 2-е изд., стер. – Гомель : ГГУ им. Ф. Скорины, 2021. – 199 с.

ISBN 978-985-577-788-6

Пособие ставит своей целью оптимизировать учебно-познавательную деятельность студентов по усвоению материала курса «Информационные технологии в биологических исследованиях». Может быть использовано как при проведении лабораторных занятий, так и для самостоятельной подготовки при работе над курсовыми и дипломными работами и проектами.

Адресовано студентам учреждений высшего образования, обучающимся по специальности 1-31 01 01 «Биология» (по направлениям), направление специальности 1-31 01 01-02 «Биология (научно-педагогическая деятельность)».

УДК 57.08:004
ББК 28с51

ISBN 978-985-577-788-6

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Современная биология не может развиваться без применения основ математической статистики. Математика требуется, прежде всего, при описании биологических множеств, популяций, штаммов, сортов, пород, линий, посевов, стад, подопытных групп. Математические методы необходимы для исчерпывающего извлечения информации о типичных объектах, их разнообразии, структуре этого разнообразия, о системах биологических взаимоотношений и взаимодействий, о разных биоценозах, о влияниях разных факторов на биологические объекты, развивающиеся в различных условиях.

Некоторые биологические вопросы не могут быть решены без применения специальных математических методов. К таким вопросам относятся сравнение выборочных групп по изучаемым показателям и определение достоверности результатов такого сравнения с заданной вероятностью безошибочных прогнозов, определение достаточной численности подопытных объектов, измерение силы влияния различных факторов на биологические процессы и явления.

Актуальность изучения данной учебной дисциплины связана с необходимостью научной корректности и экономической обусловленности выводов и прогнозов в биологии, сельском хозяйстве и медицине при помощи средств вычислительной техники.

Целью представленного учебного пособия является усвоение студентами методических подходов к статистической обработке данных и обучение современным методам обработки исходной информации с использованием персональных компьютеров.

Учебная дисциплина «Информационные технологии в биологических исследованиях» базируется на ранее полученных студентами знаниях по таким дисциплинам, как «Физика», «Высшая математика» и связана с такими смежными дисциплинами, как «Ботаника», «Зоология», «Химия».

Учебное пособие построено на выполнении конкретных заданий в каждой из 9 тем. Задания содержат определённый набор данных, взятых как из литературы, так и непосредственно из исследований, проводимых авторами пособия. При рассмотрении каждой из тем предварительно подробнейшим образом, по шагам алгоритма, рассматривается проведение того или иного анализа с использованием программных пакетов Excel и STATISTICA.

ТЕМА 1. ЗНАКОМСТВО С ПРОГРАММНЫМ ПАКЕТОМ STATISTICA 7.0

1.1 Создание нового файла данных в STATISTICA 7.0.
1.2 Основы графического отображения данных в STATISTICA 7.0.

1.1 Создание нового файла данных в STATISTICA 7.0

Рассмотрим создание нового файла с данными на примере таблицы стоимости рекламы в газете (таблица 1.1).

Таблица 1.1 – Стоимость рекламы в газете

Длина (мм)	Ширина (мм)	Площадь (мм. кв.)	Цена (руб.)
47	35	1 645	1 446 000
47	73	3 431	2 768 000
47	111	5 217	3 974 000
47	149	7 003	5 147 000
47	209	9 823	6 290 000
47	225	10 575	7 537 000
47	263	12 361	8 828 000
47	301	14 147	10 260 000

Для удобства обучения тем или иным возможностям изучаемых программ мы приводим процесс создания нового файла или проведения анализа в алгоритмическом пошаговом режиме.

Шаг 1. Создание электронной таблицы.

Выберите команду **New** (Новый) из меню **File** (Файл). Эта команда доступна также по комбинации клавиш **CTRL+N** (рисунок 1.1А). На экране появится диалоговое окно **Create New Document** (Создайте новый документ). В данном окне нас интересует закладка **Spreadsheet** (Таблица) (рисунок 1.1Б), которая имеет следующие элементы:

– Number of variables (Количество вариант);

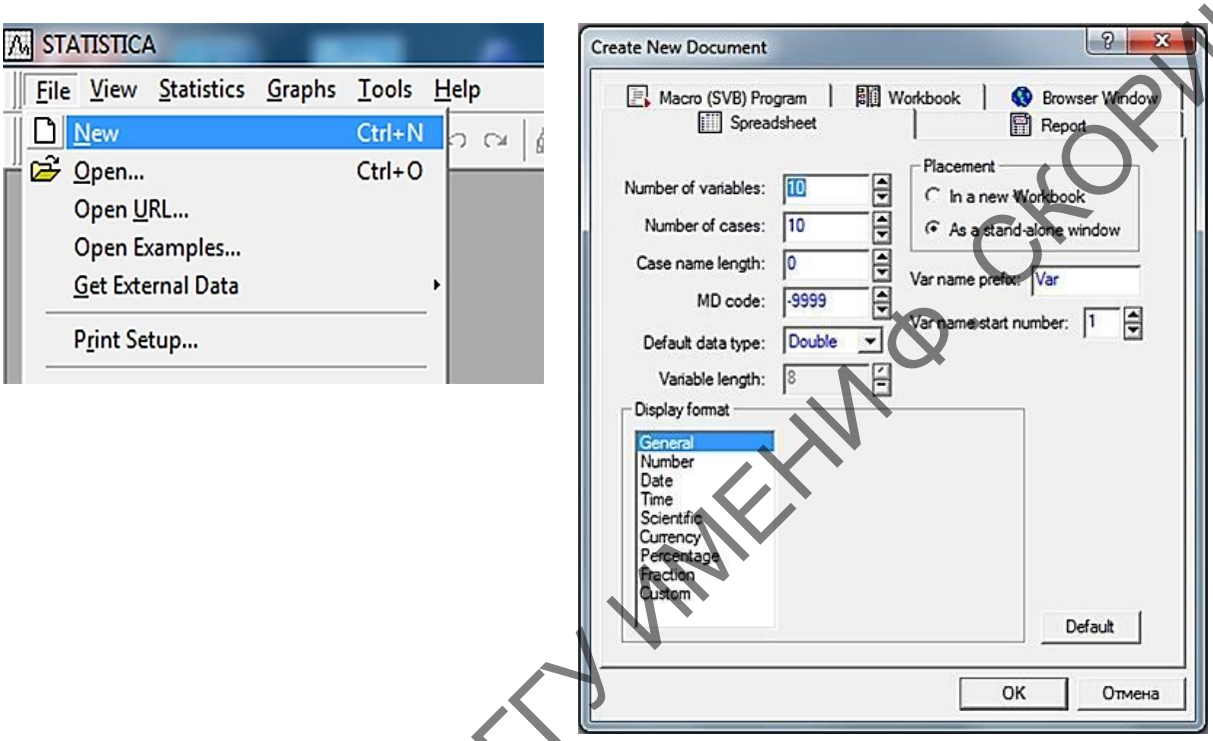
– Number of cases (Количество значений варианты);

– Case name length (Длина названия столбца значений варианты);

– MD code (Отсутствующий код данных по умолчанию для пустых ячеек или конкретных значений, которые вы намерены игнорировать при расчетах);

– Default data type (Тип данных по умолчанию);

– **Variable length** (Длина текстовой переменной в символах);
– **Display format** (Отображение данных в выбранном формате);
– **Placement** (Расположение таблицы данных):
In a new Workbook (В новой рабочей книге);
As a stand-alone window (Как отдельное окно);
– **Var name prefix** (Название переменных по умолчанию);
– **Var name start number** (Начальный номер списка).



А Б

А – опция меню **New**;
Б – диалоговое окно **Create New Document**

Рисунок 1.1 – Создание нового файла данных в STATISTICA 7.0

Шаг 2. Задание свойств электронной таблицы.

В диалоговом окне **Create New Document** (Создайте новый документ) необходимо поставить курсор в ячейку **Number of variables** (Количество вариант) и указать количество столбцов в нашей таблице, т. е. 4. Далее в ячейке **Number of cases** (Количество значений варианты) нужно указать количество строк в таблице. В нашем случае это 8 (без учёта заголовка!). Тип данных оставляем по умолчанию, т. е. **Double** (Двоичный). Остальные показатели также оставляем без изменений и далее необходимо нажать кнопку **OK**. Появится наша рабочая таблица (рисунок 1.2).

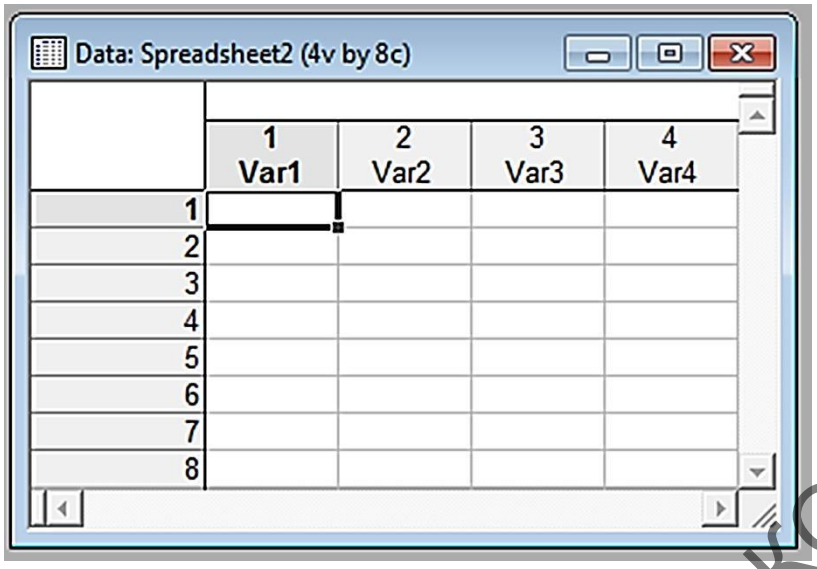


Рисунок 1.2 – Заготовка рабочей таблицы

Для придания нашей таблице соответствующего вида необходимо озаглавить столбцы (варианты). Для этого нужно дважды кликнуть мышкой по серому полю заголовка столбца, после чего перейти в диалоговое окно варианты (рисунок 1.3).

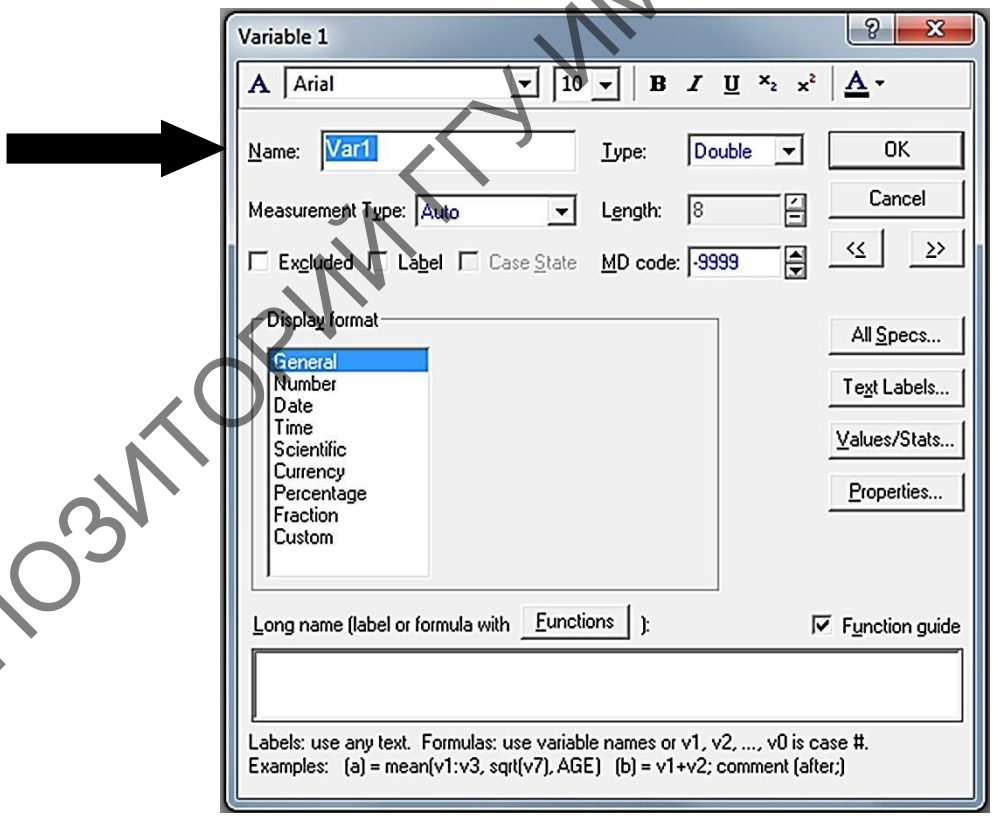
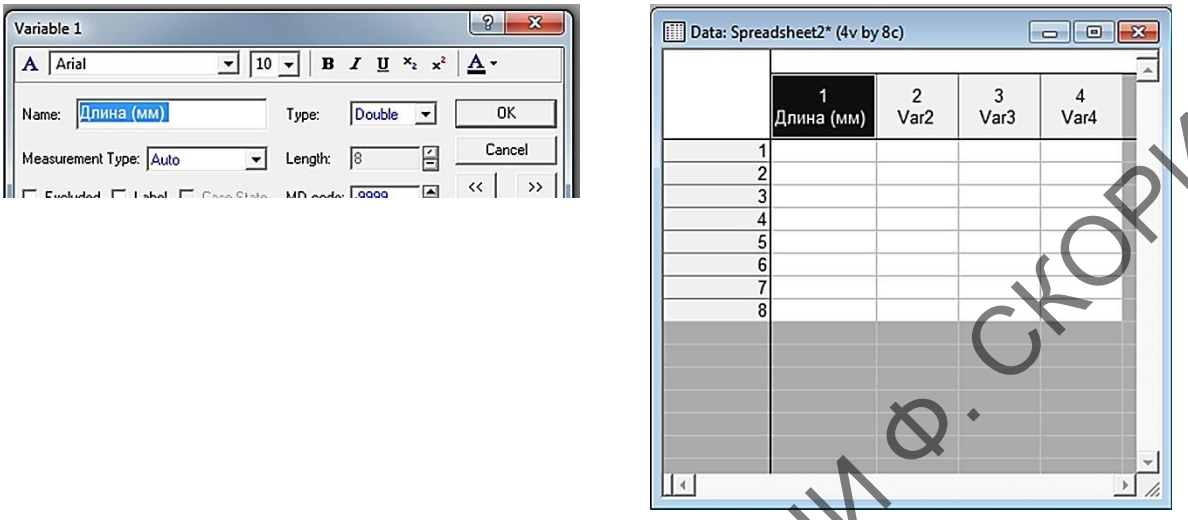


Рисунок 1.3 – Диалоговое окно **Variable 1**

Для переименования переменной необходимо поставить курсор в поле **Name** (Имя), выделить старое название (рисунок 1.3), затем ввести необходимое новое название переменной, в нашем случае – «Длина (мм)» (рисунок 1.4А) и нажать **OK**. Результат отображён на рисунке 1.4Б.



А Б

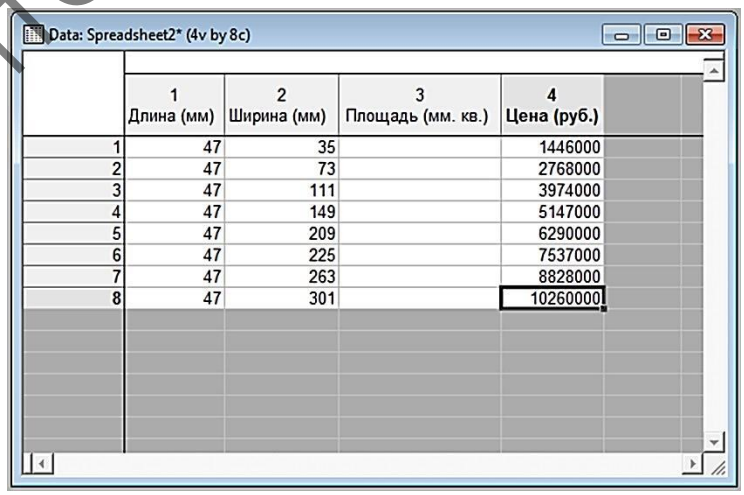
А – изменение названия в поле **Name**; Б – изменённый заголовок в таблице

Рисунок 1.4 – Переименование переменных в STATISTICA 7.0

Аналогично переименовываем все остальные столбцы.

Шаг 3. Заполнение электронной таблицы данными.

Вносим в рабочую таблицу данные из таблицы 1.1, за исключением переменной «Площадь (мм. кв.)». Рабочая таблица будет иметь вид как на рисунке 1.5.

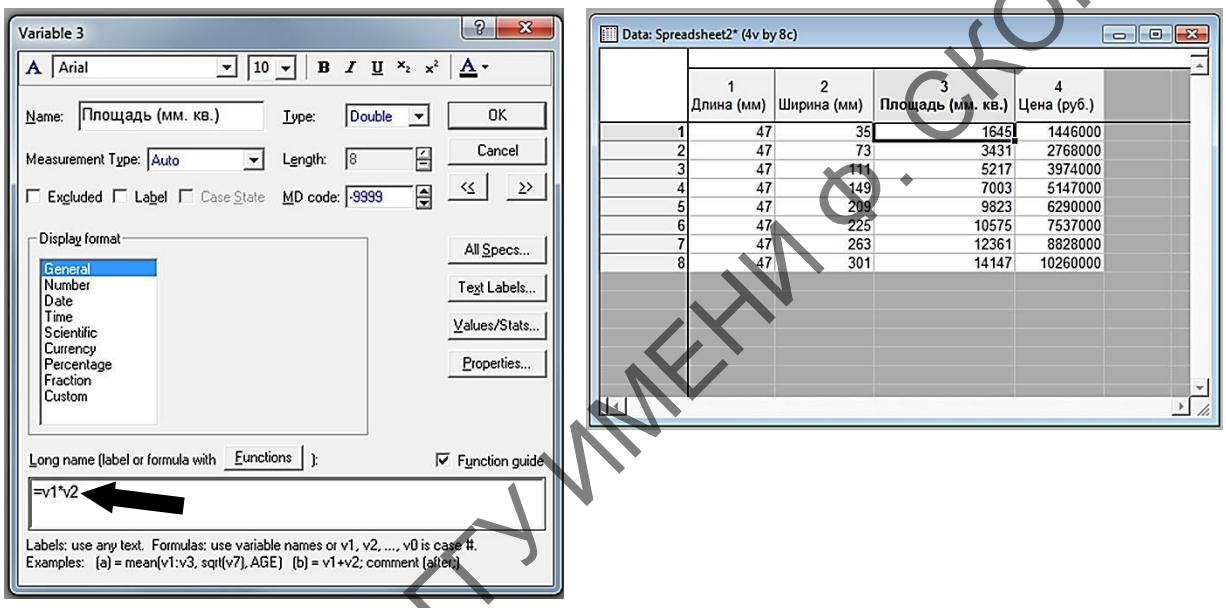


	1 Длина (мм)	2 Ширина (мм)	3 Площадь (мм. кв.)	4 Цена (руб.)
1	47	35		1446000
2	47	73		2768000
3	47	111		3974000
4	47	149		5147000
5	47	209		6290000
6	47	225		7537000
7	47	263		8828000
8	47	301		10260000

Рисунок 1.5 – Сформированная таблица данных в STATISTICA 7.0

Для заполнения данных переменной «Площадь (мм. кв.)» воспользуемся возможностями программного пакета STATISTICA 7.0, который рассчитает площадь по заранее введённой нами формуле.

Для этого необходимо дважды кликнуть по заголовку столбца «Площадь (мм. кв.)», выйти в диалоговое окно третьей переменной и в поле **Long name (label or formula with Functions)** (Длинное имя (метка или формула с функцией)) ввести формулу расчёта площади, умножив данные первой переменной на данные второй переменной: «**=v1*v2**» (рисунок 1.6А), в результате таблица будет полностью заполнена (рисунок 1.6Б).



А Б

А – ввод формулы; Б – окончательно сформированная таблица

Рисунок 1.6 – Ввод формул в свойства переменных в STATISTICA 7.0

1.2 Основы графического отображения данных в STATISTICA 7.0

1.2.1 Полигон распределения и гистограмма распределения

Программный пакет STATISTICA 7.0 обладает очень широкими возможностями графического отражения информации и результатов проведённого анализа. Эти возможности заложены как внутри блоков соответствующего анализа, так и собраны отдельно в меню **Graphs** (Графики) (рисунок 1.7).

Рассмотрим способы построения некоторых типов графиков и диаграмм на конкретном примере экспериментальных данных замера коренных зубов ископаемого млекопитающего (таблица 1.2).

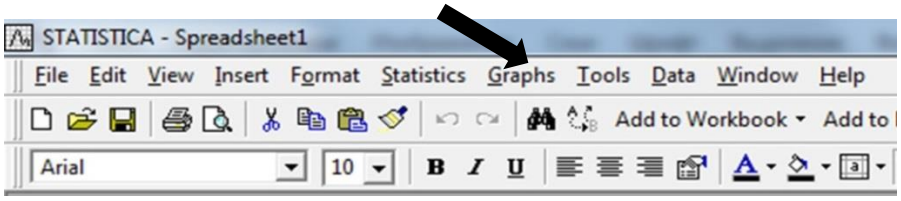


Рисунок 1.7 – Пункт меню **Graphs** в STATISTICA 7.0

Таблица 1.2 – Ширина верхнего последнего моляра у ископаемого млекопитающего

5,8	6,2	6,3	6,1	6,1	5,7	6,2	6,1	5,9
6,5	6,0	6,1	5,8	6,3	6,2	5,7	6,1	6,0
6,2	5,4	5,9	6,0	5,7	5,9	5,7	5,7	6,0
6,1	6,7	6,2	6,5	6,2	6,1	5,9	5,9	6,1

Шаг 1. Составление таблицы с данными.

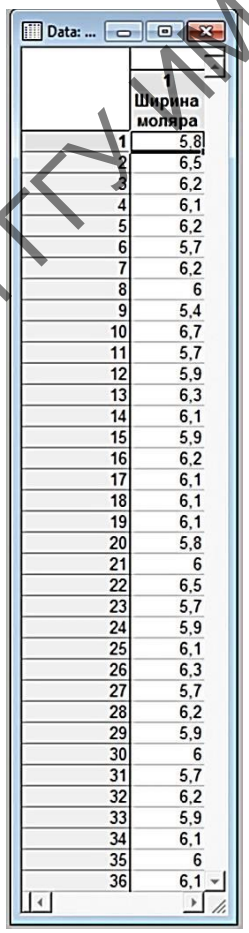


Рисунок 1.8 – Электронная таблица переменной «Ширина моляра»

Составим электронную таблицу данных, учитывая, что переменная у нас будет только одна – «Ширина моляра». Соответствующим образом зададим и имя переменной (рисунок 1.8).

Шаг 2. Добавление переменных.

В целом для построения полигона распределения необходимо предварительно из имеющихся данных составить вариационный ряд (значения признака и частоты их встречаемости). В связи с этим необходимо добавить ещё две переменные (два столбца) в нашу таблицу.

Для этого следует подвести курсор к заголовку переменной «Ширина моляра» и нажать правую кнопку мыши (рисунок 1.9А). Далее в контекстном меню выбрать пункт **Add variables** (Добавить переменные) и в появившемся диалоговом окне в поле **How many** (Сколько) указать количество добавляемых переменных (в нашем случае – 2), а в поле **After** (После) и их положение в таблице (в нашем случае – «Ширина моляра»). Остальное оставляем без изменений (рисунок 1.9).

В таблице появятся 2 новых столбца с названиями **NewVar1** и **NewVar2**. Необходимо их переименовать в «Ширина» и «Частота» соответственно.

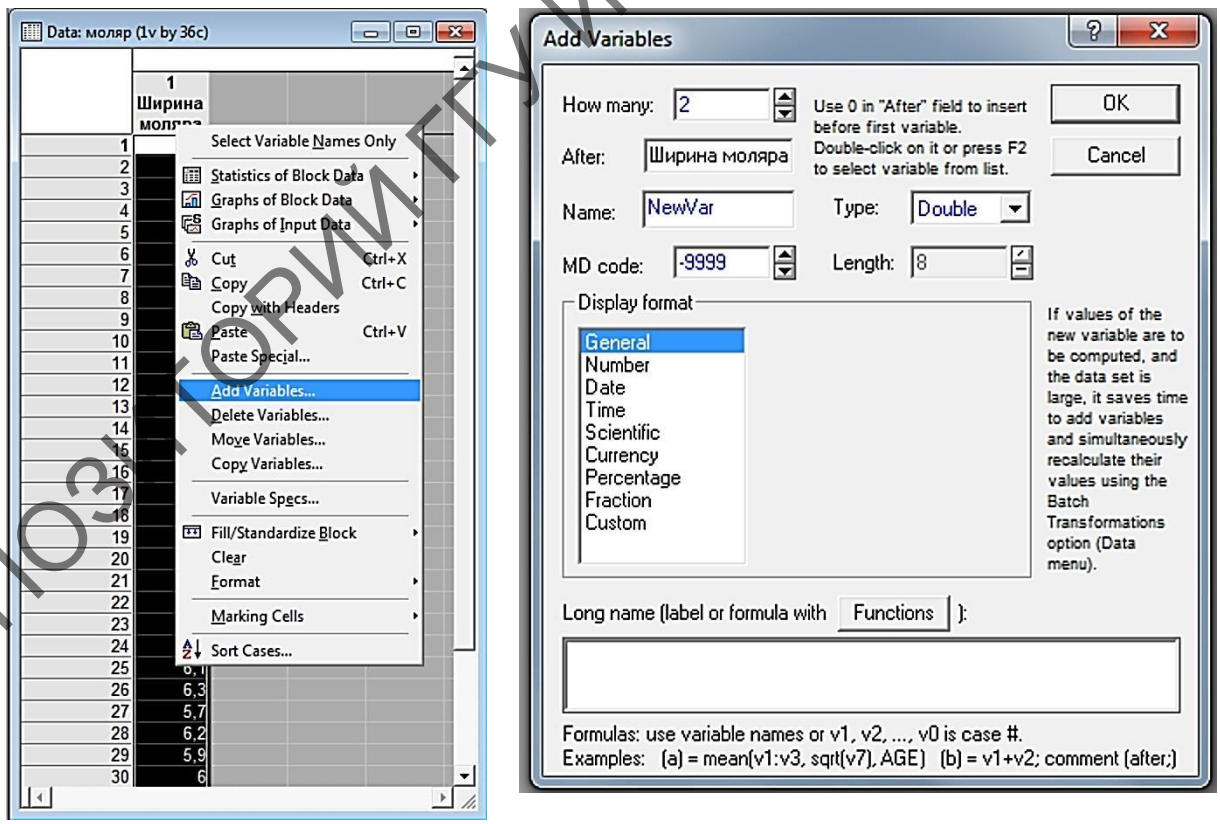


Рисунок 1.9 – Добавление переменных в таблицу

Шаг 3. Расчёт частот и построение вариационного ряда.

Теперь всё готово для формирования вариационного ряда. Для этого необходимо зайти в пункт главного меню **Statistics** (Статистические процедуры) и выбрать в нём модуль **Basic Statistics/Tables** (Основные статистические показатели / Таблицы), и далее – опцию **Frequency tables** (Таблицы частот). В появившемся диалоговом окне нужно указать, какую переменную мы собираемся анализировать. Чтобы это сделать, следует в данном и последующих анализах в дальнейшем нажать кнопку **Variables** (Переменные) (рисунок 1.10).

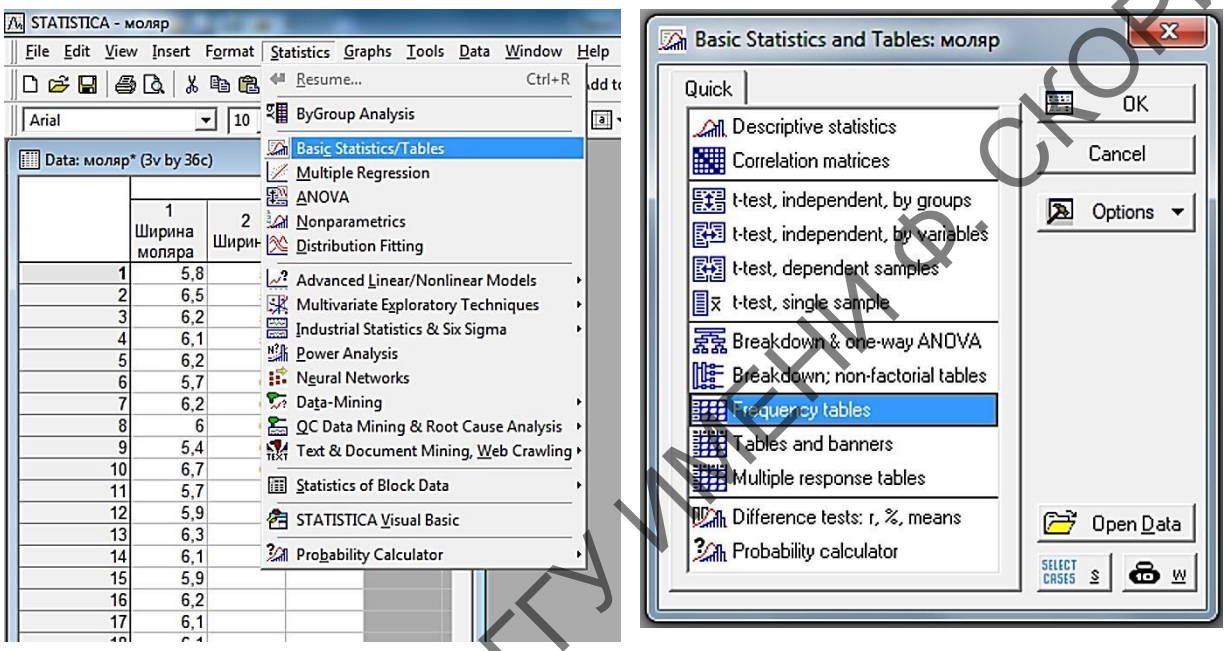


Рисунок 1.10 – Меню **Basic Statistics/Tables**

При нажатии на нее появится окно **Select the variables for the analysis** (Выбор переменных для анализа). Нужно указать программе, что необходимо обработать переменную «Ширина моляра», либо дважды кликнув по пункту «Ширина моляра», либо выделив его курсором и нажав кнопку **ОК**. Увидеть результаты можно тремя способами:

– нажать кнопку **Summary: Frequency tables** (Результат: Таблицы частот);

– нажать кнопку **Summary** (Результат);

– нажать клавишу **Enter** (Ввод) на клавиатуре.

В итоге программа покажет таблицу, представленную на рисунке 1.11.

Category	Frequency table: Ширина моляра (моляр)			
	Count	Cumulative Count	Percent	Cumulative Percent
5,4000000000	1	1	2,77778	2,7778
5,7000000000	5	6	13,88889	16,6667
5,8000000000	2	8	5,55556	22,2222
5,9000000000	5	13	13,88889	36,1111
6	4	17	11,11111	47,2222
6,1000000000	8	25	22,22222	69,4444
6,2000000000	6	31	16,66667	86,1111
6,3000000000	2	33	5,55556	91,6667
6,5000000000	2	35	5,55556	97,2222
6,7000000000	1	36	2,77778	100,0000
Missing	0	36	0,00000	100,0000

Рисунок 1.11 – Таблица частот переменной «Ширина моляра»

В этой таблице имеются следующие столбцы:

– **Category** (Категория) – содержит ранжированные значения анализируемой переменной, отмеченные в выборке. В нашем случае ширина моляра изменялась от 5,4 до 6,7;

– **Count** (Счет) – приведены частоты, с которыми встречались отмеченные значения переменной (значению ширины моляра 5,4 соответствует одно значение, 5,7 – 5 значений, 5,8 – 2 значения и т. д.) ;

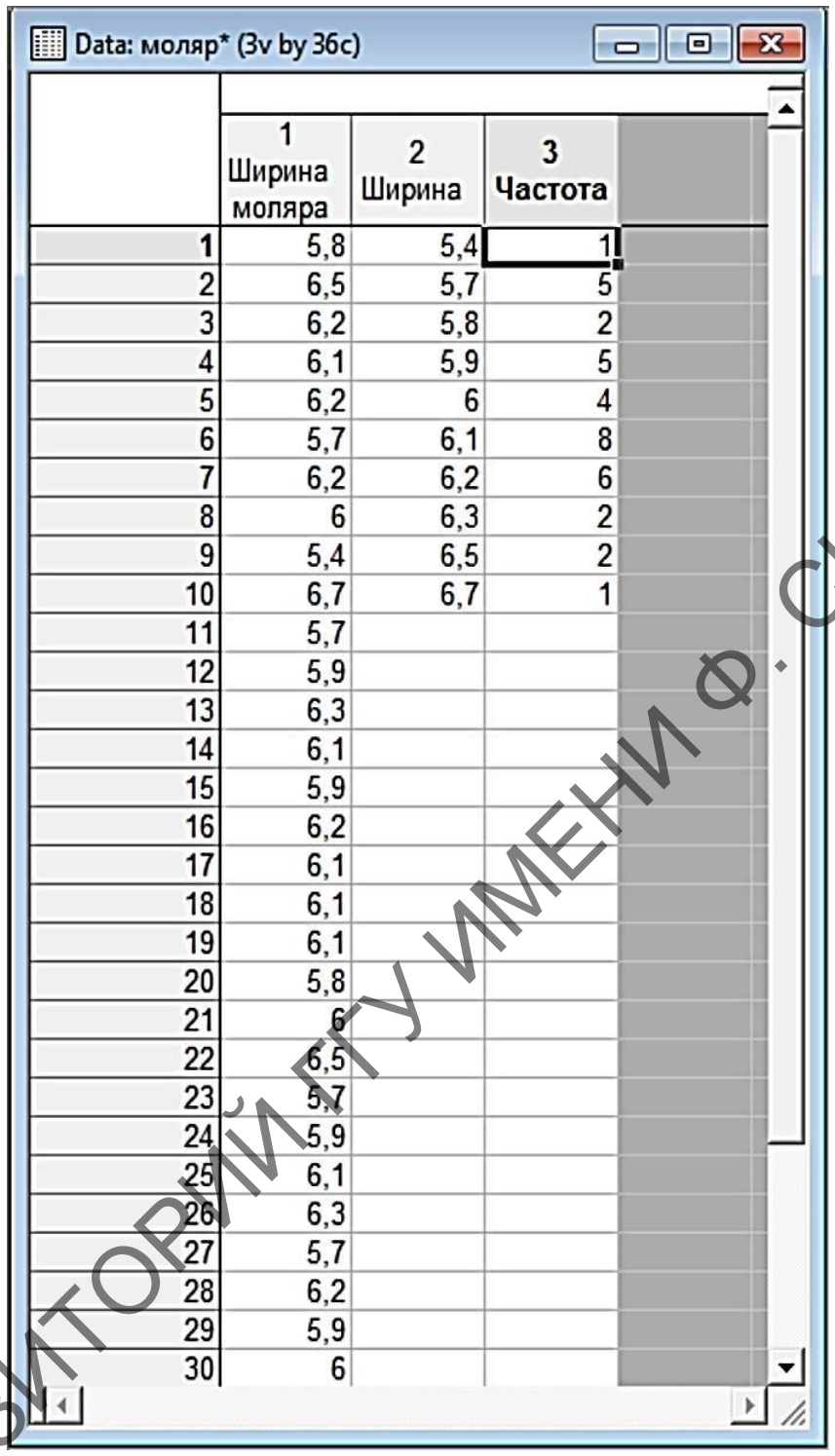
– **Cumulative count** (Кумулятивный счёт) – накопленные частоты;

– **Percent** (Процент) – процент, который составляет каждая из частот от общего числа наблюдений;

– **Cumulative percent** (Кумулятивный процент) – накопленные процентные доли частот;

– **Missing** (Отсутствующие) – имеет отношение к неотмеченным в выборке значениям переменной. В нашем случае таковых нет, поэтому на пересечении столбца **Count** и строки **Missing** отображён 0.

Для дальнейшего построения графика распределения необходимо перенести полученный результат из таблицы частот в нашу таблицу любым способом: либо набрав вручную, либо скопировав через буфер обмена. В итоге таблица будет иметь вид, показанный на рисунке 1.12.



	1 Ширина моляра	2 Ширина	3 Частота
1	5,8	5,4	1
2	6,5	5,7	5
3	6,2	5,8	2
4	6,1	5,9	5
5	6,2	6	4
6	5,7	6,1	8
7	6,2	6,2	6
8	6	6,3	2
9	5,4	6,5	2
10	6,7	6,7	1
11	5,7		
12	5,9		
13	6,3		
14	6,1		
15	5,9		
16	6,2		
17	6,1		
18	6,1		
19	6,1		
20	5,8		
21	6		
22	6,5		
23	5,7		
24	5,9		
25	6,1		
26	6,3		
27	5,7		
28	6,2		
29	5,9		
30	6		

Рисунок 1.12 – Подготовленная электронная таблица переменной «Ширина моляра»

Шаг 4. Построение полигона распределения.

Для построения полигона распределения в пункте главного меню **Graphs** (Графики) необходимо выбрать подпункт **2D Graphs** (Двухмерные графики) и **Line plots (Variables)** (Линейные графики (по переменным)) (рисунок 1.13).

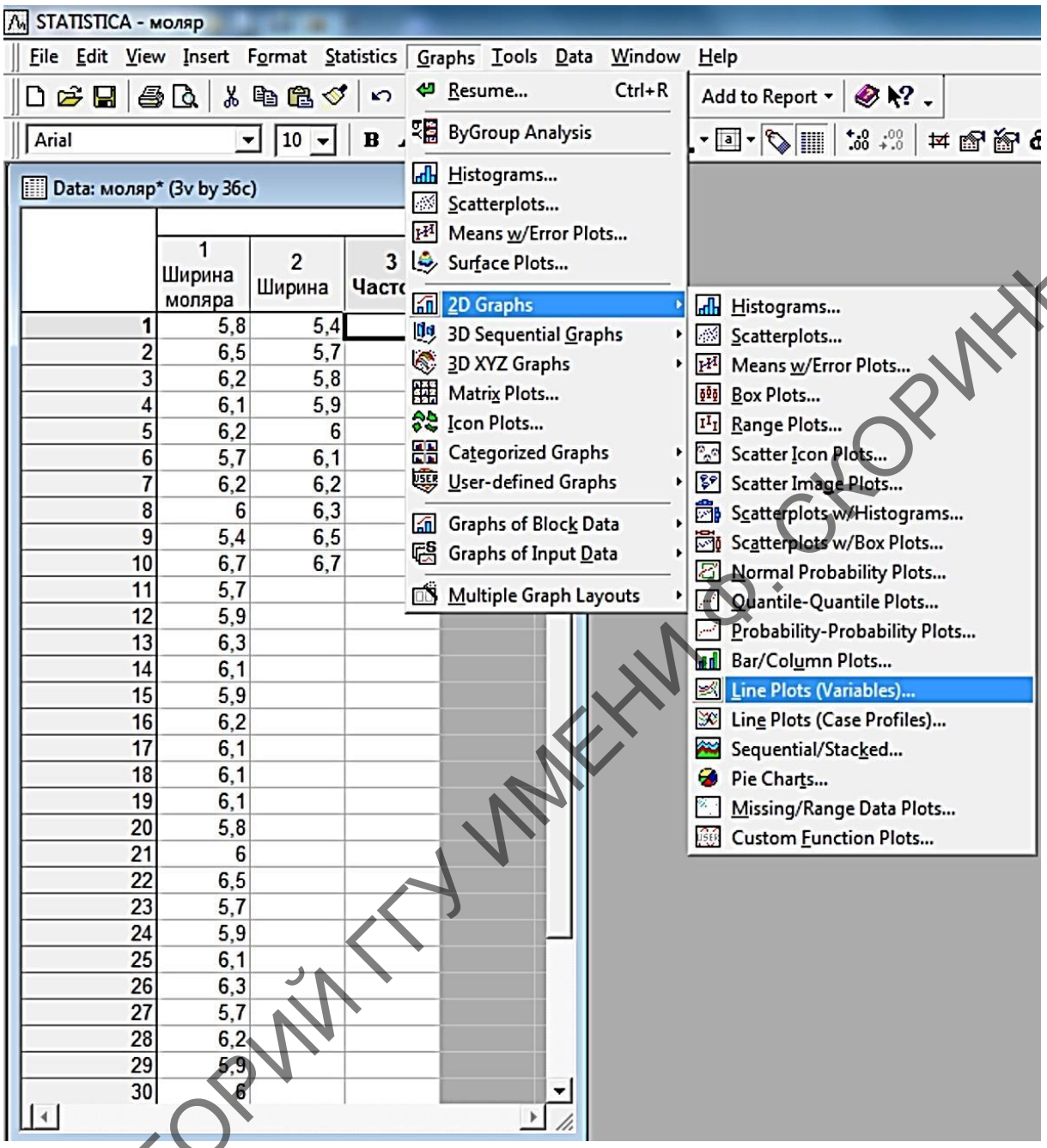


Рисунок 1.13 – Выбор типа графика

В новом диалоговом окне (рисунок 1.14) нужно выбрать закладку **Advanced** (Расширенные настройки). И далее на ней в поле **Graph type** (Тип графика) выбрать **XY Trace**, а в выпадающем меню **Display points** (Отображать точки) – выбрать опцию **On** (Включить). И, напоследок, открыть закладку **Options 1** и снять флажок **Display case labels** (Отобразить подписи значений). Впрочем, если последнюю операцию не выполнять, то значение ширины моляров на графике будет отражено.

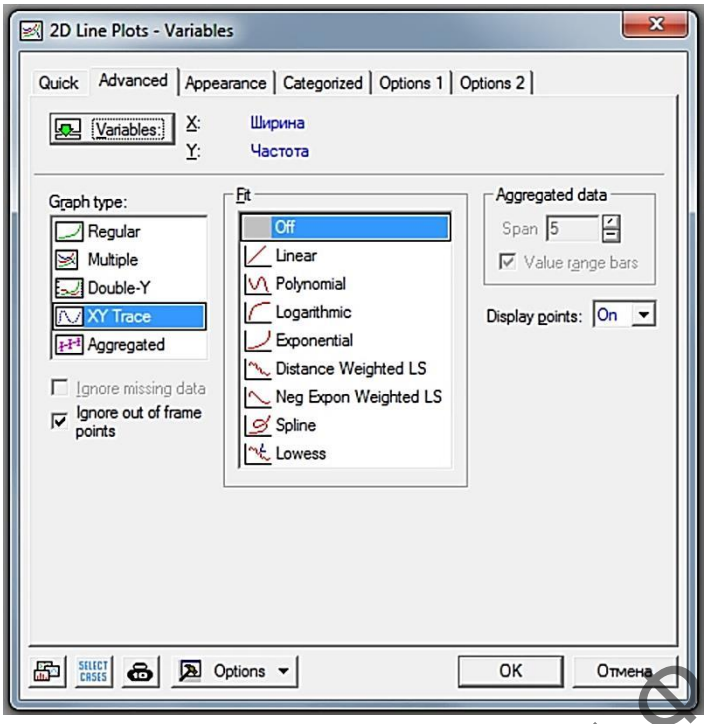


Рисунок 1.14 – Диалоговое окно опции **2D Line plots (Variables)**

Теперь необходимо указать программе, какой из столбцов нашей таблицы с данными соответствуют ширине моляров, т. е. представляет ось X, а какой – частоте встречаемости (ось Y). Для этого нужно снова перейти на закладку **Advanced** (Расширенные настройки) и нажать кнопку **Variables** (Переменные). Далее в появившемся окне в левом списке выделить пункт «Ширина», а в правом – «Частота» и нажать кнопку **ОК** (рисунок 1.14), а затем нажать кнопку **ОК** ещё раз, и в конечном итоге график будет отображён (рисунок 1.15).

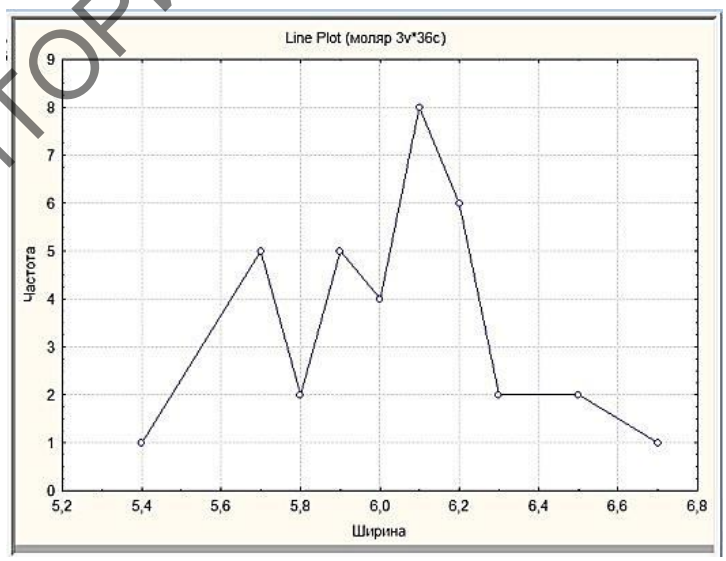


Рисунок 1.15 – Полигон распределения переменной «Ширина моляра»

Шаг 5. Построение гистограммы распределения.

Для построения гистограммы распределения переменной «Ширина моляра» необходимо в главном меню выбрать пункт **Graphs** (Графики), подпункт **2D Graphs** (Двухмерные графики) и затем – **Histograms** (Гистограммы) (рисунок 1.16).

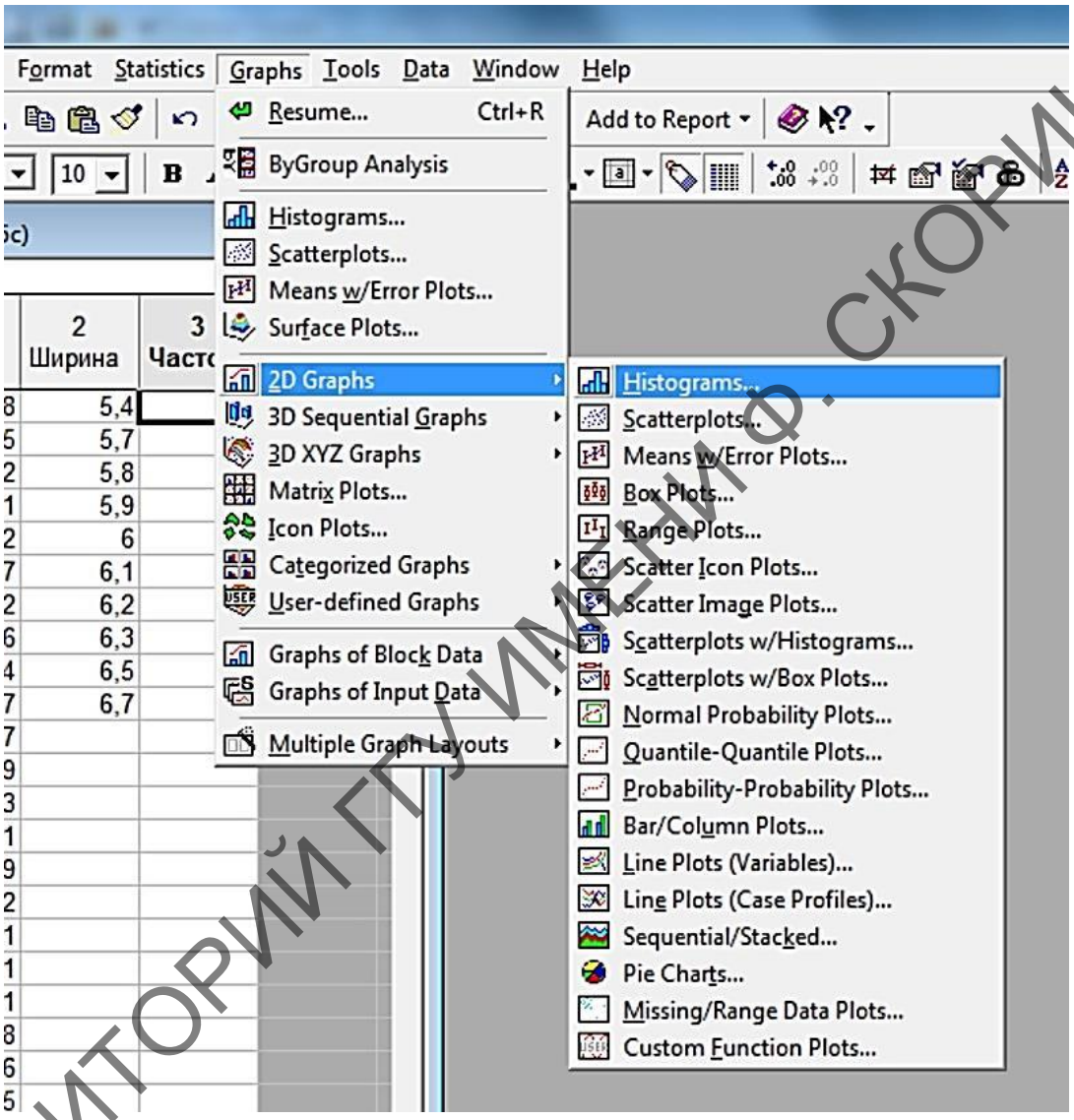


Рисунок 1.16 – Выбор пункта меню **Histograms**

В появившемся диалоговом окне (рисунок 1.17) нужно выбрать закладку **Advanced** (Расширенные настройки). Далее, нажав на кнопку **Variables** (Переменные), необходимо выбрать для анализа переменную «Ширина моляра». После этого в поле **Fit type** (Тип подгонки) выставить значение **Off**, а в выпадающем меню **Y axis** (Ось Y) – **N** (Количество). Остальные настройки можно оставить без изменений, затем нажать на кнопку **OK**. В результате будет отображена гистограмма, приведённая на рисунке 1.18.

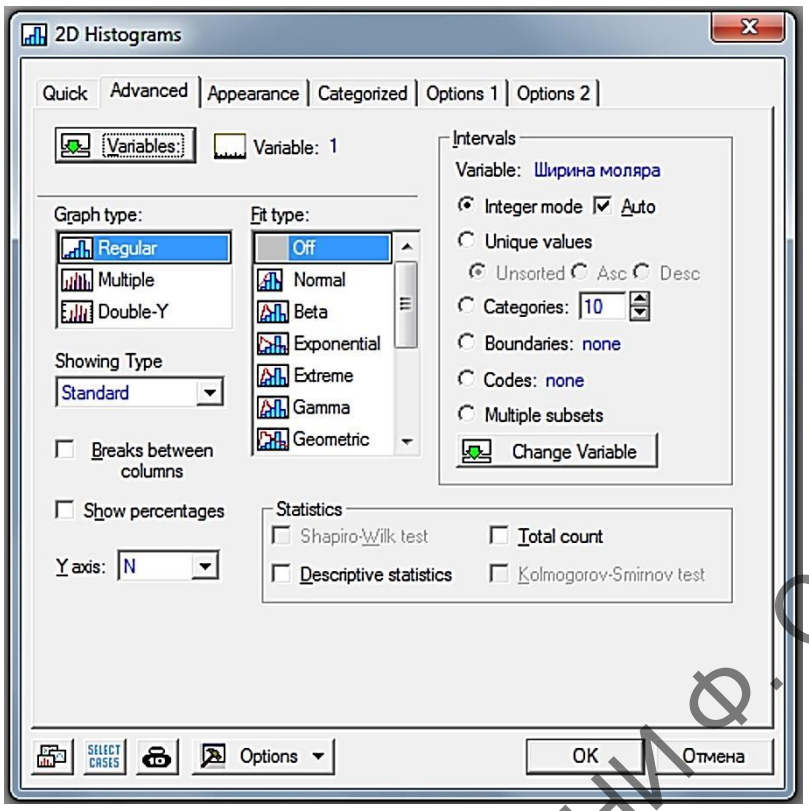


Рисунок 1.17 – Диалоговое окно опции меню **Histograms**

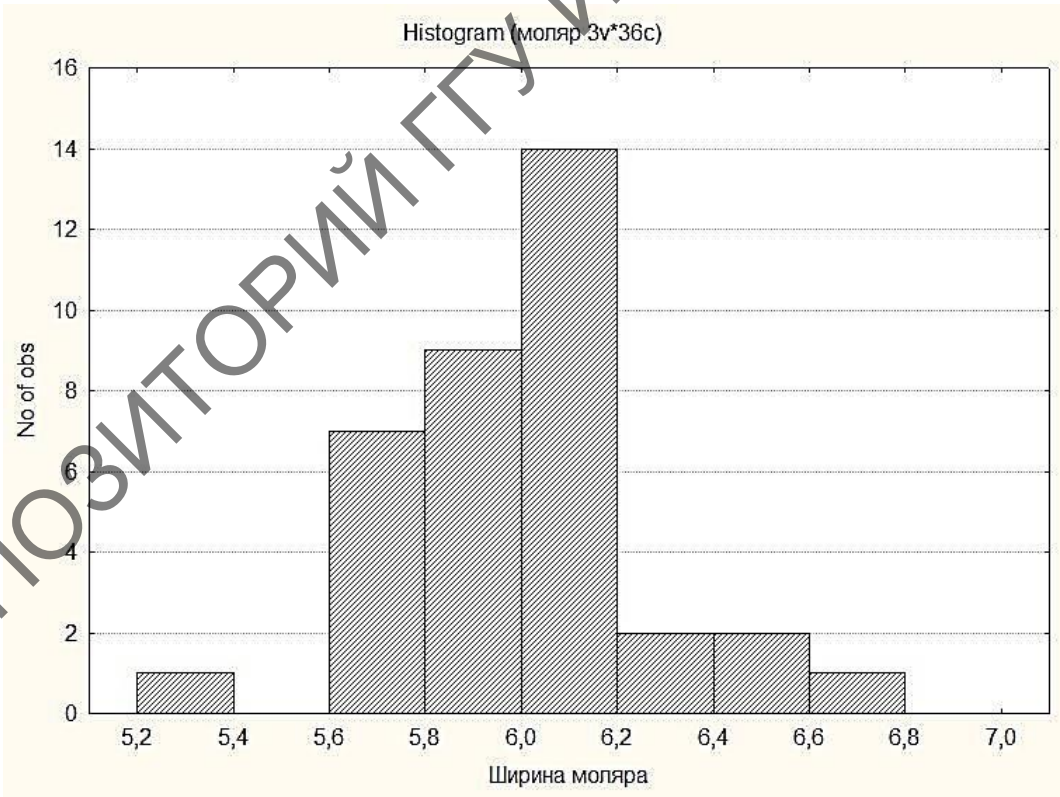


Рисунок 1.18 – Гистограмма распределения переменной «Ширина моляра»

Программный пакет STATISTICA 7.0 предоставляет достаточно широкие возможности для оформления графиков. Для этого достаточно кликнуть мышкой по интересующему вас элементу, и появится диалоговое окно со множеством опций по его настройке (заголовки, оси и их названия, маркеры и их форма, цвет и размер, и т. п.).

1.2.2 Графики диапазонов

Графики диапазонов (wisker plots) удобны для описания динамики изменений анализируемых признаков. Точки на таких графиках обычно соответствуют средней величине, чаще всего – средней арифметической, реже – медиане исследуемого признака. Отличительная особенность такого графика – это отходящие от точки вертикальные или горизонтальные линии, т. е. так называемые «усы» (whiskers). Длина этих линий соответствует либо величине выбранного показателя разброса данных (минимум и максимум, среднеквадратичное отклонение, дисперсия, квантили), либо точности оценки генеральных параметров (стандартная ошибка, доверительный интервал).

Рассмотрим построение такого графика на примере учёта числа видов жесткокрылых на одном из стационаров за сезон исследований. В течение 5 месяцев (с мая по сентябрь) на стационаре ежедекадно каждый месяц проводили сбор жесткокрылых и фиксировали количество видов.

Data: виды_жуков (2v by...

	1 Месяц	2 Число видов
1	май	18
2	май	18
3	май	16
4	июнь	19
5	июнь	22
6	июнь	25
7	июль	15
8	июль	21
9	июль	17
10	август	15
11	август	14
12	август	13
13	сентябрь	9
14	сентябрь	11
15	сентябрь	8

Рисунок 1.19 – Оформление данных для построения диаграммы диапазонов

Шаг 1. Создание файла данных.

Чтобы «объяснить» программе, какие из учётов относятся к конкретному месяцу, необходимо добавить дополнительный столбец «Месяц», в котором перечислены названия месяцев, а в следующем за ним столбце «Число видов» указаны значения исследуемой нами переменной (рисунок 1.19).

Столбец «Месяц» в программе STATISTICA является так называемой «группирующей переменной» (grouping variable), а столбец, в котором непосредственно находятся значения исследуемого признака (в нашем случае – число видов) – зависимой переменной (dependent variable). То есть, другими словами, мы определяем, что число видов жуков зависит от месяца учёта.

Следует заметить, что подобный способ оформления таблицы с данными характерен для многих видов анализа, применяемых в STATISTICA, и будет часто встречаться нам в дальнейшем.

Шаг 2. Выбор графика диапазонов.

В разделе **Graphs** (Графики) главного меню выберите опцию **2D Graphs** (Двухмерные графики), а затем – **Means w/Error plots** (Графики средних с ошибками) (рисунок 1.20).

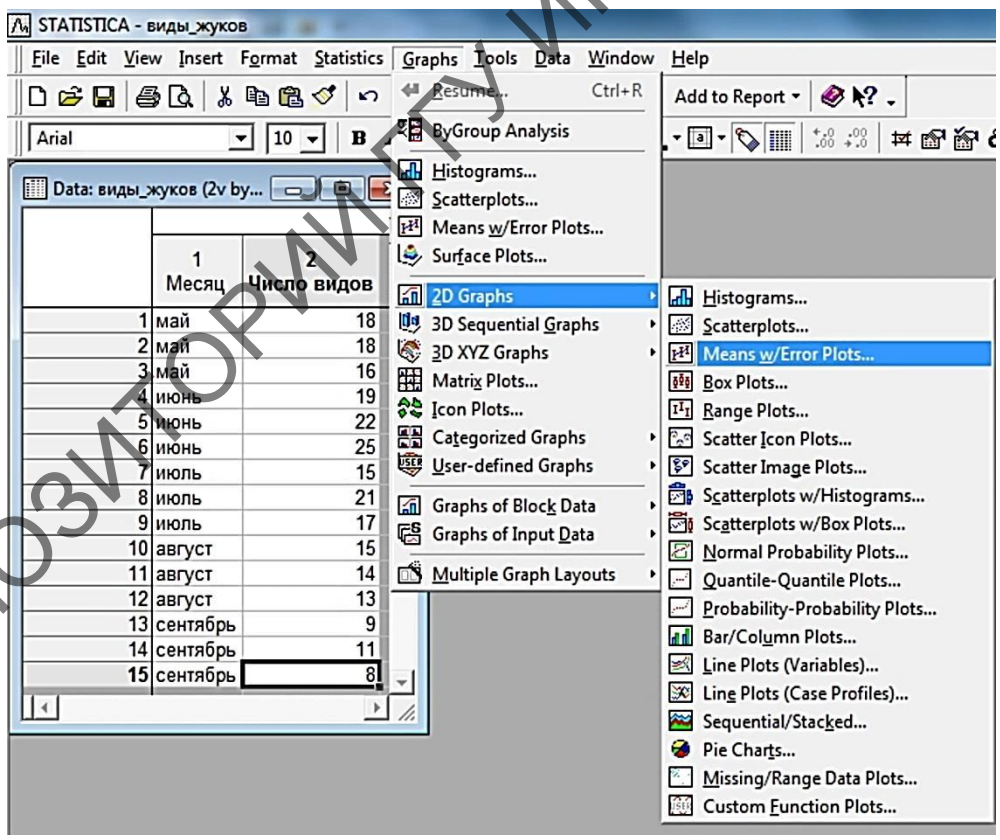


Рисунок 1.20 – Выбор опции **Means w/Error plots** (Графики средних с ошибками) в STATISTICA 7.0

Шаг 3. Указание переменных.

Сначала нужно указать программе, с какими переменными нам надо работать. Для этого кликаем мышкой по кнопке **Variables** (Переменные) и указываем группирующую и зависимую переменные в диалоговом окне **Select variables for means with error plots** (Выбор переменных для графика средней с ошибкой), как показано на рисунке 1.21. После этого нажимаем кнопку **ОК**.

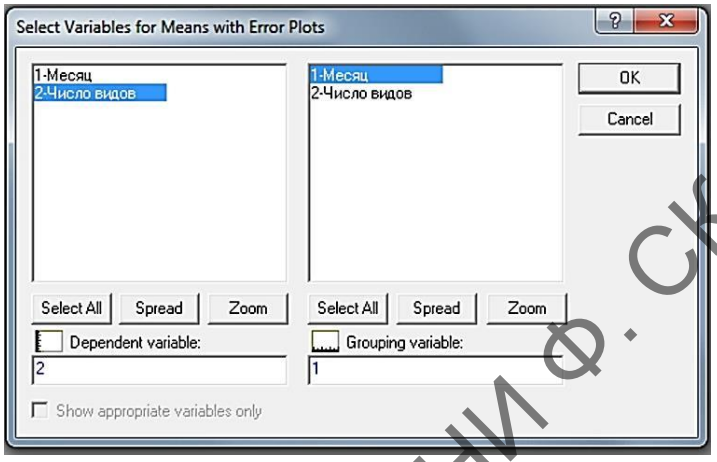


Рисунок 1.21 – Виды переменных диапазона «Число видов»

Шаг 4. Разметка оси X.

После нажатия кнопки **ОК** на предыдущем шаге мы возвращаемся в ранее открытое диалоговое окно (рисунок 1.22), где необходимо перейти в закладку **Advanced** (Расширенные настройки).

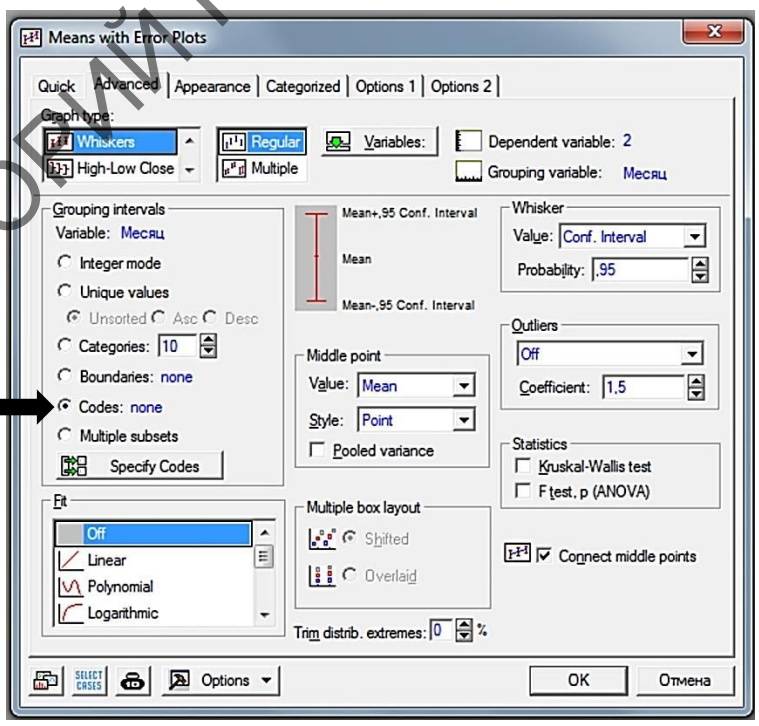


Рисунок 1.22 – Окно настройки параметров графика диапазонов

Далее в поле **Grouping intervals** (Группирующие интервалы) необходимо указать, на какие интервалы следует разбить программе ось X (в нашем случае ось X – это названия месяцев). Для этого нужно выбрать пункт **Codes** (Коды), указанный на рисунке 1.22 стрелкой, и нажать кнопку **Specify Codes** (Специфические коды), расположенную чуть ниже. В появляющемся окне (рисунок 1.23) следует нажать кнопки **All** (Все) и **ОК**.

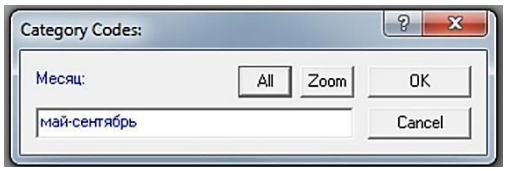


Рисунок 1.23 – Диалоговое окно **Category Codes**

В данном случае в качестве кодов используется название месяцев, и, нажимая кнопку **All** (Все), мы указываем программе, что хотим обсчитать данные за все месяцы. В противном случае необходимо уточнить, какие конкретные месяцы программе нужно учесть в расчётах.

Шаг 5. Указание диапазона.

В поле **Whisker** (Усы) в выпадающем меню **Value** (Значение) необходимо указать, что мы хотим видеть в качестве диапазона. В нашем случае – стандартную ошибку средней. В связи с этим в выпадающем списке выбираем опцию **Std error** (Стандартная ошибка), а в поле **Coefficient** (Коэффициент) необходимо выставить 1. Таким образом, основные настройки завершены, что отображено на рисунке 1.24.

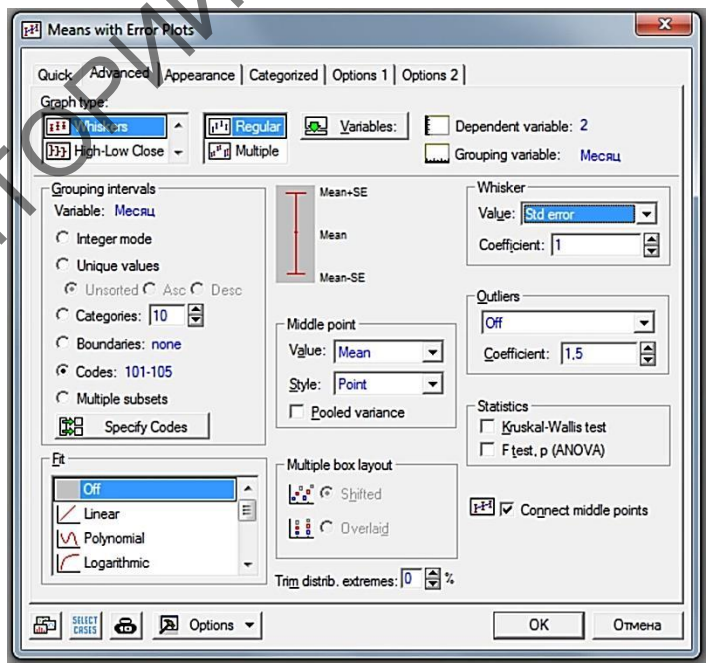


Рисунок 1.24 – Основные настройки графика диапазонов

Шаг 6. Построение графика.

Для построения графика необходимо в диалоговом окне нажать на кнопку **ОК**. Получившийся график отражён на рисунке 25.

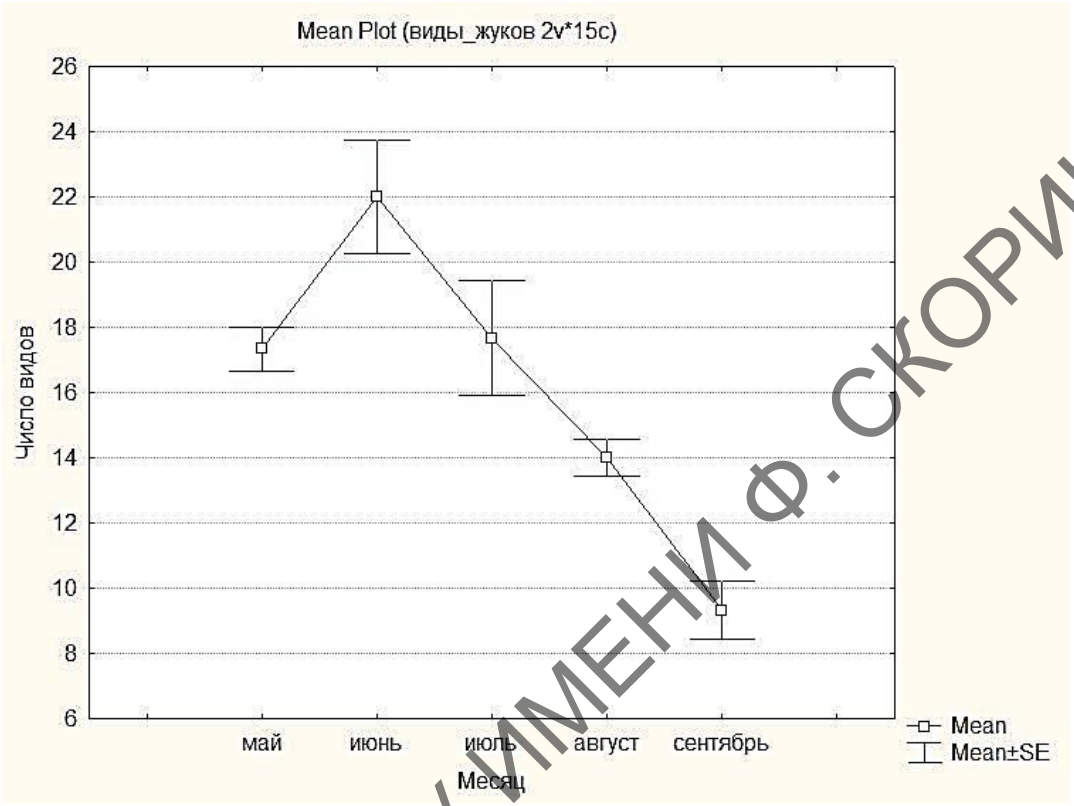


Рисунок 1.25 – График диапазонов средней со стандартной ошибкой числа видов жуков на протяжении сезона учётов

1.2.3 Диаграммы размахов

Подобные диаграммы получили свое негласное название «ящики с усами» (box-whisker plots) за своеобразный вид: точку, соответствующую или средней арифметической, или медиане, окружает прямоугольник – «ящик» (box), от которого отходят показатели разброса или точности в виде длинных «усов» (whiskers).

Графики этого типа позволяют дать очень полную статистическую характеристику для анализируемой исследователем выборки. Диаграммы размахов обычно используют для визуальной быстрой оценки разницы между двумя или более переменными.

Для примера построения подобного рода диаграмм используем рассмотренные выше данные о числе видов жесткокрылых.

Шаг 1. Создание таблицы с данными.

Откроем ранее сохранённую таблицу из предыдущего раздела по количеству видов жесткокрылых в различные месяцы (рисунок 1.19) или создадим ее заново.

Шаг 2. Загрузка модуля диаграммы.

Необходимо в главном меню выбрать раздел **Graphs** (Графики) и опцию **2D Graphs** (Двухмерные графики), затем – **Box plots** (Графики в виде ящиков) (рисунок 1.26).

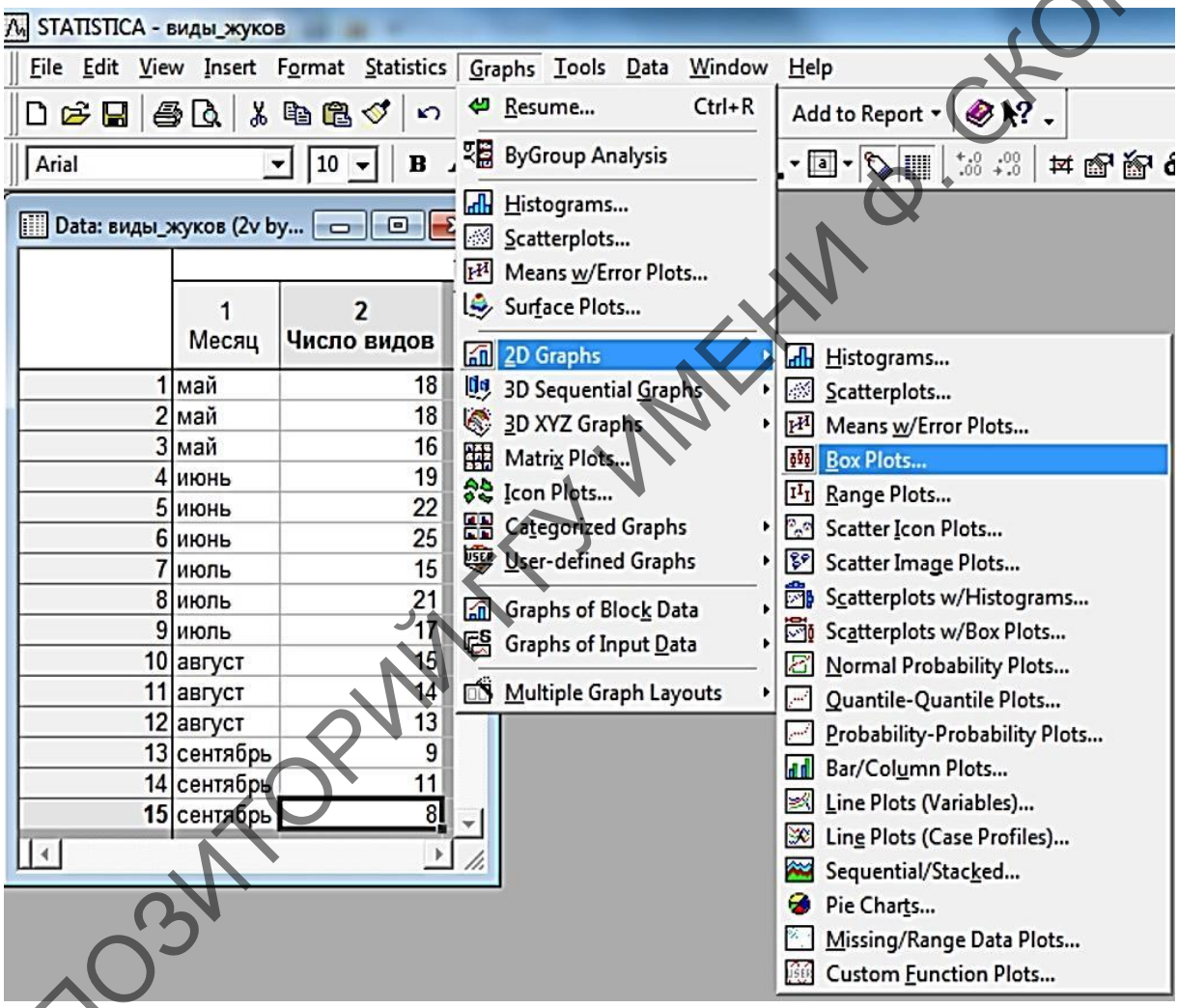


Рисунок 1.26 – Выбор опции **Box plots** (Графики в виде ящиков) в STATISTICA 7.0

Шаг 3. Выбор опций диаграммы

Прежде чем выбрать соответствующие опции, необходимо зайти в диалоговом окне на закладку **Advanced** (Расширенные настройки) и указать программе необходимые переменные, нажав кнопку **Variables**

(Переменные). Переменные и их коды указываются так же, как описано в пункте 1.2.2. Диалоговое окно отображено на рисунке 1.27.

Далее в поле **Middle point** (Срединная точка) нужно выбрать тот показатель (среднюю арифметическую или медиану), который мы хотим отобразить. В нашем случае это средняя арифметическая (Mean).

Затем в поле **Box** (Ящик) выбрать показатель, который будет отображаться прямоугольником на диаграмме. В данном случае это стандартная ошибка средней арифметической (Std error) с коэффициентом, равным единице.

После этого в поле **Whisker** (Усы) необходимо выбрать тот показатель, который будет изображён на диаграмме как «усы размаха». Это будет стандартное (среднеквадратичное) отклонение (Std dev), также с коэффициентом, равным единице.

В заключение настроек в поле **Outliers** (Выбросы) установить значение **Off** (Отключить). Остальные параметры оставить по умолчанию и нажать **ОК**.

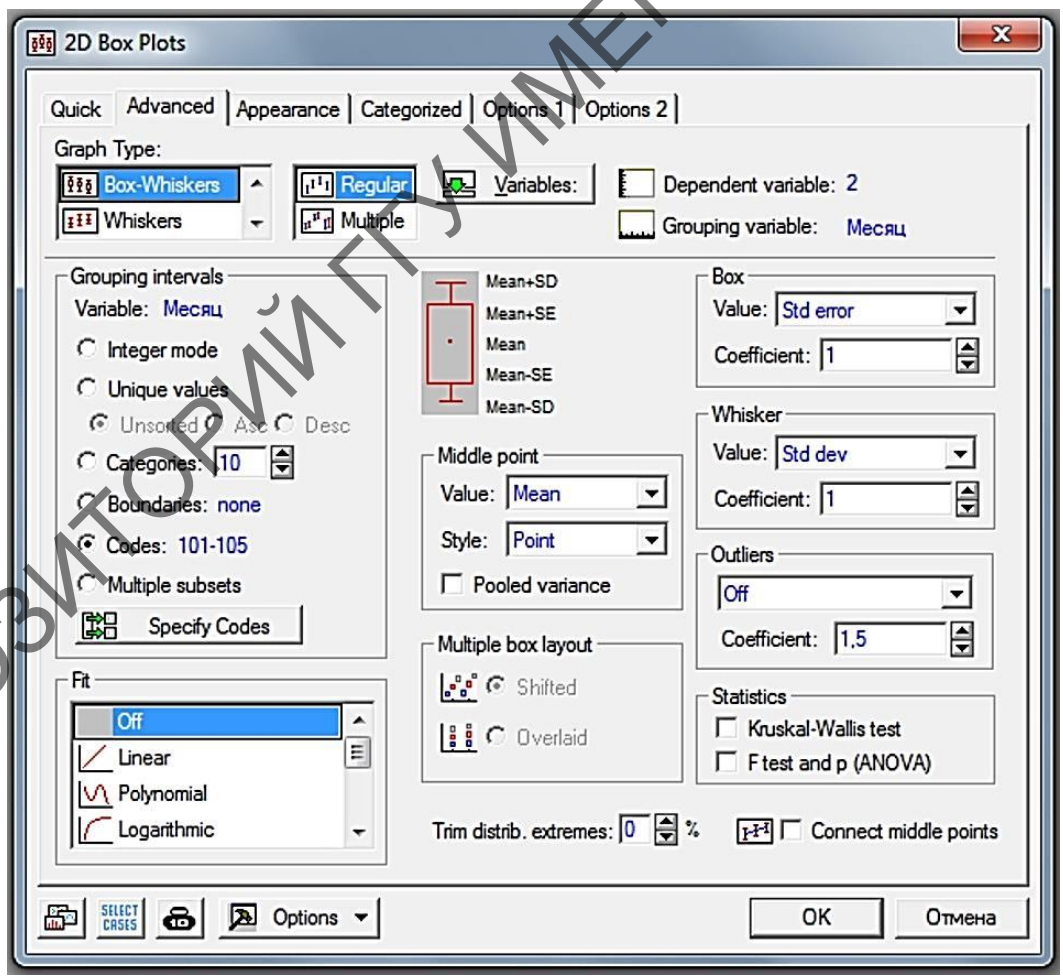


Рисунок 1.27 – Диалоговое окно опции **2D Box plots**

Искомая диаграмма будет иметь вид, представленный на рисунке 1.28.

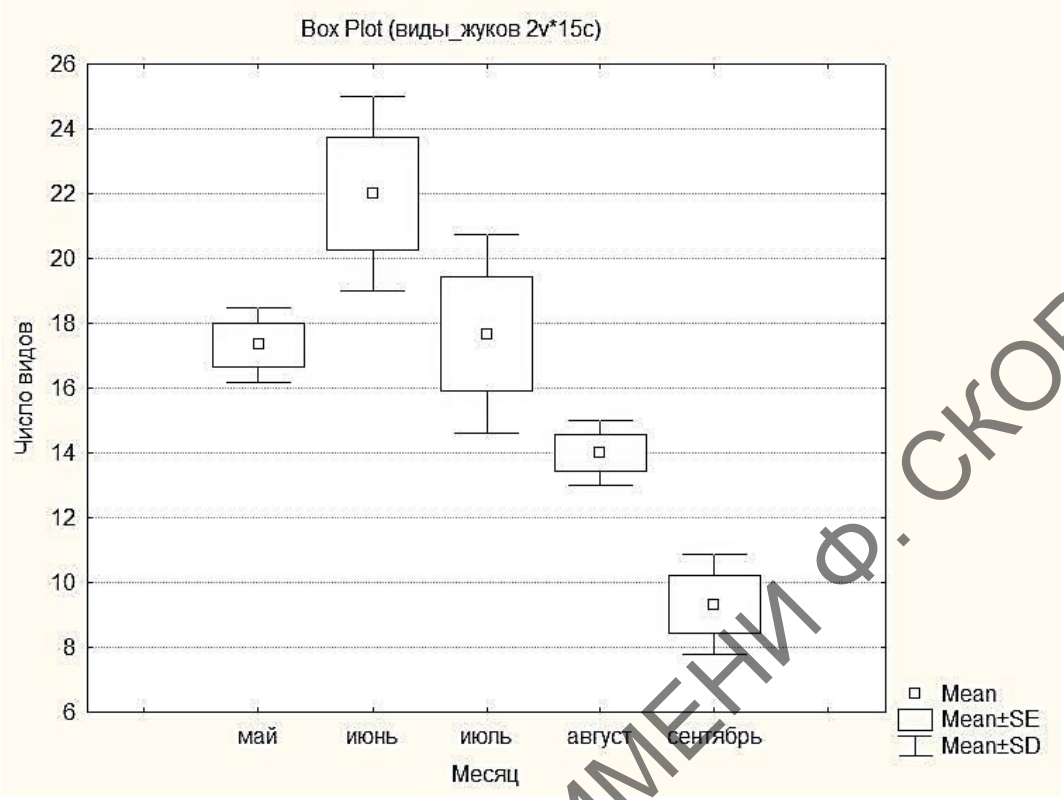


Рисунок 1.28 – Диаграмма размахов числа видов жуков на протяжении сезона учётов

Для экспресс-анализа данной диаграммы необходимо учитывать, что если прямоугольники (т.е. стандартная ошибка средней арифметической) сравниваемых переменных не пересекаются, то с высокой долей вероятности можно утверждать о достоверном отличии одной переменной от другой.

1.2.4 Круговые диаграммы

Подобного рода круговые и эллиптические диаграммы (pie charts) очень удобны при анализе каких-либо качественных признаков. Для примера рассмотрим распределение в сообществе 15 видов животных с различной биотопической приуроченностью.

Шаг 1. Создание таблицы с данными.

Создадим новую таблицу с данными, отражающими биотопическую приуроченность (биопреферендум) каждого из 15 обнаруженных видов животных (рисунок 1.29).

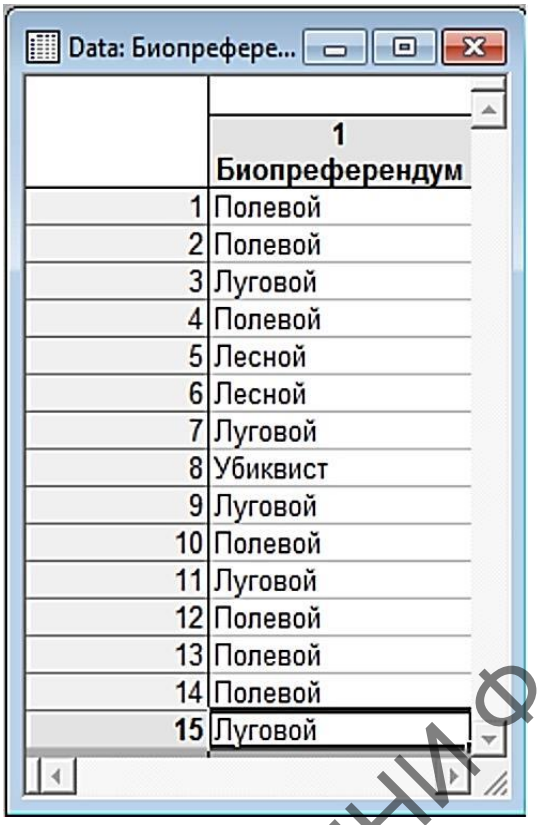


Рисунок 1.29 – Оформление данных для построения круговой диаграммы

Шаг 2. Выставление параметров.

Для построения круговой диаграммы долей необходимо:

– в главном меню выбрать раздел **Graphs** (Графики) и опцию **2D Graphs** (Двухмерные графики) и затем – **Pie charts** (Круговые диаграммы) (рисунок 1.30);

– в появившемся диалоговом окне перейти на закладку **Advanced** (Расширенные настройки);

– в поле **Frequency intervals** (Интервалы частот) выбрать опцию **Codes** (Коды) и нажать кнопку **Specify Codes** (Специфические коды). На появившейся панели нажать кнопку **All** (Все), а затем **ОК**;

– в поле **Pie legend** (Легенда диаграммы) выбрать вариант подписи сегментов круговой диаграммы. В нашем случае – **Text and Percent** (Текст и проценты). Таким образом, будут отображены названия биопреферендума и частота его встречаемости (%) в исследованном сообществе;

– остальные параметры выставить, как указано на рисунке 1.31. Данные параметры рассматриваются для круговой диаграммы и расчёта долей самой программой.

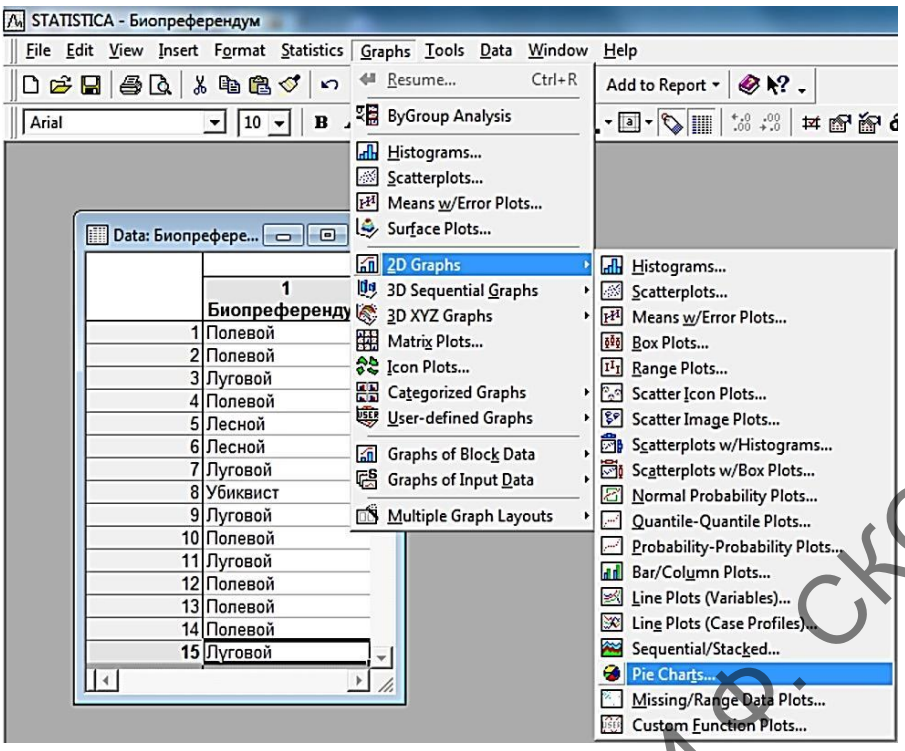


Рисунок 1.30 – Выбор раздела меню **Pie chart** (Круговая диаграмма) в STATISTICA 7.0

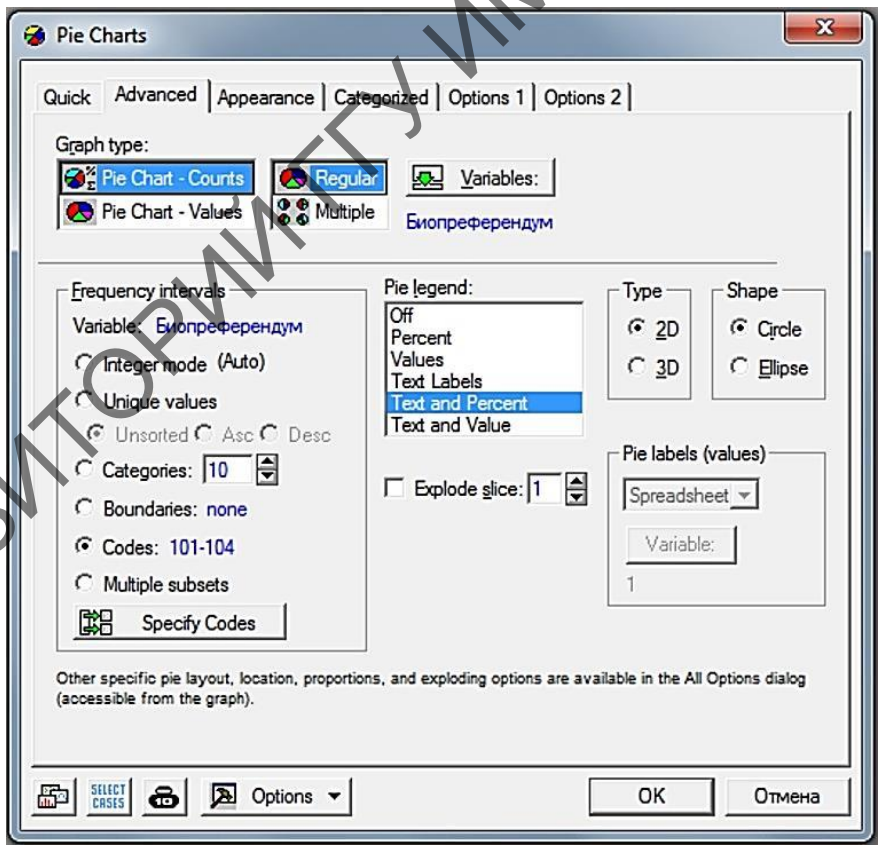


Рисунок 1.31 – Оформление диалогового окна **Pie chart** (Круговая диаграмма) в STATISTICA 7.0

Шаг 3. Построение диаграммы.

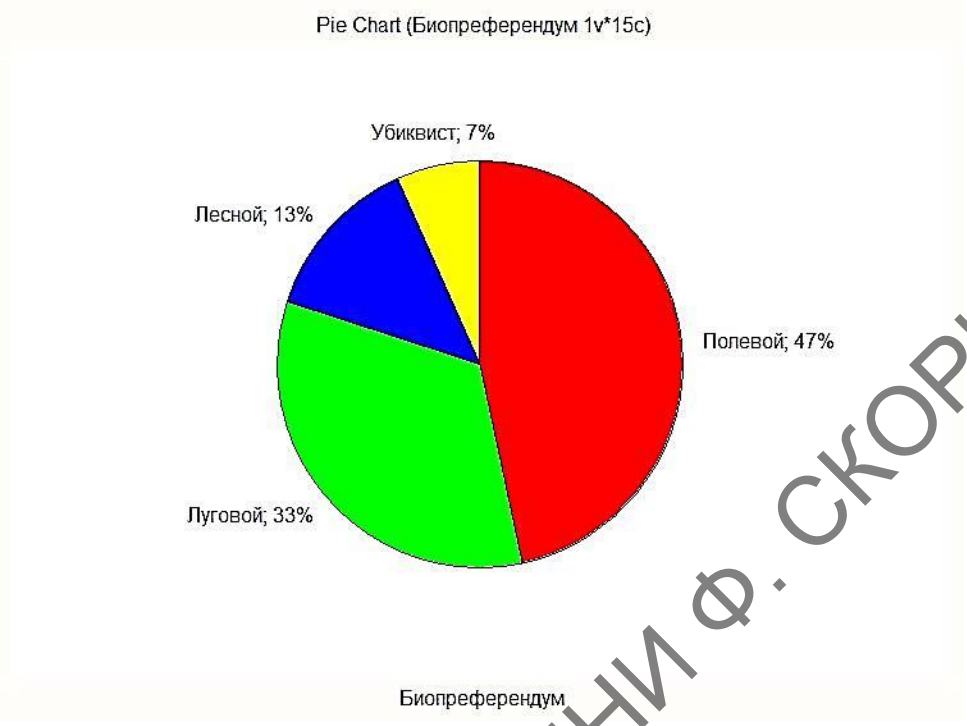


Рисунок 1.32 – Круговая диаграмма в STATISTICA 7.0

При нажатии кнопки ОК после выставления параметров произойдет построение и отображение искомой диаграммы (рисунок 32).

Задания для самоконтроля

1) Создайте таблицу с результатами олимпийских чемпионов в беге на 100 м.

Год	Чемпион	Страна	Время	Год	Чемпион	Страна	Время
1896	Вэрк	США	12,0	1936	Оуэнс	США	10,3
1900	Джервис	США	10,8	1948	Диллард	США	10,3
1904	Хан	США	11,0	1952	Реминджил	США	10,4
1906	Хан	США	11,2	1956	Морроу	США	10,5
1908	Уолкер	ЮАР	10,8	1960	Хари	ФРГ	10,2
1912	Крейг	США	10,8	1964	Хейес	США	10,0
1920	Пэддок	США	10,8	1968	Хайнс	США	09,9
1924	Абрахаме	Англия	10,6	1972	Борзов	СССР	10,1
1928	Уияльямс	Канада	10,8	1976	Кроуфорд	Тринидад	10,6
1932	Тоулэн	США	10,3	1936	Оуэнс	США	10,3

2) Количество щенков у 160 самок лисиц было следующим.

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

Составьте таблицу с данными, на основе которой постройте вариационный ряд, полигон распределения и гистограмму.

3) Были проведены промеры пескарей, обитающих в одном из водоёмов Гомельской области (в мм).

30	81	52	47	51	62	34	50	25	71	51	61	48	50	79
71	42	47	44	59	39	40	39	65	42	74	64	70	24	9

Составьте таблицу с данными, на основе которой постройте вариационный ряд, полигон распределения и гистограмму.

4) Был проведён учёт численности врановых птиц в центральном парке города в различные месяцы года.

Месяц	Количество	Месяц	Количество	Месяц	Количество
январь	15	апрель	21	июль	45
январь	12	апрель	17	июль	66
февраль	9	май	33	август	45
февраль	11	май	39	август	46
март	14	июнь	49	сентябрь	37
март	16	июнь	60	сентябрь	47

Постройте график диапазонов (wisker plot) и диаграмму размахов (box-whisker plot) колебаний численности врановых птиц в исследованный период.

5) Были проведены учёты численности *Daphnia pulex* на 3 створах р. Уза.

Створ	Численность	Створ	Численность	Створ	Численность
1	154	2	111	3	200
1	128	2	95	3	193
1	134	2	101	3	199

Постройте график диапазонов (wisker plot) и диаграмму размахов (box-whisker plot) колебаний численности рачка на исследованных участках реки.

6) Были проведены учёты встречаемости сороки в различных биотопах окрестностей села на протяжении месяца.

Число месяца	Биотоп	Число месяца	Биотоп	Число месяца	Биотоп	Число месяца	Биотоп
1	Берег	8	Ельник	15	Сосняк	22	Ольс
2	Берег	9	Ельник	16	Ольс	23	Ельник
3	Сад	10	Ельник	17	Берег	24	Берег
4	Сад	11	Березняк	18	Сад	25	Сад
5	Сад	12	Березняк	19	Сад	26	Березняк
6	Сосняк	13	Ольс	20	Сосняк	27	Парк
7	Сосняк	14	Сад	21	Сосняк	28	Парк

Постройте круговую диаграмму частоты встречаемости (в %) сороки в различных биотопах окрестностей села.

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

ТЕМА 2. ПЕРВИЧНЫЙ АНАЛИЗ ДАННЫХ В EXCEL И STATISTICA 7.0

2.1 Описательная статистика.
2.2 Проверка распределения на нормальность.
2.3 Сравнение выборочных средних.

2.1 Описательная статистика

2.1.1 Первичный анализ статистических данных в Excel

Под описательной статистикой обычно понимают расчёт средних, стандартной ошибки, стандартного отклонения, асимметрии, эксцесса, медианы, моды и ряда других основных показателей выборки.

В качестве примера расчёта описательной статистики в электронных таблицах Excel мы возьмём вес бурозубок (в граммах) на участке смешанного леса:

7,1 7,7 3,6 8,3 8,8 10,4 8,9 9,0 8,9 14,0 9,7 9,4 8,5 15,9 12,6 7,1
9,1 6,2 10,7 13,8 13,6 15,2 3,4 9,3 13,3 6,7 7,9 4,9 4,5 8,0 17,1 9,1

Имеющиеся данные необходимо предварительно представить либо в виде одной строки, либо одного столбца (рисунок 2.1).

Рисунок 2.1 – Создание ряда данных в книге Excel

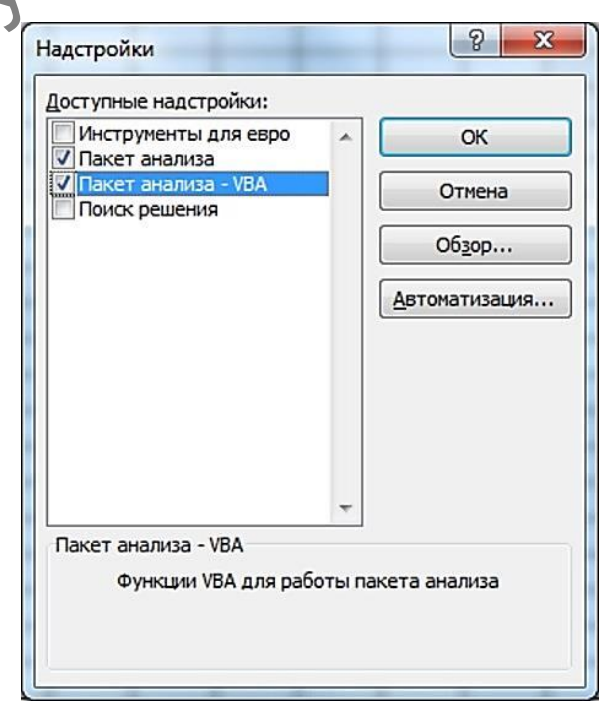


Рисунок 2.2 – Включение пакета анализа данных в надстройках Excel

Затем нужно включить пакет «Анализ данных» в надстройках Excel. Для этого необходимо перейти в раздел меню **Файл → Параметры → Надстройки** и в диалоговом окне внизу около окошка **Надстройки Excel** нажать кнопку **Перейти**, в результате чего появится диалоговое окно (рисунок 2.2), в котором необходимо включить опции **Пакет анализа** и **Пакет анализа – VBA**. Всё. Теперь ваш Excel готов к статистической обработке данных.

Прежде чем провести первичную статистическую обработку наших данных по бурозубкам, необходимо определить величину класса (межклассовый интервал) по формуле:

$$i = (X_{max} - X_{min})/(1 + 3{,}322 \times lgN), \quad (1)$$

где N – число наблюдений.

В нашем случае размер межклассового интервала будет равен 2,32, т. е. ≈ 2,3. Представим размеры интервалов в виде таблицы (таблица 2.1).

Таблица 2.1 – Размер интервалов

классы	3,4	5,7	8,0	10,4	12,7	15,0	17,3

Полученная таблица понадобится нам позже при построении гистограммы. Теперь мы можем перейти уже непосредственно к расчёту описательной статистики наших данных.

Шаг 1. Указание вида анализа.

Для расчёта описательной статистики необходимо перейти в раздел меню **Данные** и нажать справа на панели кнопку **Анализ данных**, а затем в выпавшем окне в списке выбрать опцию **Описательная статистка** (рисунок 2.3) и нажать кнопку **ОК**.

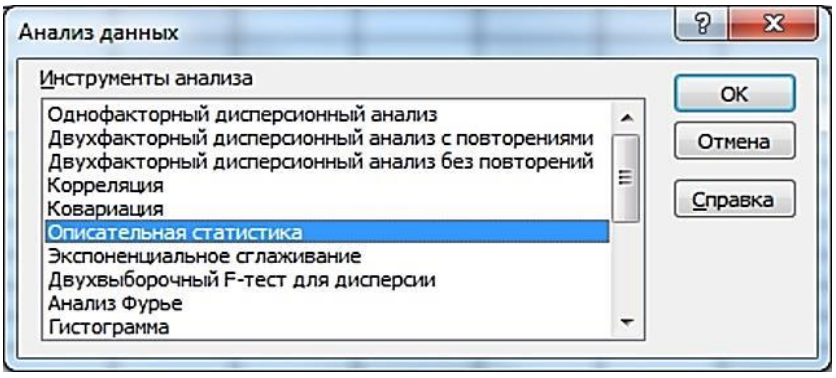


Рисунок 2.3 – Диалоговое окно раздела меню **Анализ данных**

Шаг 2. *Выбор входного диапазона данных.*

В появившемся диалоговом окне (рисунок 2.4) программы необходимо указать диапазон ваших данных в виде интервала абсолютных адресов ячеек (в нашем случае $A$1:$A$31), либо набрав его вручную в окошке **Входной интервал**, либо установив курсор внутрь окошка, указателем мыши выделить этот диапазон непосредственно в книге.

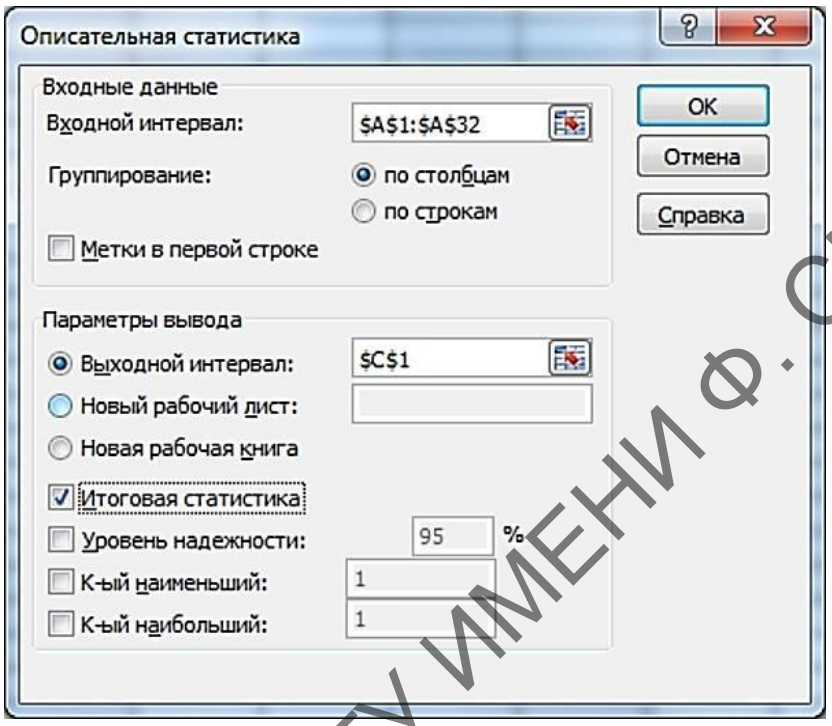


Рисунок 2.4 – Диалоговое окно **Описательная статистика**

Группирование данных следует проводить так, как они представлены. В нашем случае – по столбцам. Метку в первой строке необходимо обозначать только в том случае, если столбец с данными озаглавлен и необходимо указать программе, что ячейку с заголовком не нужно учитывать в расчётах.

Шаг 3. Выбор параметров вывода результатов анализа.

Для того чтобы указать программе, где необходимо вывести результаты анализа, в диалоговом окне **Описательная статистика** нужно выбрать один из трёх вариантов:

– **Выходной интервал** (необходимо указать адрес конкретной ячейки текущего рабочего листа книги Excel, где будет выведен левый верхний угол таблицы с результатами анализа);

– **Новый рабочий лист** (результаты будут выведены на новом рабочем листе текущей книги Excel);

– **Новая рабочая книга** (результаты будут выведены в новой книге Excel, т. е. в новом файле).

В нашем примере выбран первый вариант с ячейкой $C$1 (рисунок 2.4).

Затем необходимо обозначить пункт **Итоговая статистика**, а остальные поля не выделять и нажать кнопку **OK**. Полученный результат отображён на рисунке 2.5.

	A	B	C	D	E
1	7,1		*Столбец1*		
2	7,7				
3	3,6		Среднее	9,459375	
4	8,3		Стандартная ошибка	0,613849885	
5	8,8		Медиана	8,95	
6	10,4		Мода	7,1	
7	8,9		Стандартное отклонени	3,472459329	
8	9		Дисперсия выборки	12,05797379	
9	8,9		Эксцесс	-0,236600693	
10	14		Асимметричность	0,411918565	
11	9,7		Интервал	13,7	
12	9,4		Минимум	3,4	
13	8,5		Максимум	17,1	
14	15,9		Сумма	302,7	
15	12,6		Счет	32	
16	7,1				
17	9,1				
18	6,2				
19	10,7				
20	13,8				

Рисунок 2.5 – Выведенные результаты опции **Описательная статистика**

Шаг 4. Формирование списка классов.

Для построения гистограммы распределения в Excel необходимо перенести данные ранее полученных межклассовых интервалов (рисунок 2.6).

Классы
3,4
5,7
8
10,4
12,7
15
17,3

Рисунок 2.6 – Классы бурозубок по массе (межклассовый интервал – 2,3)

Шаг 5. Построение гистограммы.

Для построения гистограммы необходимо в пункте меню **Данные** выбрать опцию **Анализ данных**. Затем в появившемся окне выбрать пункт **Гистограмма** (рисунок 2.7) и нажать **ОК**.

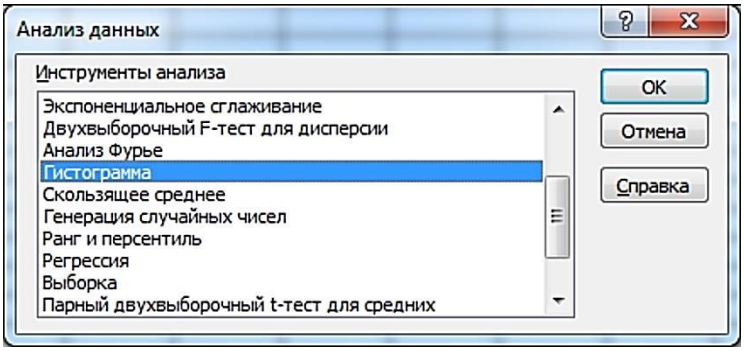


Рисунок 2.7 – Диалоговое окно раздела меню **Анализ данных**

В появившемся диалоговом окне **Гистограмма** необходимо указать поочерёдно:

а) входные данные, а именно:

– **Входной интервал** (диапазон значений признака, т. е. выборку; в нашем случае – $A$1:$A$31);

– **Интервал карманов** (диапазон значений классов, в нашем случае – $C$19:$C$25);

– **Метки** (учитывать или нет метку первой строки таблицы с данными);

б) параметры вывода (значения аналогичны шагу 3, только измените адрес или место вывода гистограммы).

Затем необходимо обозначить пункт **Вывод графика**, а остальные пункты оставить без изменений (рисунок 2.8) и нажать **ОК**.

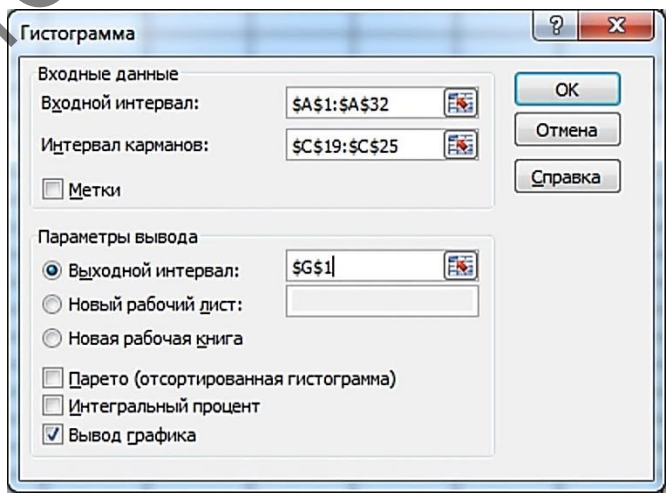


Рисунок 2.8 – Диалоговое окно **Гистограмма**

После нажатия кнопки **ОК** в указанном месте появляется гистограмма и сопровождающая её таблица с частотами каждого класса выборки (рисунок 2.9).

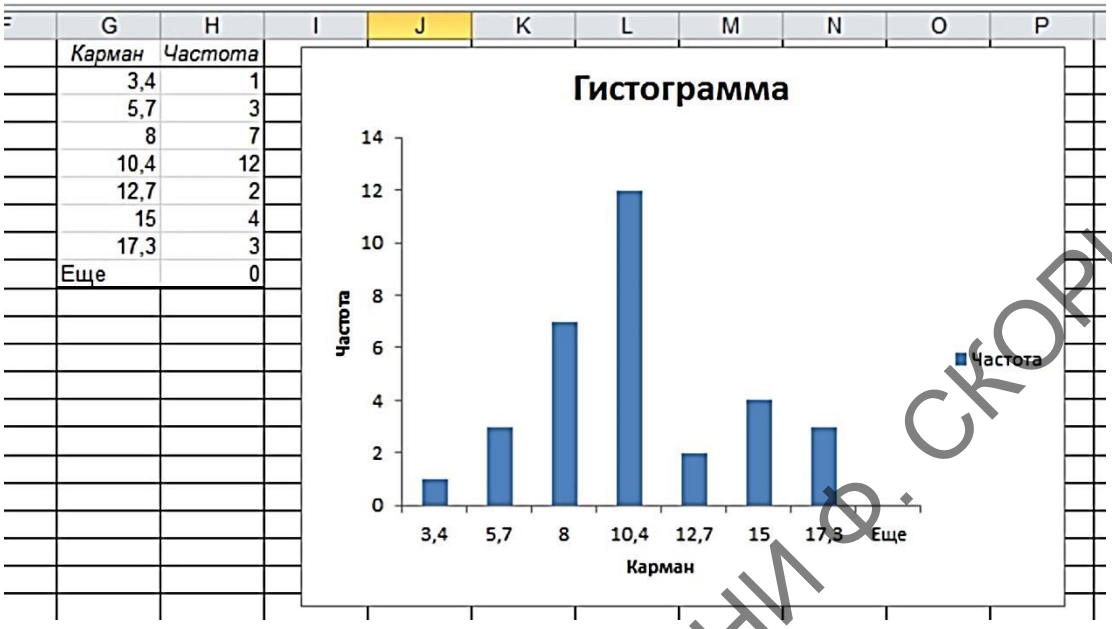


Рисунок 2.9 – Результаты построения гистограммы в Excel

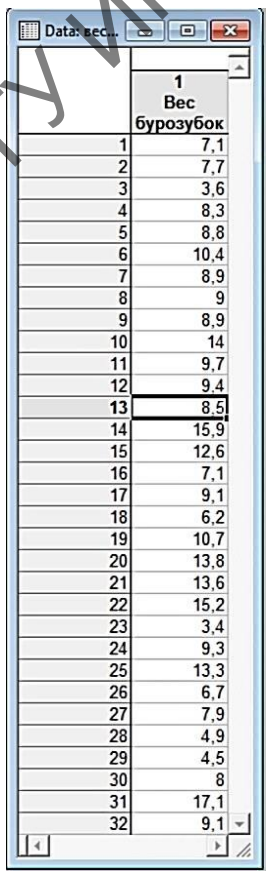


Рисунок 2.10 – Электронная таблица данных переменной «Вес бурозубок»

2.1.2 Первичный анализ статистических данных в Statistica

Шаг 1. Создание файла с данными.

Необходимо создать новую таблицу с данными в программе Statistica (см. Тема 1), либо скопировав данные по весу бурозубок из предыдущего файла Excel, либо набрав вручную. Единственную переменную можно назвать, к примеру, «Вес бурозубок» (рисунок 2.10).

Шаг 2. Выставление параметров анализа.

Для дальнейшего выяснения результатов описательной статистики изучаемой выборки необходимо зайти в пункт главного меню **Statistics** (Статистические процедуры) и выбрать в нём модуль **Basic Statistics/Tables** (Основные статистические показатели / Таблицы); далее – опцию **Descriptive statistics** (Описательная статистика) и нажать **ОК**. Появится диалоговое окно **Descriptive statistics** (рисунок 2.11), в котором необходимо перейти на закладку **Advanced** (Расширенные настройки) (рисунок 2.12), не забыв указать имя переменной.

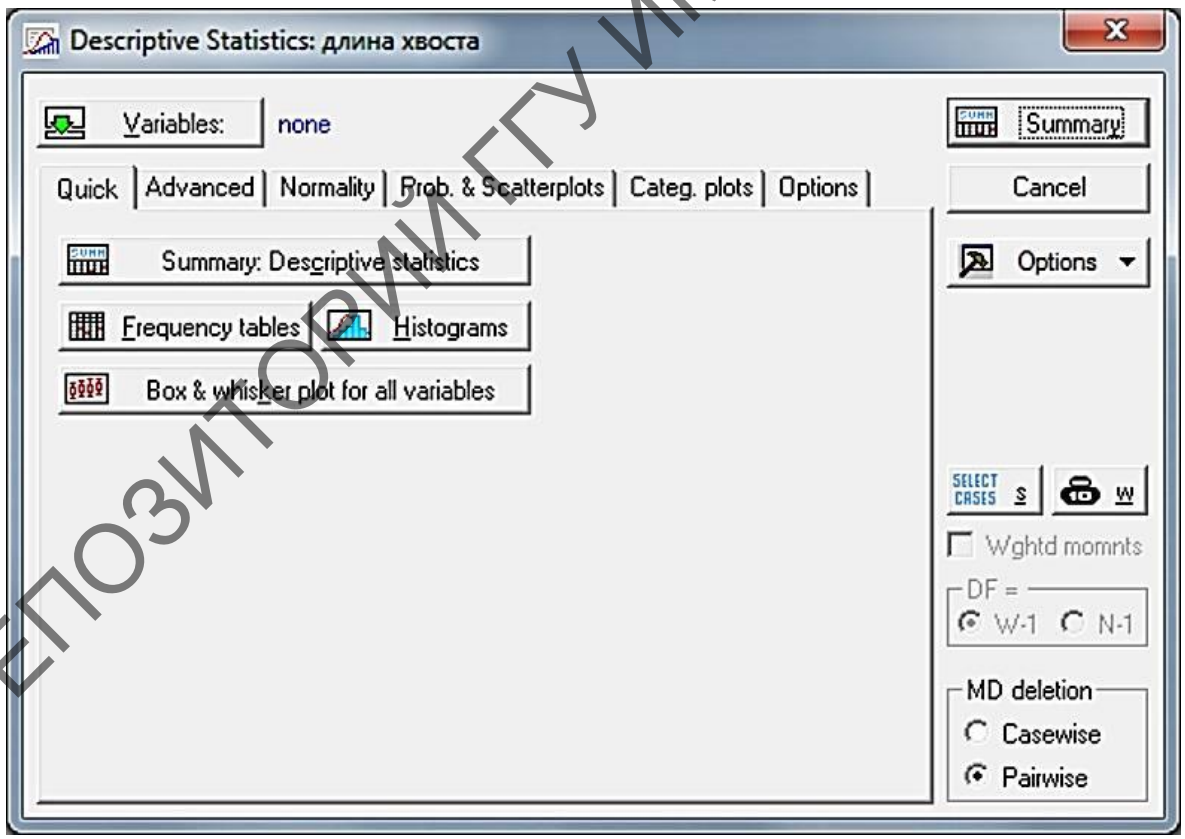


Рисунок 2.11 – Диалоговое окно **Descriptive statistics**

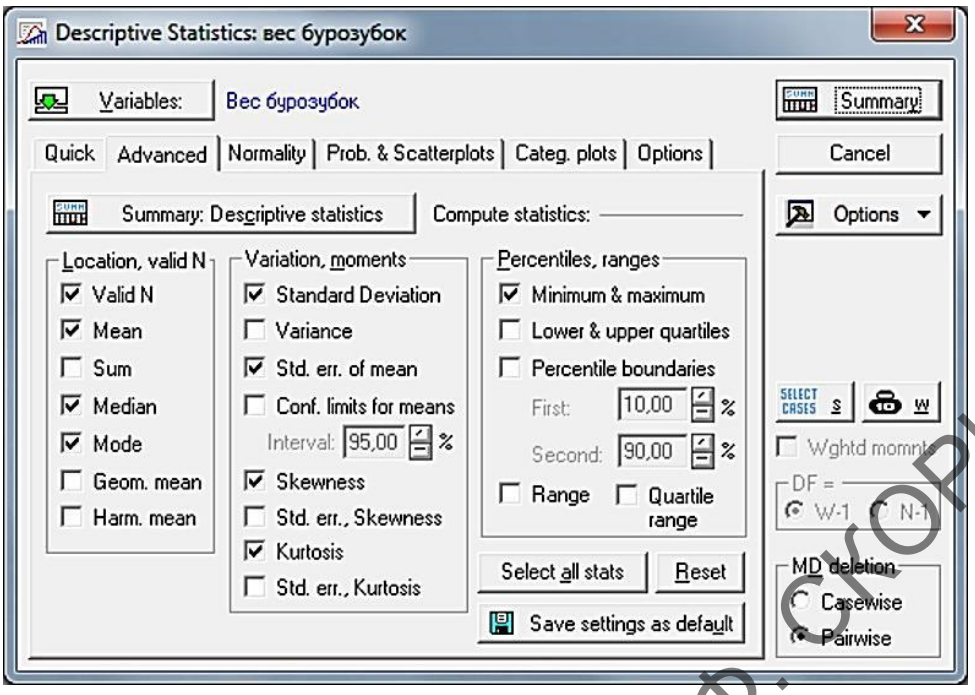


Рисунок 2.12 – Диалоговое окно **Descriptive statistics**, закладка **Advanced**

В закладке **Advanced** (Расширенные настройки) имеются следующие статистические показатели:

– **Valid N** – объём выборки (число значений признака);

– **Mean** – средняя арифметическая выборки;

– **Sum** – сумма всех значений выборки;

– **Median** – медиана выборки;

– **Mode** – мода выборки;

– **Geom. mean** – геометрическая средняя выборки;

– **Harm. mean** – гармоническая средняя выборки;

– **Standard Deviation** – стандартное отклонение выборки;

– **Variance** – дисперсия выборки;

– **Std. err. of mean** – стандартная ошибка средней выборки;

– **Conf. limits for means: Interval %** – доверительные пределы для средних: ширина доверительного интервала;

– **Skewness** – значение асимметрии распределения;

– **Std. err., Skewness** – стандартная ошибка асимметрии;

– **Kurtosis** – значение эксцесса распределения;

– **Std. err., Kurtosis** – стандартная ошибка эксцесса;

– **Minimum & maximum** – минимальное и максимальное значения в выборке;

– **Lower & upper quartiles** – нижний и верхний квартили;

– **Percentile boundaries: First & Second**: первый и второй перцентили;

– **Range** – размах выборки;

– **Quartile range** – межквартильный размах.

На этой закладке также имеется несколько кнопок:

– **Select all stats** – выбор сразу всех имеющихся статистических показателей;

– **Reset** – сброс всех включённых в расчёт показателей;

– **Save settings as default** – сохранить выставленные показатели как используемые по умолчанию, т. е. программа будет их предлагать для расчета по умолчанию при каждом запуске модуля **Descriptive Statistics**.

Нужно выставить все параметры, которые необходимо выяснить, используя в качестве образца рисунок 2.12.

После этого для получения долгожданных результатов анализа необходимо нажать кнопку **Summary** (Результат), которые отразятся в итоговой таблице (рисунок 2.13).

	Descriptive Statistics (вес бурозубок)										
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Std.Dev.	Standard Error	Skewness	Kurtosis
Вес бурозубок	32	9,459375	8,950000	Multiple	2	3,400000	17,10000	3,472459	0,613850	0,411919	-0,236601

Рисунок 2.13 – Итоги анализа описательной статистики

2.2 Проверка распределения на нормальность

Все существующие на сегодня методы статистической обработки данных подразделяют на 2 группы:

– параметрические;

– непараметрические.

Необходимым условием, определяющим возможность применения параметрических методов, является подчинение исследуемых признаков закону нормального распределения, имеющему характерный колоколообразный вид. Непараметрические же методы выполнения такого условия не требуют. Выявлено, что в 2/3 всех случаев распределения биологических признаков существенно отличаются от нормального.

Для того, чтобы избежать в будущем ошибок и неверного истолкования полученных результатов из-за применения параметрических методов анализа для ненормально распределенных данных, любой анализ биологических признаков должен начинаться с проверки нормальности их распределения. Для этого рассмотрим 3 способа, реализованные в программном пакете STATISTICA:

а) подгонка распределения (тест χ^2);

б) Тест Колмогорова – Смирнова и Лиллифорса; W-тест Шапиро – Уилка;

в) график нормальных вероятностей.

2.2.1. Подгонка распределения (критерий χ^2)

Шаг 1. Создание электронной таблицы с данными.

В качестве примера используем данные по длине хвоста (в мм) у оленьих мышей в возрасте одного года:

58	57	64	61	56	65	63	58	63	62	63
60	59	61	54	58	66	67	63	63	63	57
61	60	58	57	65	61	60	68	64	62	62
63	56	59	64	61	64	57	60	63	63	61
58	52	60	59	57	61	54	58	64	60	62
62	59	60	63	60	60	64	59	63	59	61
63	59	62	63	61	65	61	64	57	57	56
59	54	64	63	57	59	59	58	63	59	60

Создадим электронную таблицу с данными (рисунок 2.14).

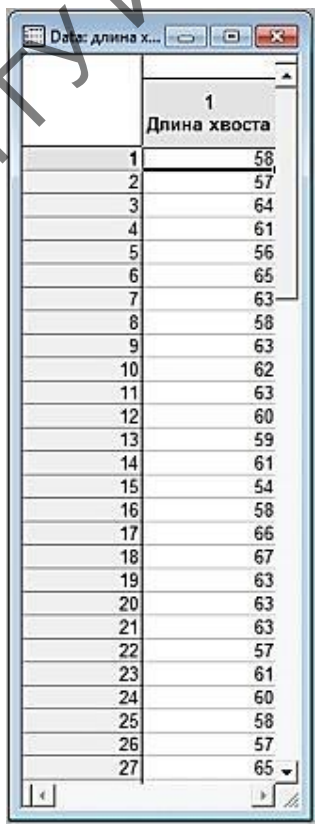


Рисунок 2.14 – Электронная таблица данных переменной «Длина хвоста»

Шаг 2. Выбор анализа.

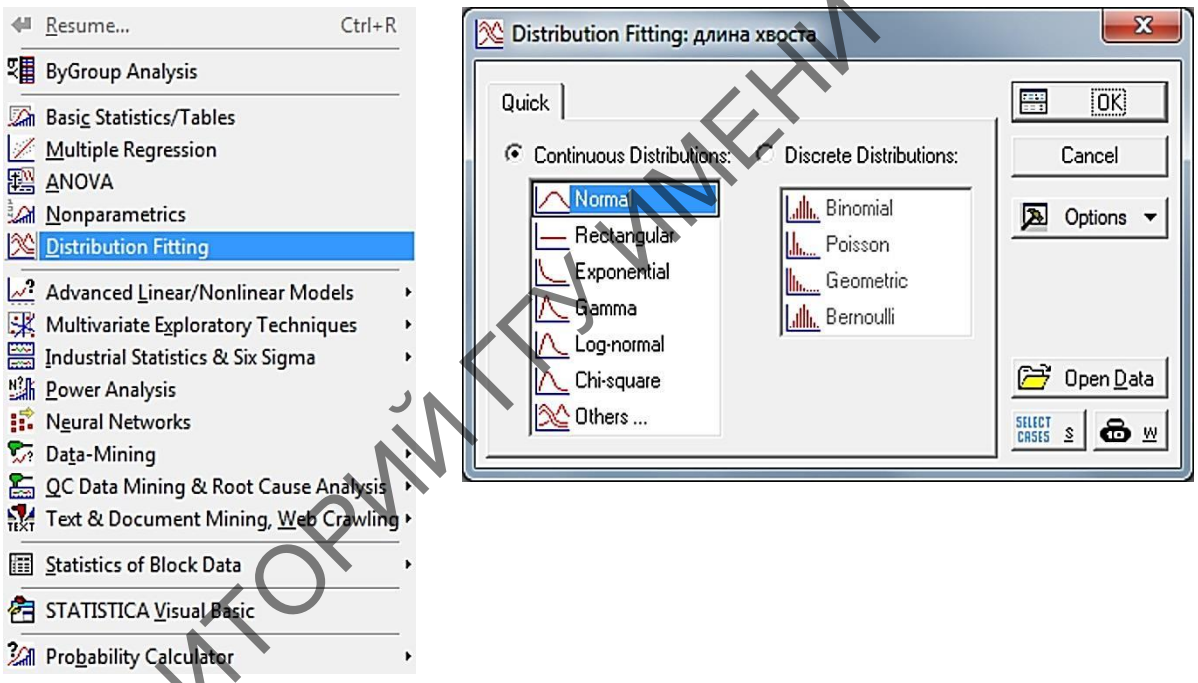
В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (Статистические процедуры) и в нём модуль **Distribution fitting** (Подгонка распределения) (рисунок 2.15А). Он позволяет проверить соответствие анализируемых данных различным математическим распределениям, в том числе и нормальному.

Шаг 3. Проведение анализа.

После выбора модуля **Distribution fitting** (Подгонка распределения) на экране появляется диалоговое окно модуля (рисунок 2.15Б), где находятся две колонки:

– **Continuous Disrtibutions** (Непрерывные распределения) – нормальное, прямое, экспоненциальное, гамма, логарифмически нормальное, хи квадрат и другие;

– **Discrete Disrtibutions** (Дискретные распределения)..



А Б

А – выбор модуля в меню Statistics, Б – выбор типа распределения

Рисунок 2.15 – Модуль **Distribution fitting**

Необходимо выбрать в разделе **Continuous Disrtibutions** (Непрерывные распределения) пункт **Normal** (Нормальное) и нажать **ОК**. Появится диалоговое окно опции **Continuous Disrtibutions** (Непрерывные распределения) (рисунок 2.16), где следует указать

нужную нам переменную для расчёта, нажав кнопку **Variables** (Переменные). Остальные настройки оставить без изменений.

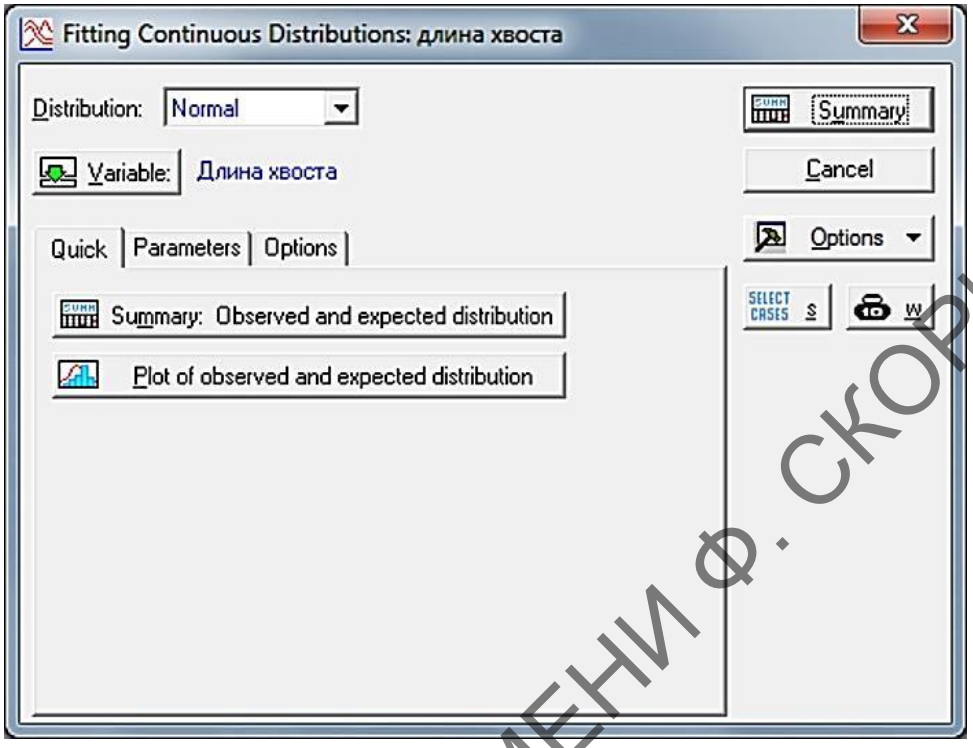


Рисунок 2.16 – Диалоговое окно **Descriptive statistics**, закладка **Quick**

При нажатии на кнопку **Plot of observed and expected distributions** (График наблюдаемого и ожидаемого распределений) получим гистограмму распределения данных о длине хвоста оленьих мышей и колоколообразную кривую красного цвета, соответствующую ожидаемому нормальному распределению (рисунок 2.17). При этом следует сказать, что ожидаемое распределение имеет ту же среднюю арифметическую и то же стандартное отклонение, что и у анализируемой выборки.

Анализируя выведенный график, можно сказать, что в целом распределение длины хвоста оленьих мышей соответствует нормальному (столбики гистограммы образуют фигуру близкую к колоколообразной). Этот вывод, основанный на чисто визуальном анализе распределения, имеет также и более серьёзное подтверждение в виде результатов теста χ^2 (Chi-square test). Они отражены в верхней части графика (рисунок 2.17). Другими словами, при проверке соответствия распределения данных текущей выборки нормальному выдвигается нулевая гитпотеза о том, что наблюдаемое распределение анализируемого признака не отличается от теоретически ожидаемого нормального распределения. Результаты этого теста проверяют

нулевую гипотезу, и выясняется, что если вероятность ошибиться, отклонив эту гипотезу, оказалась больше 0,05 (p = 0,18763), то принимается тот факт, что гипотеза действительно верна. То есть распределение значений длины хвоста оленьих мышей в возрасте одного года статистически не отличается от нормального распределения.

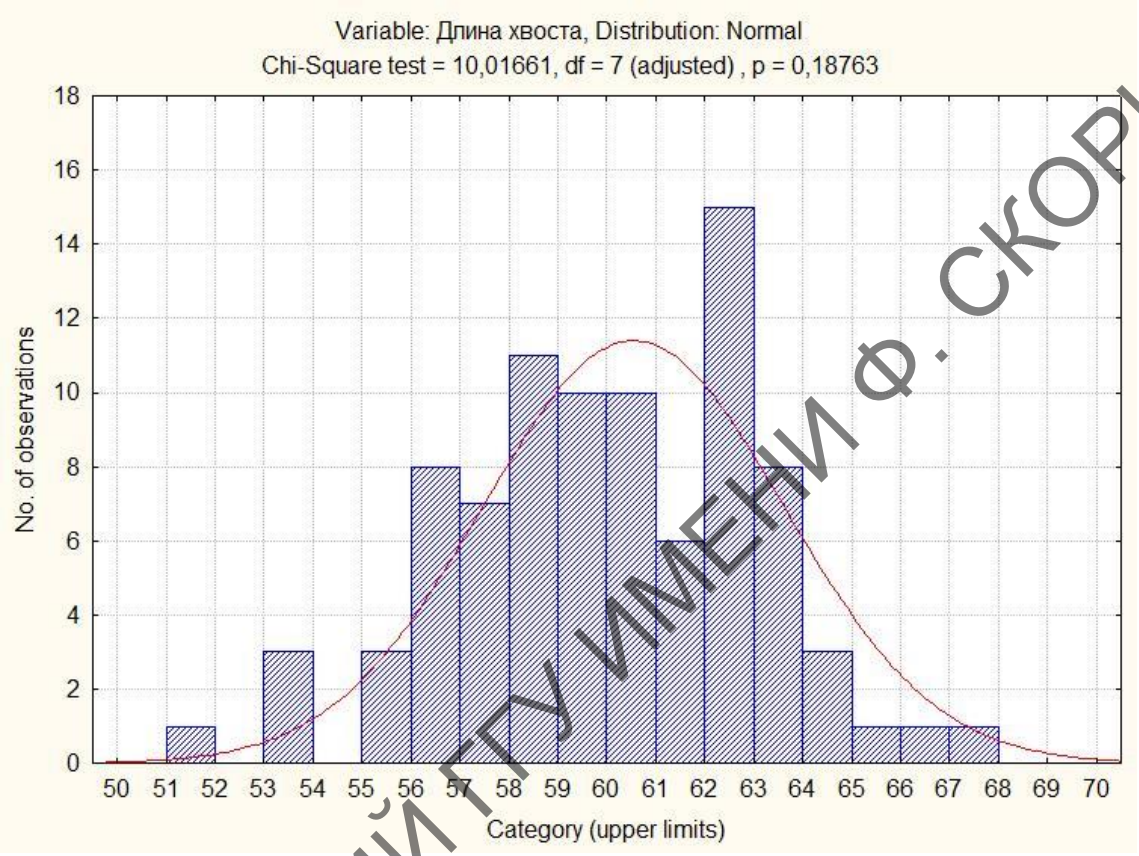


Рисунок 2.17 – Итоги анализа, выполненные при помощи модуля **Distribution fitting**

В то же время следует отметить, что мощность теста χ^2 (хи-квадрат) при проверке нормальности распределения данных выборки не имеет достаточной силы (т. е. его применение может привести к ошибочному выводу о нормальности распределения и годится только для предварительного анализа). Поэтому для серьёзных расчётов лучше воспользоваться другими тестами.

2.2.2 Тесты Колмогорова – Смирнова и Шапиро – Уилка

Рассмотрим работу этих тестов на примере уже известной нам длины хвоста оленьих мышей (пункт 2.2.1).

Шаг 1. Выбор модуля.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (Статистические процедуры), в нём – модуль **Basic Statistics/Tables** (Основные статистические показатели/ Таблицы) и далее – опцию **Descriptive Statistics** (Описательная статистика).

Шаг 2. Выставление параметров проверки на нормальность.

Для этой процедуры необходимо перейти на закладку **Normality** (Нормальность) (рисунок 2.18).

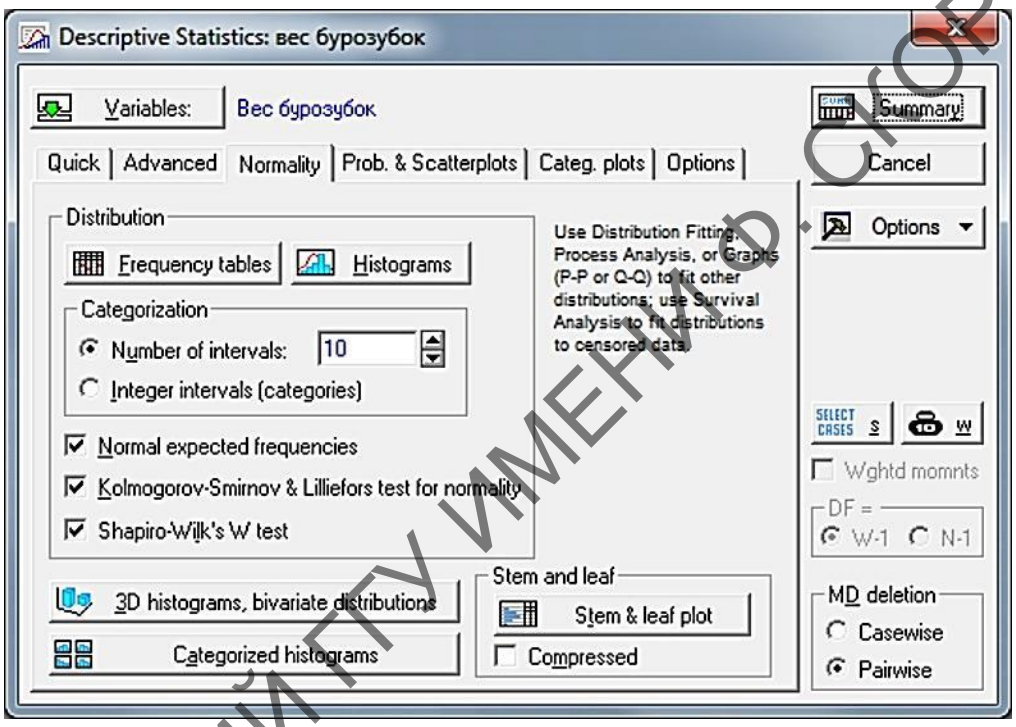


Рисунок 2.18 – Диалоговое окно **Descriptive statistics**, закладка **Normality**

Наиболее значимые элементы этой закладки собраны в поле **Distribution** (Распределение):

– уже известные по предыдущей теме кнопки **Frequency tables** (Таблица частот) и **Histograms** (Гистограммы);

– **Categorization** (Категоризация):

Number of intervals (Количество интервалов) – задаёт количество «столбиков» на гистограмме в случае непрерывности анализируемого биологического признака;

Integer intervals (Categories) (Целочисленные интервалы (Категории)) – задаёт количество «столбиков» на гистограмме в случае дискретности (выраженного только целыми числами) анализируемого биологического признака.

– **Normal expected frequencies** (Ожидаемые нормальные частоты): используется для определения теоретически ожидаемых нормальных частот;

– **Kolmogorov – Smirnov & Lilliefors test for normality** (Тест на нормальность распределения Колмогорова – Смирнова и Лиллифорса) – тест, применяемый для проверки соответствия анализируемых данных закону нормального распределения;

– **Shapiro – Wilk's W test** (W-тест Шапиро – Уилка) – тест, применяемый для проверки соответствия анализируемых данных закону нормального распределения.

Для анализа необходимо выставить необходимые параметры, указать переменную и вид теста (рисунок 2.18).

Шаг 3. Интерпретация результатов.

Для получения результатов анализа после выставления всех необходимых параметров (указания видов тестов, которые необходимо провести) следует нажать кнопку **Histograms** (Гистограммы). Программа выведет на экран результаты анализа (рисунок 2.19).

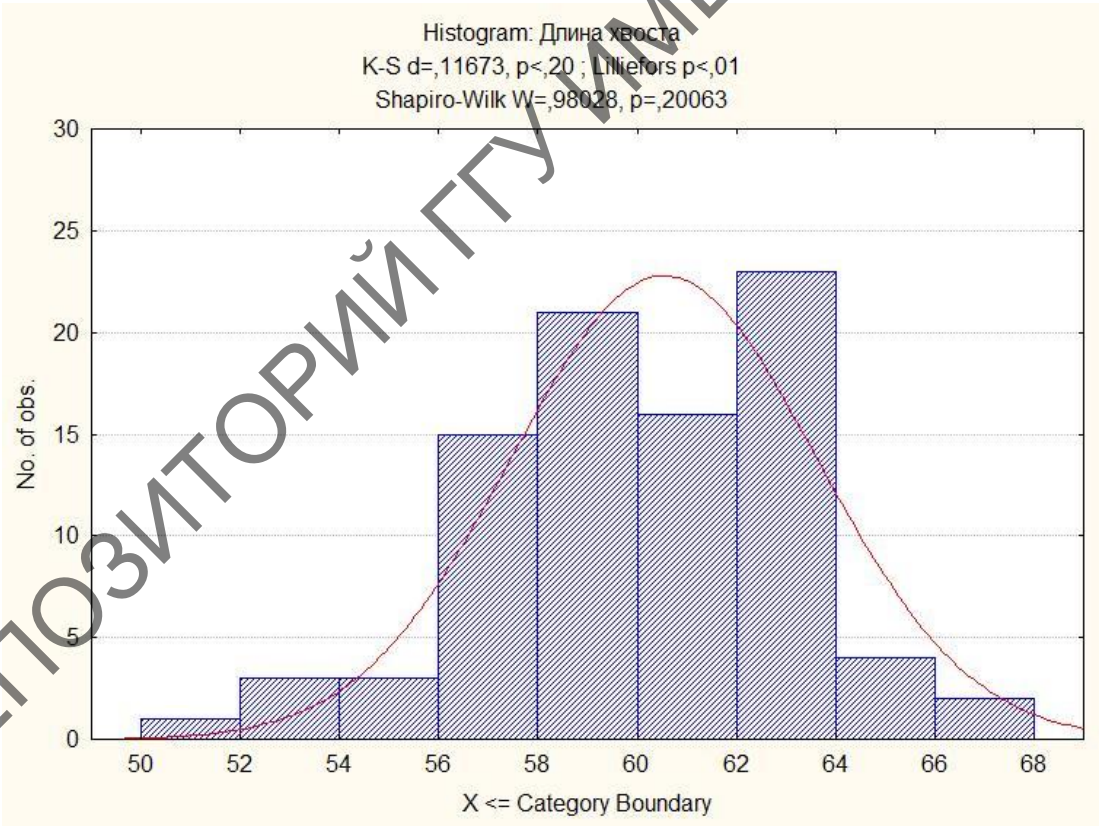


Рисунок 2.19 – Итоги анализа, выполненные при помощи тестов на нормальность

Результаты выбранных тестов на нормальность расположены в заголовке полученного графика. При p > 0,05 можно смело сделать вывод, что рассматриваемое распределение выборки не отличается от нормального. В примере с данными о длине хвоста оленьих мышей для теста Колмогорова – Смирнова получаем p < 0,2, для теста Лиллифорса p < 0,2, а для теста Шапиро – Уилка p = 0,7979 (рисунок 2.19), что подтверждает сделанный ранее вывод о нормальности распределения этих данных.

2.2.3 График нормальных вероятностей

Этот способ проверки данных на нормальность распределения заключается в использовании графика нормальных вероятностей. Такой график изображает зависимость ожидаемых нормальных частот значений признака от их реальных частот. Другими словами, если между наблюдаемым и ожидаемым распределениями нет никакой разницы, то точки на этом графике выстроятся строго вдоль прямой. Иначе – образуют фигуру, отличную от прямой.

Рассмотрим работу этих тестов на примере уже известной нам длины хвоста оленьих мышей (пункт 2.2.1).

Шаг 1. Выбор модуля.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (Статистические процедуры), в нём – модуль **Basic Statistics/Tables** (Основные статистические показатели / Таблицы), далее – опцию **Descriptive Statistics** (Описательная статистика).

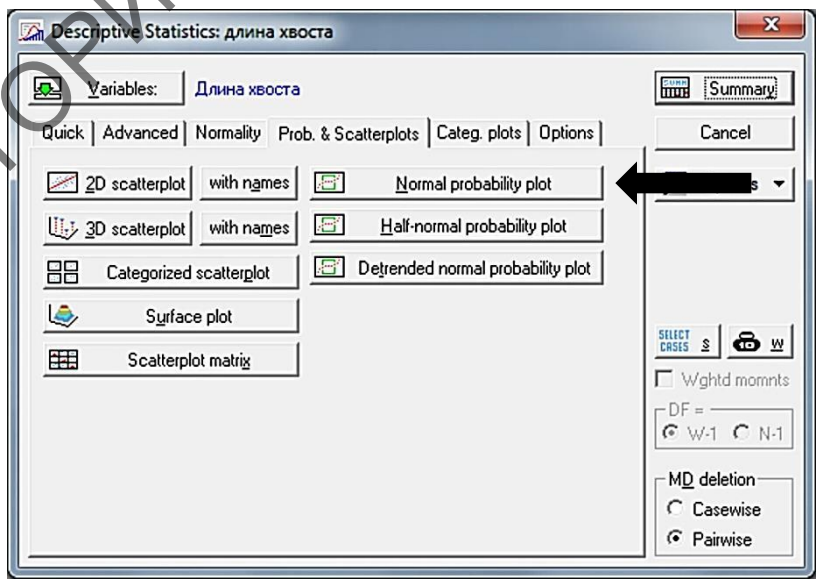


Рисунок 2.20– Диалоговое окно **Descriptive statistics**, закладка **Prob. & Scatterplots**

Шаг 2. Проведение анализа.

Для этой процедуры необходимо перейти на закладку **Prob. & Scatterplots** (Вероятностные графики и диаграммы рассеяния) (рисунок 2.20) и нажать на кнопку **Normal probability plot** (График нормальных вероятностей). В результате на экран будет выведен график (рисунок 2.21).

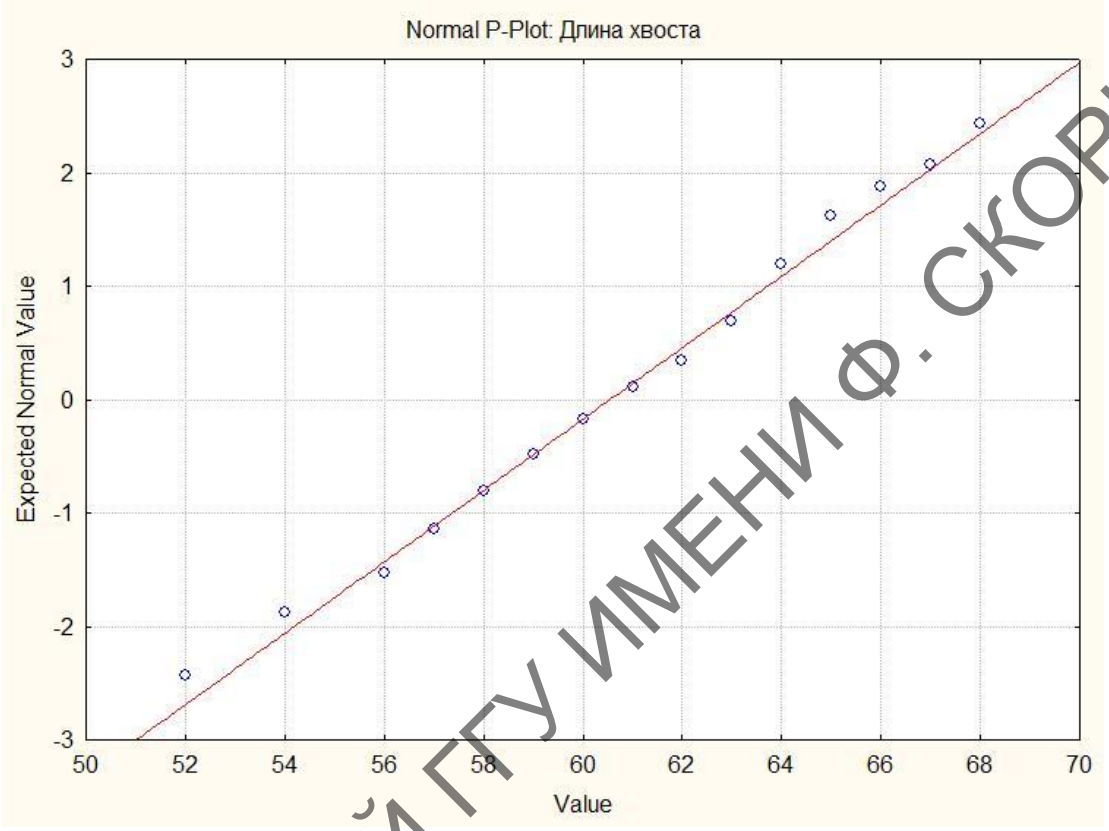


Рисунок 2.21 – Итоги анализа, выполненные при помощи графика нормальных вероятностей

Точки на этом графике достаточно плотно выстраиваются вдоль теоретически ожидаемой прямой, что еще раз подтверждает нормальность распределения данных о длине хвоста оленьих мышей.

2.3 Сравнение выборочных средних

2.3.1 Параметрический *t*-тест Стьюдента для независимых переменных

При проведении биологических исследований довольно часто необходимо проводить сравнение средних арифметических двух групп (экспериментальной и контрольной и т. п.). Для проведения

подобного рода сравнений уже довольно продолжительное время является *t*-тест Стьюдента, или просто «*t*-тест». В качестве нулевой гипотезы при проведении данного теста принимается то допущение, что наблюдаемые различия между средними значениями сравниваемых выборок случайны и не вызваны действием изучаемого фактора (т. е. обе анализируемые выборки принадлежат одной генеральной совокупности).

Следует, однако, учесть тот момент, что тест Стьюдента является параметрическим методом анализа и его корректное применение требует выполнения нескольких условий:

1) обе сравниваемые выборки должны быть независимыми, т. е. свойства одной из них никак не должны быть связаны со свойствами другой и тем более влиять на них;

2) обе выборки должны иметь нормальное распределение;

3) между дисперсиями выборок не должно быть статистически значимой разницы, т. е. должен соблюдаться принцип однородности дисперсий.

Игнорирование перечисленных условий при проведении теста Стьюдента часто приводит к неверной интерпретации полученных результатов и ошибочным выводам по итогам исследований. При этом следует учитывать тот момент, что наиболее значимым является условие нормальности распределения выборок.

2.3.1.1 Парный *t*-тест Стьюдента в Excel

Рассмотрим расчёт парного *t*-теста Стьюдента в Excel на примере данных о количестве экземпляров жужелиц в двух выборках из 10 почвенных ловушек из сходных биотопов, но расположенных на значительной удалённости одна от другой, и, следовательно, выборки никак не влияют друг на друга:

Биотоп 1	**Биотоп 2**
3	15
4	19
5	3
8	2
9	14
1	4
2	5
4	17
5	1
6	11

Шаг 1. Создание электронной таблицы с данными.

Для проведения анализа в Excel необходимо создать в новом файле таблицу, состоящую из 2 столбцов (так как анализируются две переменные) (рисунок 2.22).

	A	B
1	**Биотоп 1**	**Биотоп 2**
2	3	15
3	4	19
4	5	3
5	8	2
6	9	14
7	1	4
8	2	5
9	4	17
10	5	1
11	6	11
12		
13		

Рисунок 2.22 – Сздание рядов данных в книге Excel

Шаг 2. Выбор *t*-теста Стьюдента

Для проведения парного анализа *t*-теста Стьюдента необходимо в пункте меню **Данные** выбрать опцию **Анализ данных**. Затем в появившемся окне выбрать пункт **Парный двухвыборочный t-тест для средних** (рисунок 2.23) и нажать **ОК**.

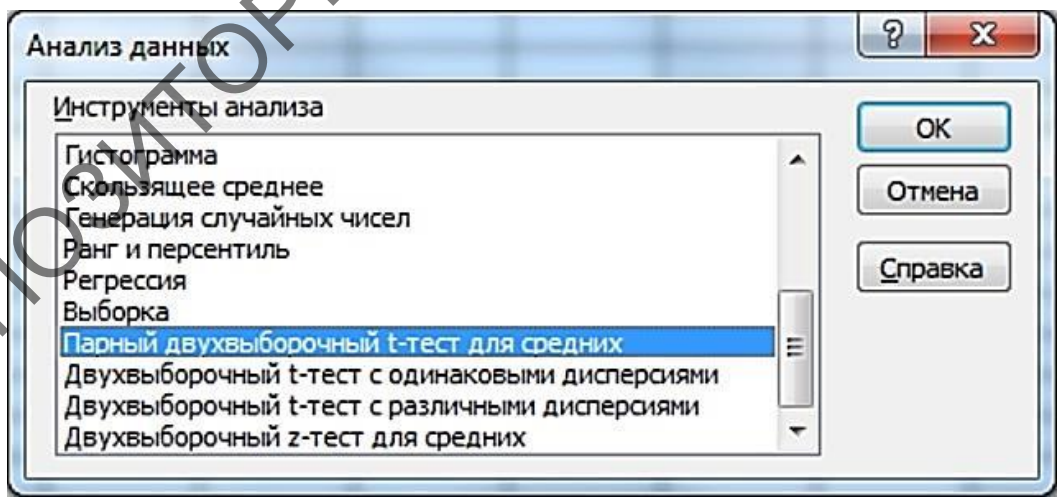


Рисунок 2.23 – Диалоговое окно раздела меню **Анализ данных**

Шаг 3. Расчёт показателей *t*-теста Стьюдента

Появившееся диалоговое окно (рисунок 2.24) имеет ряд параметров:

– **Интервал переменной 1** – ссылка на значения первой выборки.

– **Интервал переменной 2** – ссылка на значения второй выборки;

– **Гипотетическая средняя разность** – укажите гипотетическое значение разности средних (в нашем случае введем 0);

– **Метки** – если в полях **Интервал переменной 1** и **Интервал переменной 2** указаны ссылки вместе с адресами ячеек заголовков столбцов, то эту галочку нужно установить. Ссылку указывать лучше с заголовком, так как в этом случае при выводе результат выглядит нагляднее;

– **Альфа**: уровень значимости (в нашем случае достаточно 95 % точности, поэтому уровень значимости будет 0,05);

– **Выходной интервал** – диапазон ячеек, куда будут помещены результаты вычислений. Достаточно указать левую верхнюю ячейку этого диапазона.

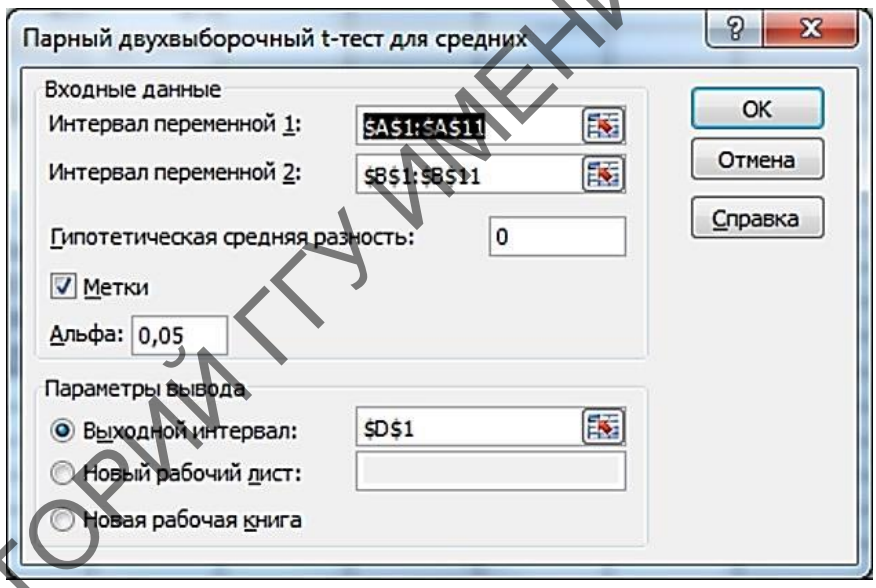


Рисунок 2.24 – Диалоговое окно
Парный двухвыборочный t-тест для средних

Выставьте все параметры, как показано на рисунке 2.24, и нажмите **ОК.**

Шаг 4. Интерпретация результатов.

Как итог анализа в указанном месте будет отображена таблица с результатами (рисунок 2.25), в которой наибольшее значение для оценки результата имеет ячейка **P(T <= t) двухстороннее**.

Парный двухвыборочный t-тест для средних		
	Биотоп 1	*Биотоп 2*
Среднее	4,7	9,1
Дисперсия	6,233333	46,54444
Наблюдения	10	10
Корреляция Пирсона	0,021527	
Гипотетическая разность средних	0	
df	9	
t-статистика	-1,9287	
P(T<=t) одностороннее	0,042925	
t критическое одностороннее	1,833113	
P(T<=t) двухстороннее	0,085851	
t критическое двухстороннее	2,262157	

Рисунок 2.25 – Итоги анализа **Парный двухвыборочный *t*-тест для средних**

В текущем примере оно равно 0,008, что больше чем 0,005. Следовательно, статистически значимой разницы между двумя анализируемыми выборками нет.

2.3.1.2 Проведение *t*-теста Стьюдента в STATISTICA

Для примера расчёта проанализируем те же данные о количестве экземпляров жужелиц в ловушках на двух стационарах.

Шаг 1. Составление электронной таблицы с данными.

При составлении этой таблицы необходимо учитывать, что переменных для анализа будет уже две, а не одна, как в предыдущих заданиях. Данные можно представить двумя способами: в виде независимых переменных и с использованием группирующей переменной (рисунок 2.26).

Data: Spreadsheet2*...

	1 Биотоп 1	2 Биотоп 2
1	3	15
2	4	19
3	5	3
4	8	2
5	9	14
6	1	4
7	2	5
8	4	17
9	5	1
10	6	11

А

Data: Spreadsheet2* (2...

	1 Биотоп	2 Количество жужелиц
1	1	3
2	1	4
3	1	5
4	1	8
5	1	9
6	1	1
7	1	2
8	1	4
9	1	5
10	1	6
11	2	15
12	2	19
13	2	3
14	2	2
15	2	14
16	2	4
17	2	5
18	2	17
19	2	1
20	2	11

Б

А – в виде независимых переменных; Б – с группирующей переменной

Рисунок 2.26 – Представление данных для анализа

Рассмотрим оба способа анализа.

1) С отражением данных в виде двух независимых переменных.

Шаг 2. Выбор вида анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (Статистические процедуры), в нём – модуль **Basic Statistics / Tables** (Основные статистические показатели/ Таблицы), далее – опцию **t-test, independent, by variables** (*t*-тест независимых переменных по переменным) (рисунок 2.27) и нажать **ОК**.

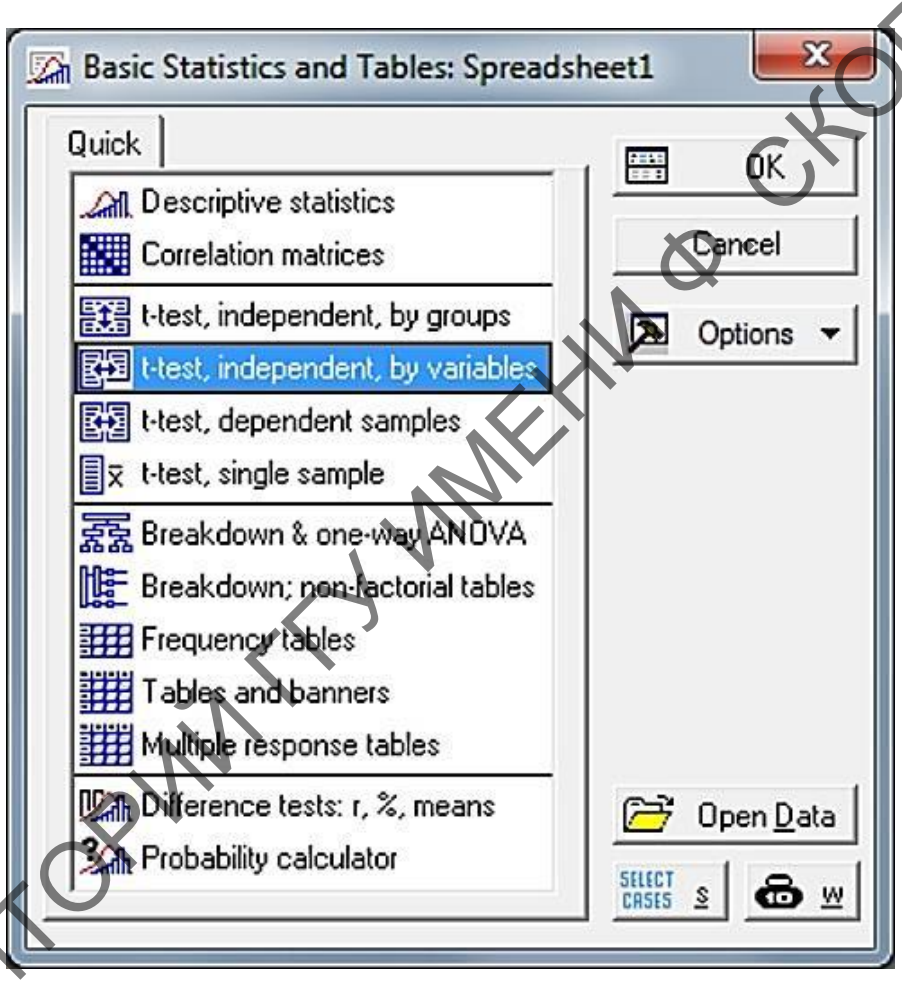


Рисунок 2.27 – Выбор опции **t-test, independent, by variables**

Шаг 3. Установка параметров анализа.

В появившемся диалоговом окне **T-test for Independent Samples by Variables** (Т-тест для независимых переменных) (рисунок 2.28) необходимо указать переменные для анализа (остальные настройки оставить без изменений), после чего нажать кнопку **Summary: T-tests** (Результаты: Т-тесты).

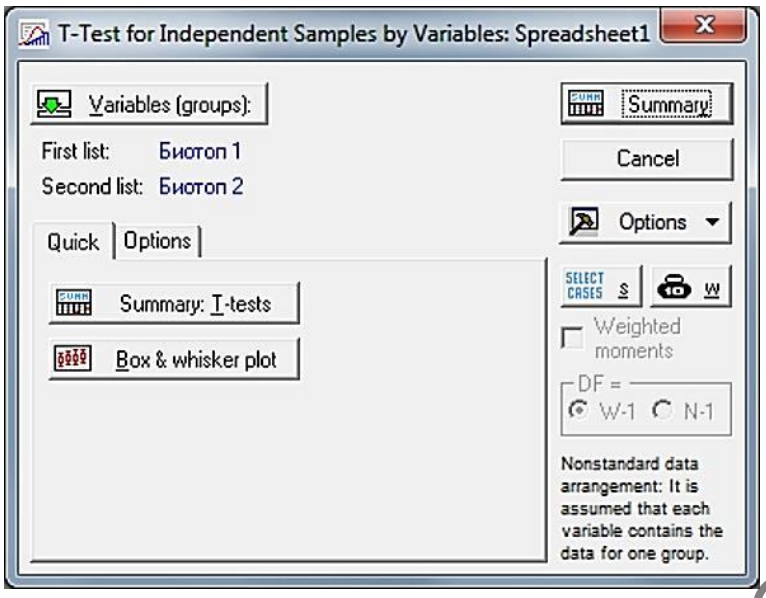


Рисунок 2.28 – Диалоговое окно **T-test for Independent Samples by Variables**

Шаг 4. Итоги анализа.

Итоги анализа будут отражены в таблице (рисунок 2.29).

	T-test for Independent Samples (Spreadsheet1) Note: Variables were treated as independent samples									
Group 1 vs. Group 2	**Mean Group 1**	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
Биотоп 1 vs. Биотоп 2	4,700000	9,100000	-1,91526	18	0,071486	10	10	2,496664	6,822349	7,467023	0,006216

Рисунок 2.29 – Итоговая таблица анализа T-test for Independent Samples

Данная таблица содержит ряд столбцов с результатами:

– **Mean Group 1**: среднее значение жужелиц в почвенных ловушках в Биотопе 1;

– **Mean Group 2**: среднее значение жужелиц в почвенных ловушках в Биотопе 2;

– **t-value**: значение рассчитанного программой t-критерия Стьюдента;

– **df**: число степеней свободы;

– **p**: уровень значимости. По сути, это наиболее интересующий нас результат анализа. В данном случае $p > 0,05$, на основании чего можно сделать вывод об отсутствии статистически значимых различий между средними значениями численности жужелиц из разных биотопов;

– **Valid N Group 1**: объем выборки «Биотоп 1»;

– **Valid N Group 2**: объем выборки «Биотоп 2»;

– **Std. dev. Group 1**: стандартное отклонение выборки «Биотоп 1»;

– **Std. dev. Group 2**: стандартное отклонение выборки «Биотоп 2»;

– **F-ratio, Variances**: значение F-критерия Фишера, с помощью которого проверяется равенство дисперсий в сравниваемых выборках (одно из условий применения теста Стьюдента);

– **p, Variances**: вероятность ошибки для F-теста Фишера. Поскольку в нашем случае $p < 0,05$, то можно заключить, что дисперсии сравниваемых выборок различаются (т. е. условие однородности дисперсий не выполняется и тест в данном случае проводить некорректно).

Шаг 5. Графические итоги анализа.

Для графического отображения анализа в виде диаграммы размахов (box-whisker plots) в диалоговом окне **T-test for Independent Samples by Variables** (Т-тест для независимых переменных) после указания переменных необходимо нажать кнопку **Box-whisker plots** (Диаграммы размахов). Результат отражён на рисунке 2.30.

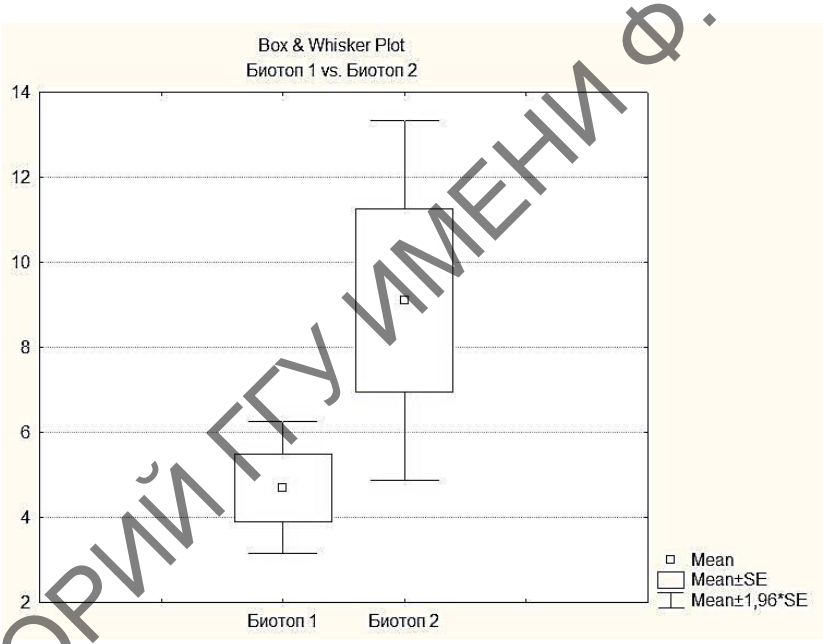


Рисунок 2.30 – Итоговый график анализа **T-test for Independent Samples**

2) С группирующей переменной.

Шаг 6. Выбор вида анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **Basic Statistics/Tables** (Основные статистические показатели / Таблицы), далее – опцию **t-test, independent, by groups** (t-тест для независимых переменных с группирующей переменной) (рисунок 2.31) и нажать **ОК**.

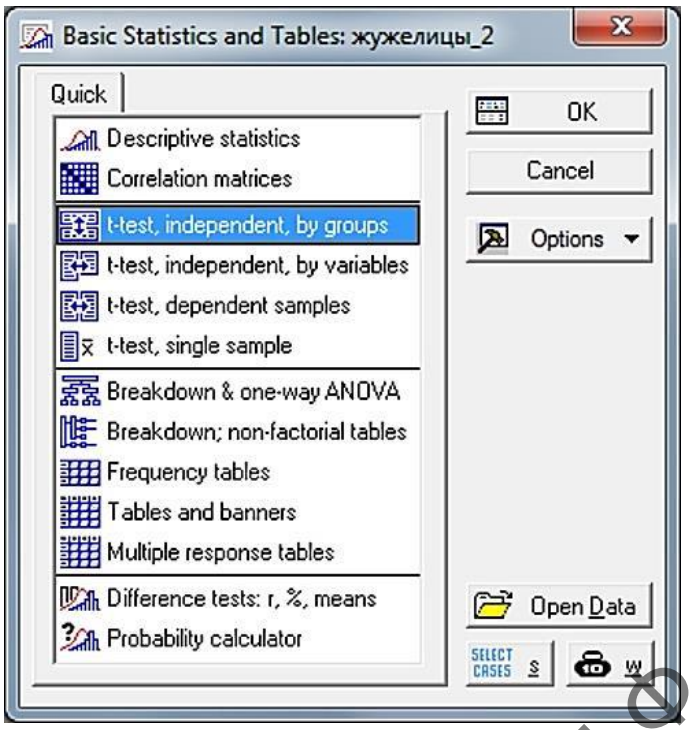


Рисунок 2.31 – Выбор опции **T-test, independent, by groups**

Шаг 7. Указание переменных.

В появившемся диалоговом окне **T-test for Independent Samples by Groups** (Т-тест для независимых переменных с группирующей переменной) необходимо указать переменные для анализа, нажав кнопку **Variables** (Переменные) (рисунок 2.32).

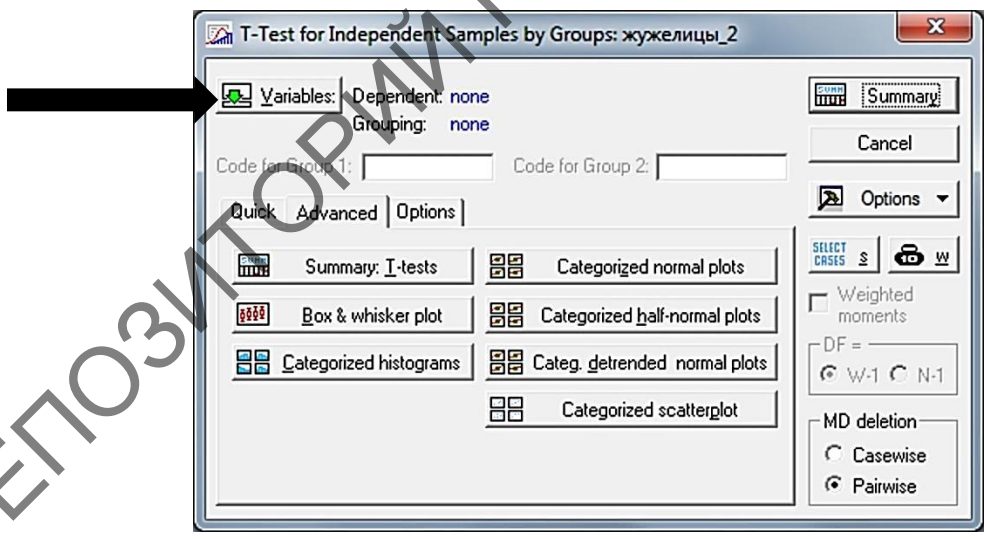


Рисунок 2.32 – Диалоговое окно **T-test for Independent Samples by Groups**

Далее в появившемся диалоговом окне **Select the dependent variables and one grouping variable** (Выбор зависимых переменных и од-

ной группирующей переменной) необходимо указать, что в данном случае зависимой переменной (слева) будет переменная **Количество жужелиц**, а группирующей (справа) – **Биотоп** (рисунок 2.33). После указания переменных следует нажать кнопку **ОК** и вернуться в предыдущее диалоговое окно.

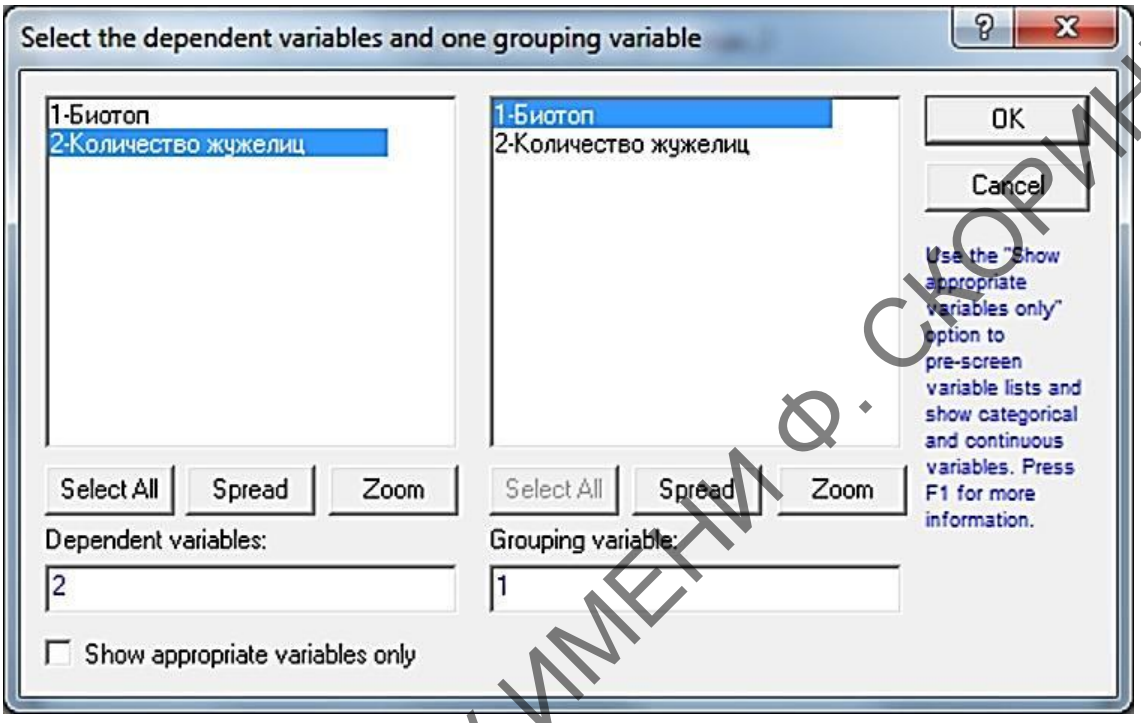


Рисунок 2.33 – Диалоговое окно **Select the dependent variables and one grouping variable**

Шаг 8. Установка параметров анализа

В появившемся диалоговом окне **T-test for Independent Samples by Groups** (Т-тест для независимых переменных с группирующей переменной) необходимо нажать кнопку **Summary: T-tests** (Результаты: Т-тесты). На экране будут отражены результаты, сходные с таблицей на рисунке 2.29.

Шаг 9. Графические итоги анализа.

Для графического отображения анализа в виде диаграммы размахов (box-whisker plots) в диалоговом окне **T-test for Independent Samples by Groups** (Т-тест для независимых переменных с группирующей переменной) после указания переменных необходимо нажать кнопку **Box-whisker plots** (Диаграммы размахов). После этого программой будет предложен выбор типа диаграммы. Следует выбрать вариант **Mean/SE/SD** (Средняя арифметическая/Стандартная ошиб-

ка/Стандартное отклонение) (рисунок 2.34) и нажать кнопку **ОК.** Результат будет отражён на рисунке 2.35.

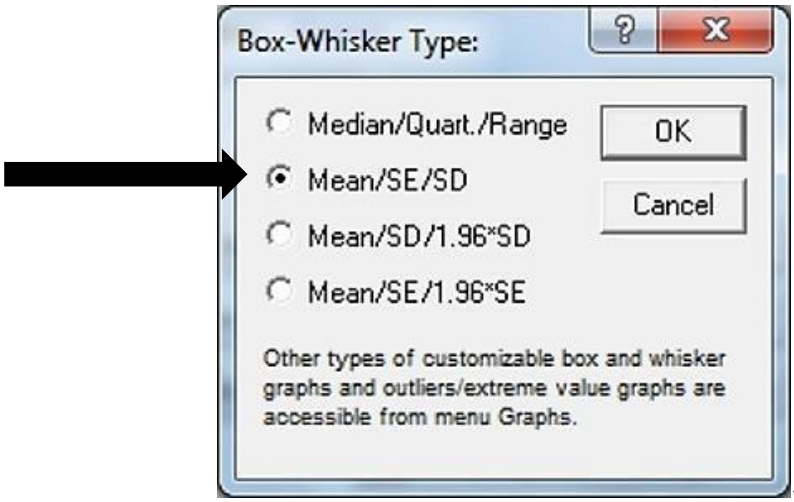


Рисунок 2.34 – Диалоговое окно **Box-Whisker Type**

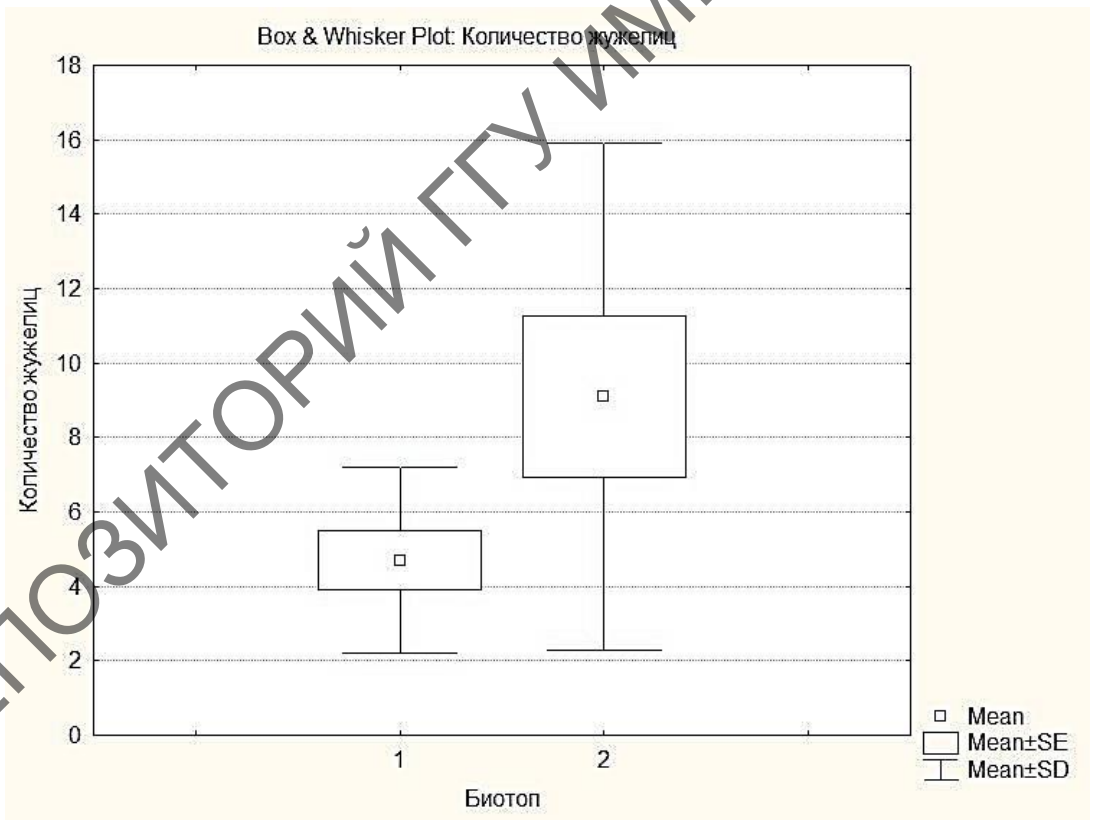


Рисунок 2.35 – Итоговый график анализа **T-test for Independent Samples**

2.3.2 Сравнение независимых переменных, не подчиняющихся закону нормального распределения

Если значения признака в двух сравниваемых группах распределены ненормально, то применение параметрического t-теста для оценки их сходства или различия между собой будет часто приводить к неверным результатам. В таких случаях следует воспользоваться непараметрическим аналогом теста Стьюдента – U-тестом Манна – Уитни (Mann – Whitney U-test).

В качестве примера рассмотрим те же данные по количеству жужелиц. При этом в учебных целях будем считать их распределение ненормальным.

Шаг 1. Составление электронной таблицы с данными.

Для наших расчётов необходимо использовать ранее составленную электронную таблицу с данными с использованием группирующей переменной (рисунок 2.26Б).

Шаг 2. Выбор вида анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (Статистические процедуры), в нём – модуль **Nonparametrics** (Непараметрические методы) (рисунок 2.36), а затем – **Comparing two independent samples (groups)** (Сравнение двух независимых выборок с группирующей переменной) (рисунок 2.37) и нажать **ОК**.

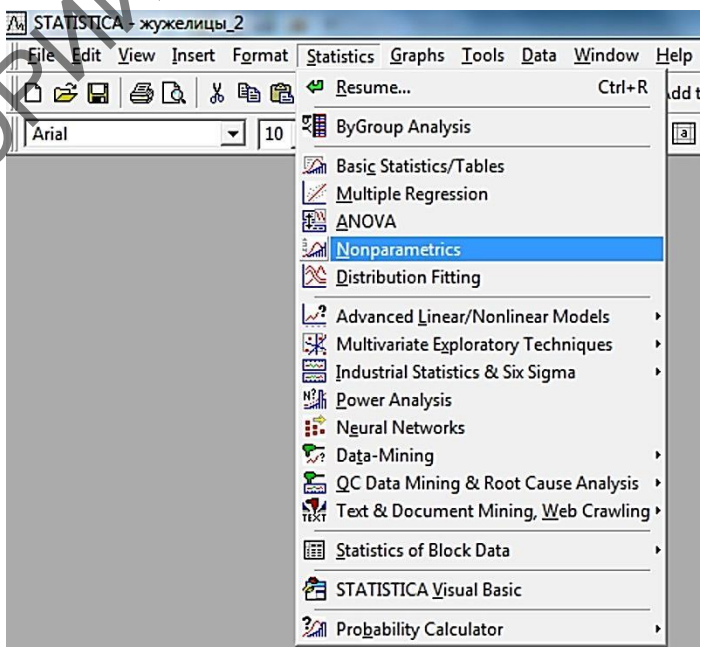


Рисунок 2.36 – Выбор непараметрических методов анализа

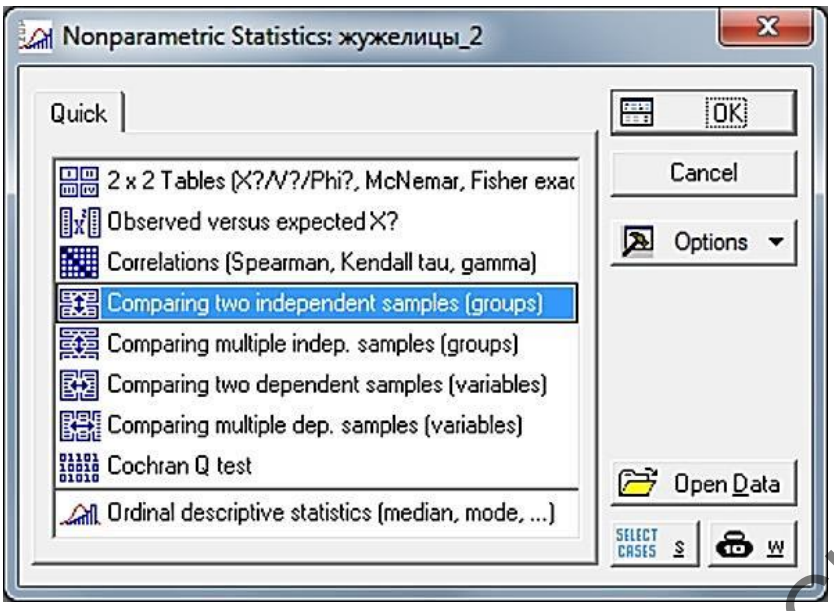


Рисунок 2.37 – Диалоговое окно **Nonparametric Statistics**

Шаг 3. Проведение анализа.

В появившемся диалоговом окне **Comparing Two Groups** (Сравнение двух групп) (рисунок 2.38) необходимо указать переменные, по аналогии с тем, как это производилось для параметрического Т-теста (рисунок 2.33), после чего нажать кнопку **Mann – Whitney U-test** (U-тест Манна – Уитни) или **M-W U test** (рисунок 2.38).

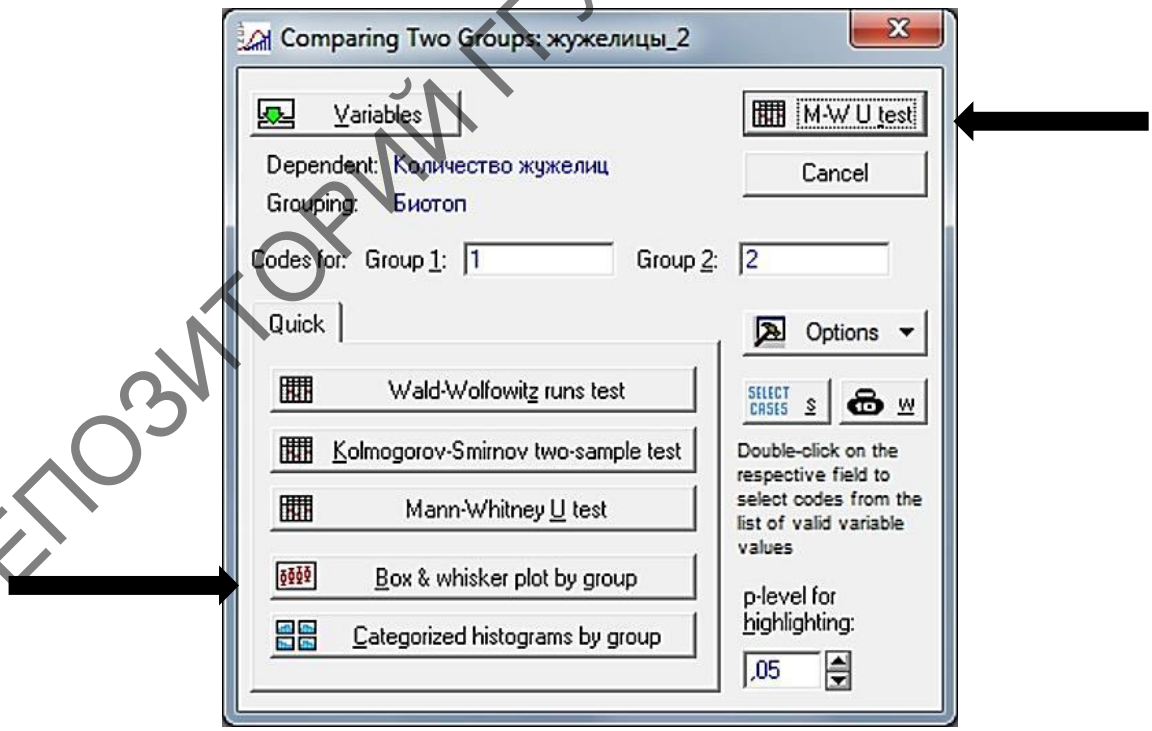


Рисунок 2.38 – Диалоговое окно **Comparing Two Groups**

Шаг 4. Интерпретация результатов.

Таблица с итоговыми результатами показана на рисунке 2.39.

Mann-Whitney U Test (жужелицы_2)
By variable Биотоп
Marked tests are significant at p <,05000

variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2	2*1sided exact p
Количество жужелиц	90,50000	119,5000	35,50000	-1,09610	0,273037	-1,10066	0,271046	10	10	0,279861

Рисунок 2.39 – Итоговая таблица Mann – Whitney U-test

Она имеет столбцы со следующими показателями:

– Rank Sum Group 1: сумма рангов группы 1 (ранг – это положение определенного значения изучаемого признака в упорядоченном по убыванию или возрастанию ряду);

– Rank Sum Group 2: сумма рангов группы 2;

– U: значение U-теста Манна – Уитни;

– Z: значение Z функции Фишера;

– p-level: уровень значимости U-теста Манна – Уитни при выборках > 20;

– Z-adjusted: уточнённое значение Z функции Фишера;

– p-level: уровень значимости U-теста Манна – Уитни при выборках <20;

– Valid N Group 1: объём выборки группы 1;

– Valid N Group 2: объём выборки группы 2;

– 2*1 sided exact p: строгий уровень значимости.

В нашем случае (объём выборки < 20) следует обратить внимание в итоговой таблице теста на величину вероятности ошибки p в седьмом столбце. При $p < 0,05$ делается вывод о наличии статистически значимой разницы между сравниваемыми выборками. Иначе (как в данном случае) – об её отсутствии.

2.3.3 Параметрический *t*-тест Стьюдента для зависимых переменных

С зависимыми выборками мы встречаемся тогда, когда измерения значений изучаемого признака выполняются на одних и тех же объектах. Рассмотрим пример.

Был проведен опыт по подкормке 32 свиноматок препаратом «Афаром», содержавшим железо и медь, в целях уменьшения доли мертворожденных поросят (таблица 2.2).

От каждой матки получали 1 опорос, когда добавляли в корм «Афаром», и 1 опорос контрольный, когда добавки препарата не было. Маток покрывали всегда одними и теми же хряками. Получены результаты.

Таблица 2.2 – Результаты внесения в корм препарата «Афаром»

Номер матки	Мертворожденные, %		Номер матки	Мертворожденные, %	
	«Афаром»	«Контроль»		«Афаром»	«Контроль»
1	0	8,3	12	11,1	0
2	0	12,5	13	11,1	0
3	0	9,1	14	0	25,0
4	18,2	22,2	15	0	9,1
5	0	10,0	16	0	14,3
6	25,0	33,3	17	0	35,7
7	10,0	0	18	0	63,6
8	11,1	0	19	0	9,1
9	0	16,7	20	0	10,0
10	0	28,6	21	22,2	40,0
11	0	25,0	22	0	0

Необходимо установить, достоверна ли разница в доле мертворожденных поросят опытной и контрольной группы?

В данном случае доля мертворожденных поросят является зависимыми переменными, так как доля может зависеть от принятия или непринятия свиноматками препарата «Афаром». В учебных целях будем считать, что условия для проведения параметрического теста выполняются.

Шаг 1. Составление электронной таблицы с данными.

Электронная таблица данных составляется с учётов двух переменных: «Афаром» и «Контроль» (рисунок 2.40).

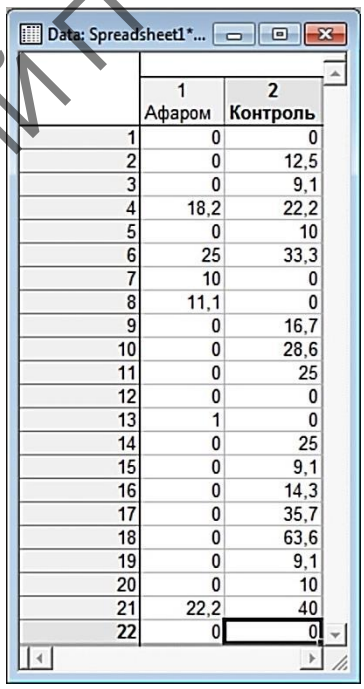


Рисунок 2.40 – Электронная таблица данных по свиноматкам

Шаг 2. Выбор вида анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (Статистические процедуры), в нём – модуль **Basic Statistics/Tables** (Основные статистические показатели/ Таблицы), далее – опцию **t-test, dependent samles** (t-тест зависимых переменных) (рисунок 2.41) и нажать **ОК**.

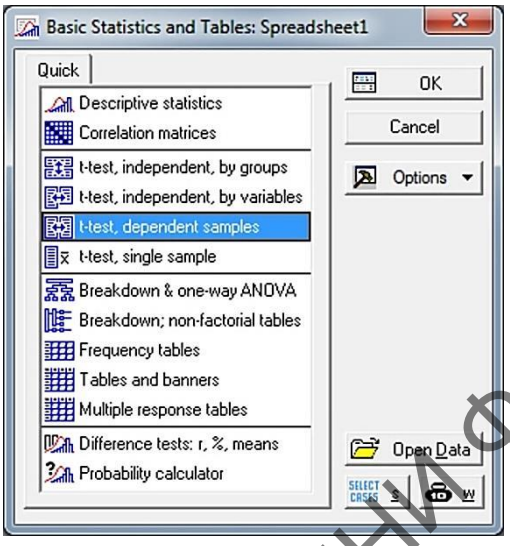


Рисунок 2.41 – Выбор опции **t-test, dependent variables**

Шаг 3. Установка параметров анализа.

В появившемся диалоговом окне **T-test for Dependent Samples** (Т-тест для зависимых переменных) (рисунок 2.42) необходимо указать переменные для анализа (первую – **First variable,** и вторую – **Second variable**), после чего нажать кнопку **Summary: T-tests** (Результаты: Т-тесты).

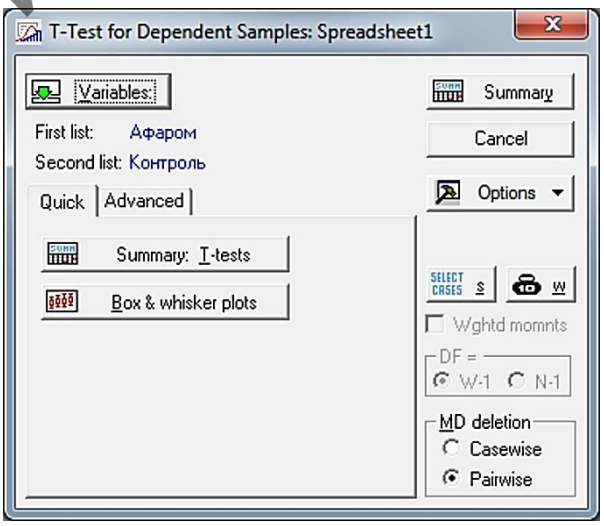


Рисунок 2.42 – Диалоговое окно **T-test for Dependent Samples**

Шаг 4. Итоги анализа.

Итоги анализа будут отражены в таблице (рисунок 2.43).

T-test for Dependent Samples (Spreadsheet1)
Marked differences are significant at p < ,05000

Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
Афаром	3,97727	7,94409						
Контроль	16,55455	16,39982	22	-12,5773	16,36093	-3,60570	21	0,001661

Рисунок 2.43 – Итоговая таблица анализа **T-test for Dependent Samples**

Эта таблица содержит следующие столбцы с результатами:

– **Mean** – средние значения доли рождения мертворожденых поросят для каждой из сравниваемых групп;

– **Std. dv.** – стандартные отклонения для каждой группы;

– **N** – число наблюдений;

– **Diff.** – средняя разница доли рождения мертворожденых поросят;

– **Std. dv. diff.** – стандартное отклонение для средней разницы;

– **t** – значение t-критерия;

– **df** – число степеней свободы;

– **p** – уровень значимости, т. е. вероятность ошибочно отвергнуть нулевую гипотезу о том, что средние величины доли рождения мертворожденых поросят в сравниваемых группах не различаются. Поскольку в нашем случае $p < 0,05$, то можно смело заключить, что средние значения доли рождения мертворожденых поросят при использовании лекарства статистически различаются. Обратите внимание, что при наличии такого рода различий результаты анализа в программе STATISTICA обычно выделяются красным цветом (как и в нашем случае).

2.3.4 Сравнение зависимых переменных, не подчиняющихся закону нормального распределения

В случае, если две зависимые выборки распределены ненормально, для их сравнения следует применить *парный тест Уилкоксона* (**Wilcoxon matched pair test**). В качестве примера рассмотрим предыдущие данные по приёму лекарства свиноматками. Для учебных целей допустим, что распределение ненормальное.

Шаг 1. Составление электронной таблицы с данными.

Необходимо использовать ранее созданную электронную таблицу.

Шаг 2. Выбор вида анализа.

В главном меню программы STATISTICA нужно выбрать пункт **Statistics** (Статистические процедуры), в нём – модуль **Nonparametrics** (Непараметрические методы) (рисунок 2.36), а затем – **Comparing dependent samples (variables)** (Сравнение двух зависимых выборок (переменных)) (рисунок 2.44) и нажать **ОК**.

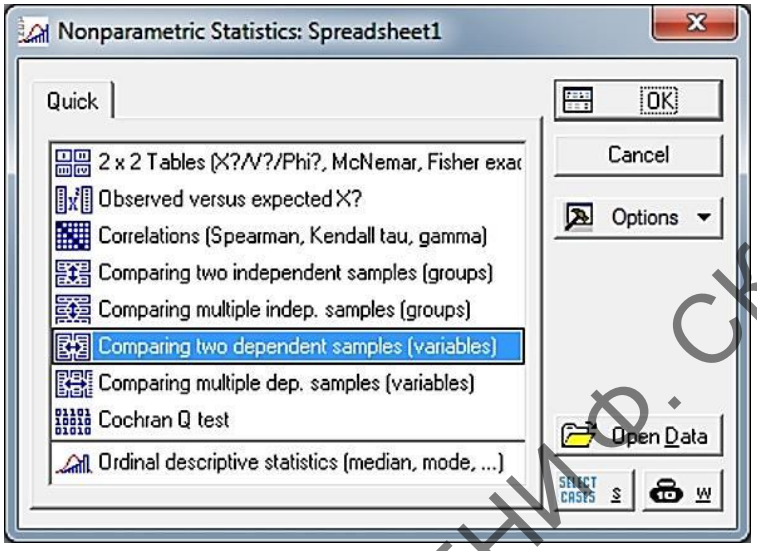


Рисунок 2.44 – Диалоговое окно **Nonparametric Statistics**

Шаг 3. Проведение анализа.

В появившемся диалоговом окне **Comparing Two Variables** (Сравнение двух переменных) (рисунок 2.45) необходимо указать переменные, по аналогии с тем, как это производилось для параметрического Т-теста (рисунок 2.33). После этого нажать кнопку **Wilcoxon matched pair test** (парный тест Уилкоксона) (рисунок 2.45).

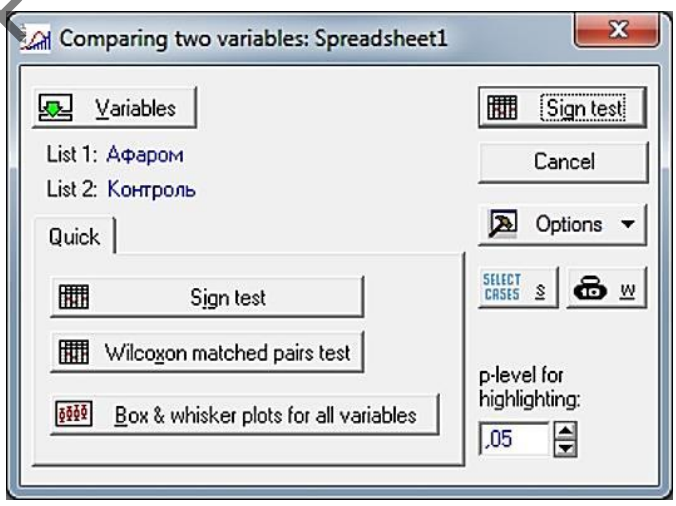


Рисунок 2.45 – Диалоговое окно **Comparing Two Groups**

Шаг 4. Итоги анализа.

Итоги анализа будут отражены в таблице (рисунок 2.46).

Wilcoxon Matched Pairs Test (Spreadsheet1) Marked tests are significant at p <,05000				
Pair of Variables	Valid N	T	Z	p-level
Афаром & Контроль	22	19,00000	3,058406	0,002225

Рисунок 2.46 – Итоговая таблица анализа **Wilcoxon matched pair test**

В итоговой таблице необходимо найти величину *p*. В том случае, если $p < 0,05$ (как в нашем примере), можно сделать вывод о наличии статистически значимой разницы между сравниваемыми выборками. Обратите внимание, что результаты подсвечены красным цветом!

Задания для самоконтроля

1) Изучен живой вес 63 телят холмогорских помесей при рождении (в кг):

27	32	32	31	32	28	37	35	26	28	36	39	36	36	24	27
32	39	34	30	37	26	27	40	35	37	32	36	28	26	32	31
28	43	26	35	45	26	35	32	32	35	28	33	30	38	32	30
35	28	32	36	32	36	37	33	23	31	33	30	23	34	36	

Рассчитайте параметры описательной статистки в Excel и STATISTICA, постройте гистограмму.

2) Были получены данные о длине третьего верхнего предкоренного зуба у 21 экземпляра ископаемого млекопитающего *Ptilodus montanus* (в мм):

3,2	2,8	2,9	3,0	3,1	3,3	2,9
3,1	2,7	3,4	2,9	3,0	2,9	2,8
2,6	3,0	2,8	3,0	3,1	2,9	3,0

Рассчитайте параметры описательной статистки в Excel и STATISTICA, постройте гистограмму.

3) Имеются следующие данные о росте (в см) взрослых мужчин:

162	151	161	170	167	164	166	164	173	172
165	153	164	169	170	154	163	159	161	167
168	164	170	166	176	157	159	158	160	161
167	155	166	167	173	165	175	165	174	167
170	169	159	159	160	156	161	162	161	181
159	169	160	169	161	161	166	164	170	180
158	167	169	165	166	172	168	171	178	178
171	165	161	162	182	164	171	169	176	177
170	169	171	160	165	165	179	161	178	173

Проверьте тремя различными способами, подчиняется ли данная выборка закону нормального распределения.

4) У баранов мериносовой породы были произведены промеры рогов (в см):

47	53	50	56	49	52	51	58	55	61
50	48	51	51	48	60	51	57	57	59
51	54	52	58	50	51	51	58	53	57
52	49	59	61	50	52	51	63	62	70
54	53	54	68	54	63	64	57	57	58
60	57	60	69	57	56	54	54	55	57

Проверьте тремя различными способами, подчиняется ли данная выборка закону нормального распределения.

5) Были получены следующие данные о весе тушканчиков (*Dipus aegyptius*):

Самцы	186	190	165	182	182	182	180
	173	157	179	164	146	173	144
	156	156	165	160	160	161	144
	153	152	151	173			
Самки	162	163	190	188	147	146	145
	157	162	186	175	147	145	145
	155	174	180	148	175	145	144
	153	165	141	164			

Сравните эти две независимые выборки между собой различными способами в Excel и STATISTICA. Выясните, есть ли различия в весе самцов и самок тушканчиков?

6) Были изучены две выборки численности жесткокрылых на участке до посева газонной травы (А) и после (Б):

А	2	0	1	0	19	2	11	16	0	0	3	0	0	0	5	1	0
Б	1	1	14	1	11	3	3	30	1	20	5	1	2	1	16	1	5

Сравните эти две зависимые выборки между собой различными способами в STATISTICA. Выясните, есть ли различия в численности жесткокрылых до и после посадки травы?

7) Была изучена длина двухнедельных проростков кукурузы (в см) на участке до внесения удобрений (a) и после (b):

a	24	16	20	17	17	15	21	18	17	12	12	12	15	30	33	15	32	40	22	25
b	22	23	17	21	8	19	29	21	20	13	24	20	19	30	26	23	32	11	21	14

Сравните эти две зависимые выборки между собой различными способами в STATISTICA. Выясните, есть ли различия в длине проростков кукурузы до и после внесения удобрений?

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

4 Мастицкий, С. Э. Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований / С. Э. Мастицкий. – Минск : РУП «Институт рыбного хозяйства», 2009. – 76 с.

ТЕМА 3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ В EXCEL И STATISTICA 7.0

3.1 Проведение корреляционного анализа в Excel.

3.2 Параметрический корреляционный анализ Пирсона в STATISTICA.

3.3 Непараметрический корреляционный анализ Спирмена в STATISTICA.

3.1 Проведение корреляционного анализа в Excel

Корреляционный анализ позволяет сделать заключение о наличии (или отсутствии) связи между анализируемыми признаками. При этом он показывает не только то, какова связь между двумя признаками по направлению (прямая или обратная), но и выражает её количественно при помощи коэффициента корреляции (r) – величины, изменяющейся от -1 до +1 (чем ближе коэффициент к 1, тем сильнее связь между признаками, а знак коэффициента корреляции указывает на направление зависимости).

Для примера проведения корреляционного анализа в электронных таблицах Excel используем данные двух признаков, обозначенных как *X* и *Y* (таблица 3.1).

Таблица 3.1 – Данные для корреляционного анализа

X	*Y*	*X*	*Y*	*X*	*Y*
3,4	14,3	8,4	19,8	10,7	21,3
3,6	14,9	8,5	19,9	11,6	21,3
4,5	17,3	8,8	19,9	12	21,8
4,8	17,3	8,9	20,1	12,3	22,0
4,9	17,4	8,9	20,1	12,6	22,1
5,2	17,5	8,9	20,1	12,7	22,4
5,4	17,6	8,9	20,1	13,3	22,7
5,7	17,6	9,0	20,2	13,6	23,5
6,2	17,6	9,0	20,3	13,8	24,2
6,7	17,8	9,1	20,3	14,0	24,4
7,1	18,0	9,3	20,5	15,0	25,2
7,5	18,0	9,4	20,6	15,2	25,2
7,7	18,1	9,7	20,9	15,8	25,3
7,8	18,1	9,7	21,0	15,9	25,7
7,9	18,6	9,9	21,1	16,6	26,8
8,0	19,7	10,1	21,1	17,1	27,5

Перед тем, как приступить непосредственно к анализу, необходимо проверить, включён ли анализ данных (о том, как включить анализ данных в Excel, см. тему 2).

Шаг 1. Создание электронной таблицы с данными.

При создании таблицы с данными необходимо каждый из признаков (другими словами – отдельную переменную) разместить в отдельном столбце. То есть, в конечном итоге получим два столбца с данными (рисунок 3.1).

	A	B
1	X	Y
2	3,4	14,3
3	3,6	14,9
4	4,5	17,3
5	4,8	17,3
6	4,9	17,4
7	5,2	17,5
8	5,4	17,6
9	5,7	17,6
10	6,2	17,6
11	6,7	17,8
12	7,1	18
13	7,5	18
14	7,7	18,1
15	7,8	18,1
16	7,9	18,6
17	8	19,7
18	8,4	19,8

Рисунок 3.1 – Создание ряда данных в книге Excel

Шаг 2. Определение способа проведения анализа.

Провести корреляционный анализ в электронных таблицах Excel можно несколькими способами: путем введения непосредственной формулы и используя готовую опцию из блока **«Анализ данных»**. Рассмотрим каждую из них.

3.1.1 Проведение корреляционного анализа в Excel при помощи формулы

Шаг 3. Введение формулы.

Для проведения анализа необходимо выбрать свободную ячейку, в которой будет отражаться в дальнейшем искомый коэффициент

корреляции. Допустим, это будет ячейка с адресом $D$1. В ней необходимо установить курсор, а затем в строке формул либо вручную написать формулу с синтаксисом **=КОРРЕЛ(A2:A49;B2:B49)** (рисунок 3.2) и нажать клавишу **Enter**, либо выбрать в главном меню программы раздел **Формулы**, а затем перейти в опцию **Вставить функцию,** и после появления диалогового окна мастера функций в верхней части окна из списка представленных функций выбрать **Статистические**, а в нижней – **КОРРЕЛ** (рисунок 3.3). После чего необходимо нажать **ОК**.

Рисунок 3.2 – Ввод формулы корреляции в строке формул Excel

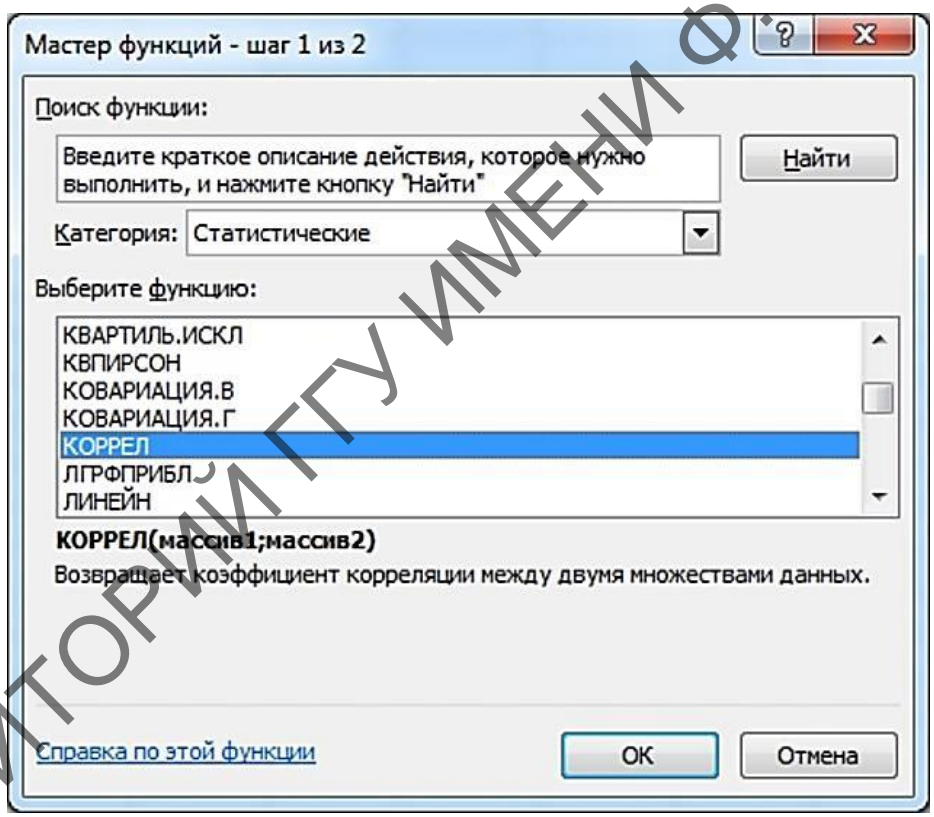


Рисунок 3.3 – Диалоговое окно **Мастер функций** в Excel

Шаг 4. Выставление параметров.

В появившемся диалоговом окне **Аргументы функции** программе необходимо указать диапазон значений первой (**Массив1**) и второй (**Массив2**) анализируемых переменных (рисунок 3.4) без заголовков (!) и нажать **ОК.**

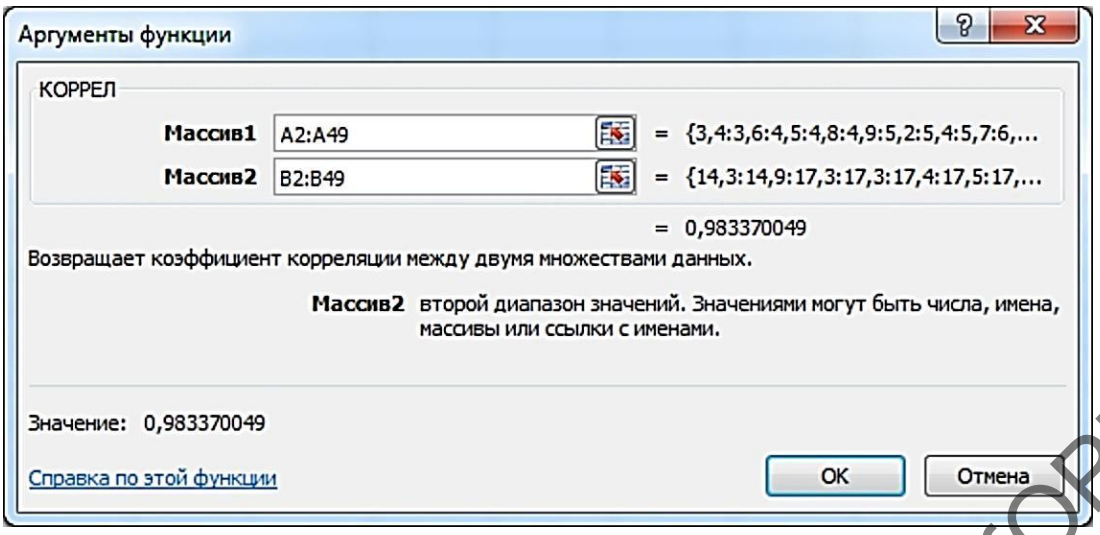


Рисунок 3.4 – Диалоговое окно **Аргументы функций** в Excel

В результате любого из выбранного действия в определённой нами ячейке отобразится значение коэффициента корреляции – 0,98337. Заметим, что это очень хороший результат.

3.1.2 Проведение корреляционного анализа в Excel при помощи предустановленного блока «Анализ данных»

Шаг 1. Выбор опции.

Для проведения корреляционного анализа этим способом необходимо перейти в пункт главного меню **Данные,** а затем открыть модуль **Анализ данных**, выбрать там из списка опцию **Корреляция** (рисунок 3.5) после чего щелкнуть мышкой **OK**.

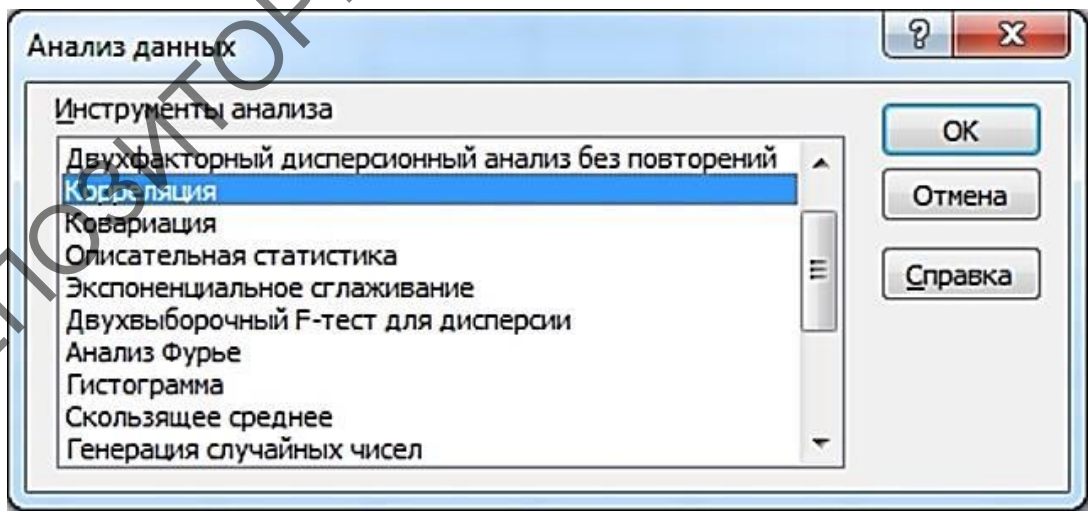


Рисунок 3.5 – Выбор опции **Корреляция**
в диалоговом окне **Анализ данных** в Excel

Шаг 2. Выставление параметров.

В появившемся окне необходимо выполнить операции и установки, как показано на рисунке 3.6 (параметры уже вам знакомы и мы на них не заостряем внимание), и щелкнуть мышкой **OK**.

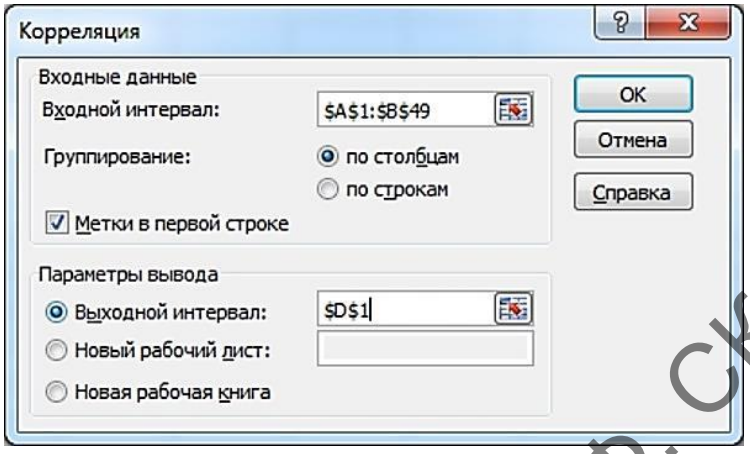


Рисунок 3.6 – Диалоговое окно **Корреляция**

В итоге будет отражена таблица с показателем коэффициента корреляции (таблица 3.2).

Таблица 3.2 – Значение коэффициента корреляции

	X	*Y*
X	1	
Y	0,98337	1

3.2 Параметрический корреляционный анализ Пирсона в STATISTICA

Коэффициент корреляции Пирсона (r) представляет собой меру линейной зависимости 2 переменных. Если возвести его в квадрат, то полученное значение коэффициента детерминации (r^2) представляет долю вариации, общую для 2 переменных (иными словами, степень зависимости или связанности двух переменных). Чтобы оценить зависимость между переменными, нужно знать как величину корреляции, так и ее значимость.

Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции. Значимость определенного коэффициента корреляции зависит от объема выборок.

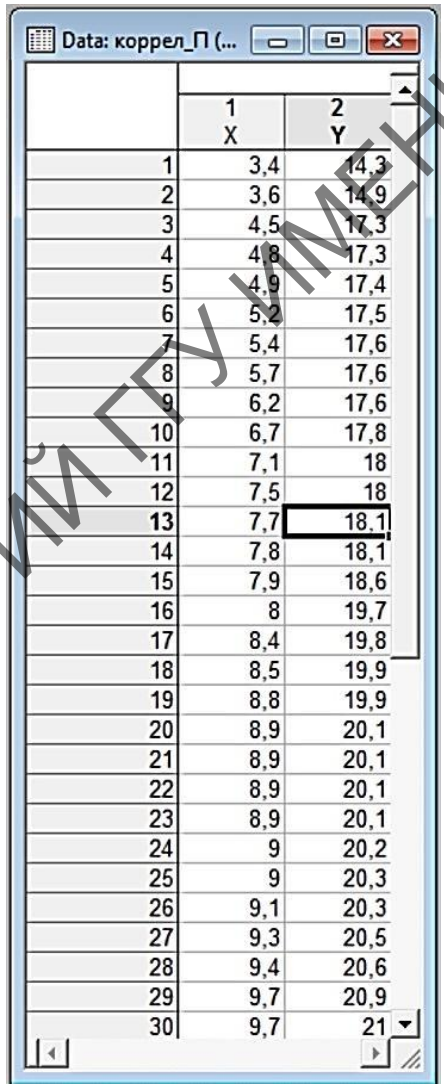
Таким образом, корреляционный анализ Пирсона для оценки связи между двумя признаками может быть проведен при наличии этих обязательных условий:

- значения обоих анализируемых признаков подчинены закону нормального распределения;
- связь между признаками линейна.

Рассмотрим расчёт коэффициента корреляции Пирсона на примере ранее рассмотренных данных в Excel (таблица 3.1).

Шаг 1. Создание электронной таблицы с данными.

Для проведения анализа необходимо предварительно составить таблицу с данными по признакам двух переменных – *X* и *Y*, представленных в виде двух столбцов. Начальная часть таблицы показана на рисунке 3.7.



	1 X	2 Y
1	3,4	14,3
2	3,6	14,9
3	4,5	17,3
4	4,8	17,3
5	4,9	17,4
6	5,2	17,5
7	5,4	17,6
8	5,7	17,6
9	6,2	17,6
10	6,7	17,8
11	7,1	18
12	7,5	18
13	7,7	18,1
14	7,8	18,1
15	7,9	18,6
16	8	19,7
17	8,4	19,8
18	8,5	19,9
19	8,8	19,9
20	8,9	20,1
21	8,9	20,1
22	8,9	20,1
23	8,9	20,1
24	9	20,2
25	9	20,3
26	9,1	20,3
27	9,3	20,5
28	9,4	20,6
29	9,7	20,9
30	9,7	21

Рисунок 3.7 – Электронная таблица данных
для расчёта коэффициента корреляции

Шаг 2. Выбор анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*) и в нём – модуль **Basic Statistics/Tables** (*Основные статистические показатели / Таблицы*), далее – опцию **Correlation Matrices** (*Корреляционные матрицы*) (рисунок 3.8) и нажать **ОК**.

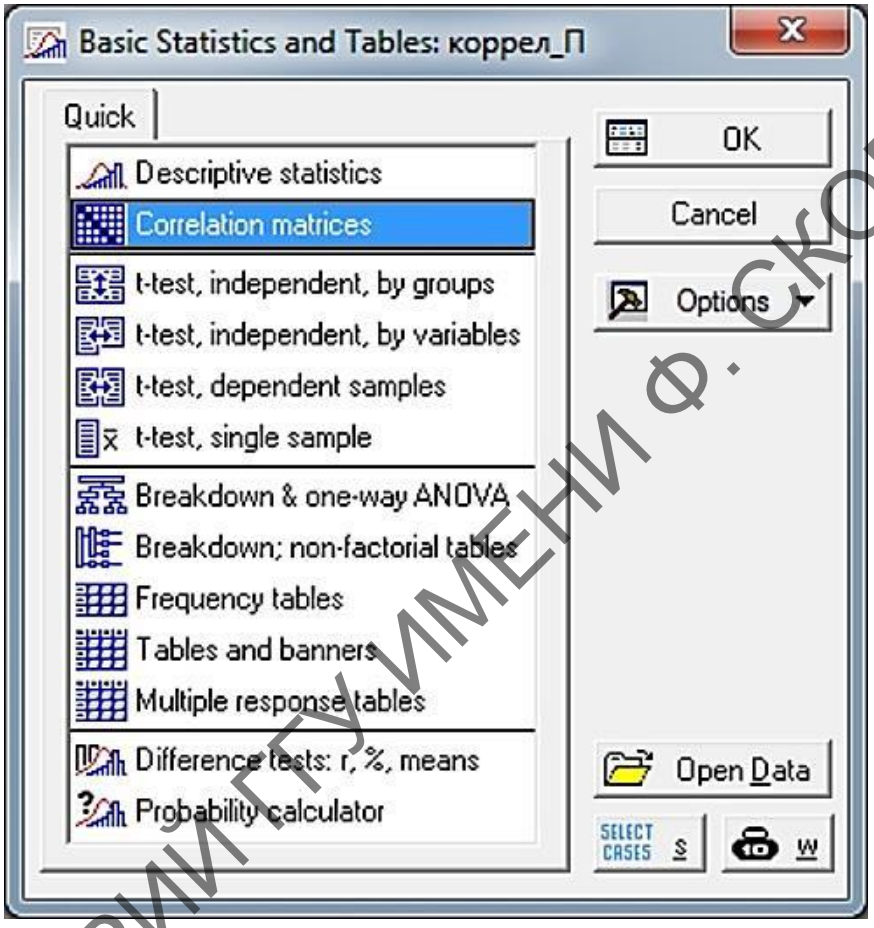


Рисунок 3.8 – Выбор модуля **Correlation Matrices**

Шаг 3. Указание переменных.

В появившемся диалоговом окне необходимо выбрать переменные, которые нужно включить в анализ. В случае, если анализируемые переменные последовательно выбираются из одного списка, нужно нажать кнопку **One variable list** (*Один список переменных*). Иначе, если анализируемые переменные выбираются из двух списков (как в нашем случае) – **Two lists (rect. matrix)** (*Два списка (прямоугольная матрица)*) (рисунок 3.9).

После этого в диалоговом окне слева необходимо указать список первой переменной (в нашем случае – *X*), а справа – список второй переменной (в нашем случае – *Y*) (рисунок 3.10).

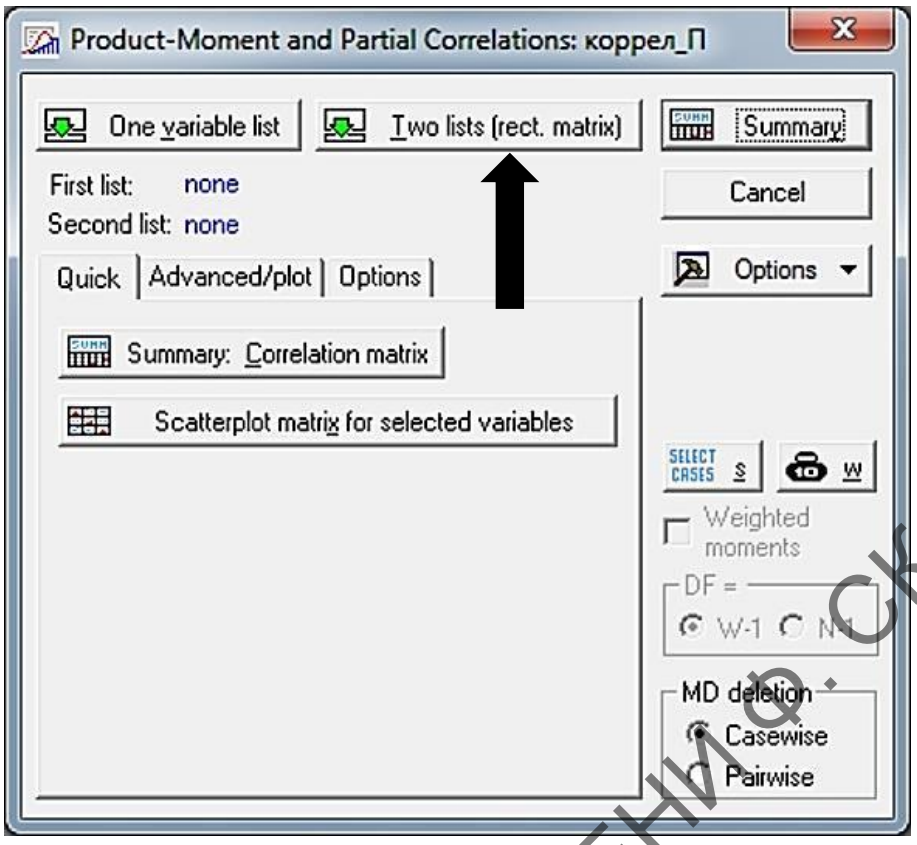


Рисунок 3.9 – Указание количества списков с переменными

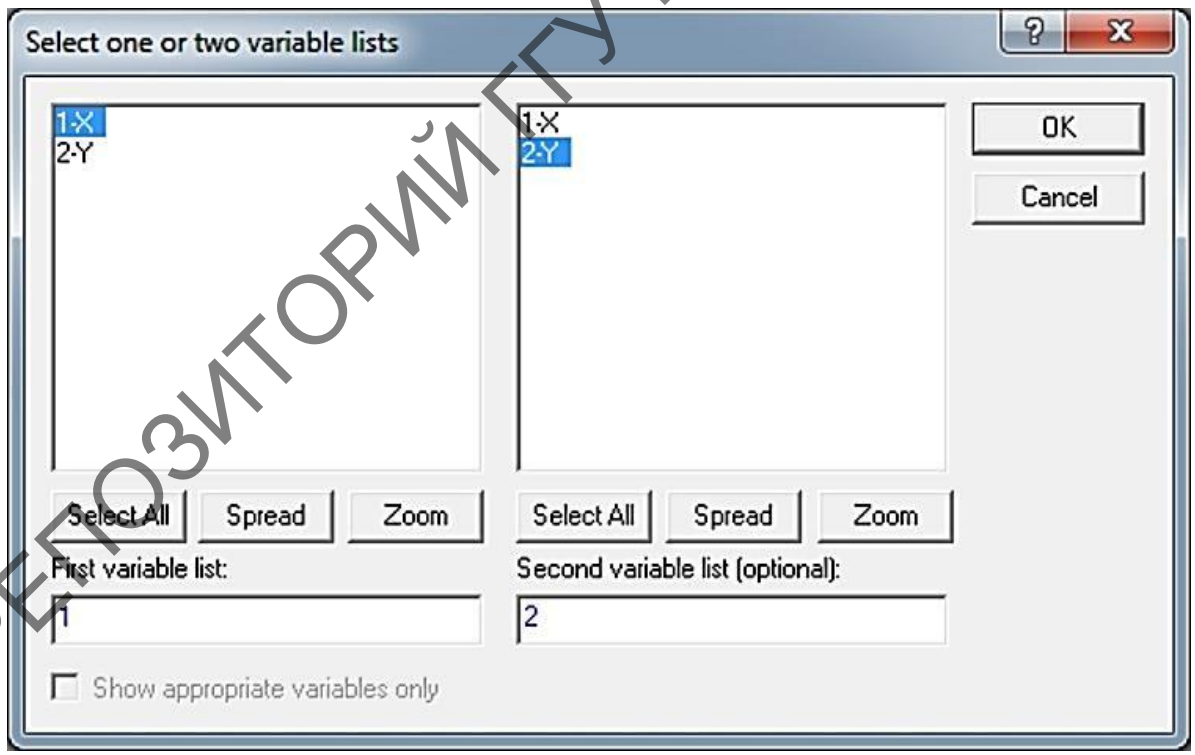


Рисунок 3.10 – Указание первого и второго списков с переменными

Шаг 4. Проверка условий для проведения анализа.

Перед проведением анализа целесообразно проверить, насколько наши данные удовлетворяют условиям для расчёта коэффициента корреляции Пирсона, т. е. насколько данные подчиняются закону нормального распределения и насколько зависимость линейна. Первое условие (нормальность распределения) для каждой из сравниваемых выборок можно проверить отдельно любым из способов, рассматриваемых ранее (тема 2), либо визуально непосредственно в рассматриваемом блоке анализа. Для этого необходимо, находясь на закладке **Quick** (*Быстрый*), выбрать кнопку **Scatterplot matrix for selected variables** (Диаграмма рассеяния для выбранных переменных) (рисунок 3.11).

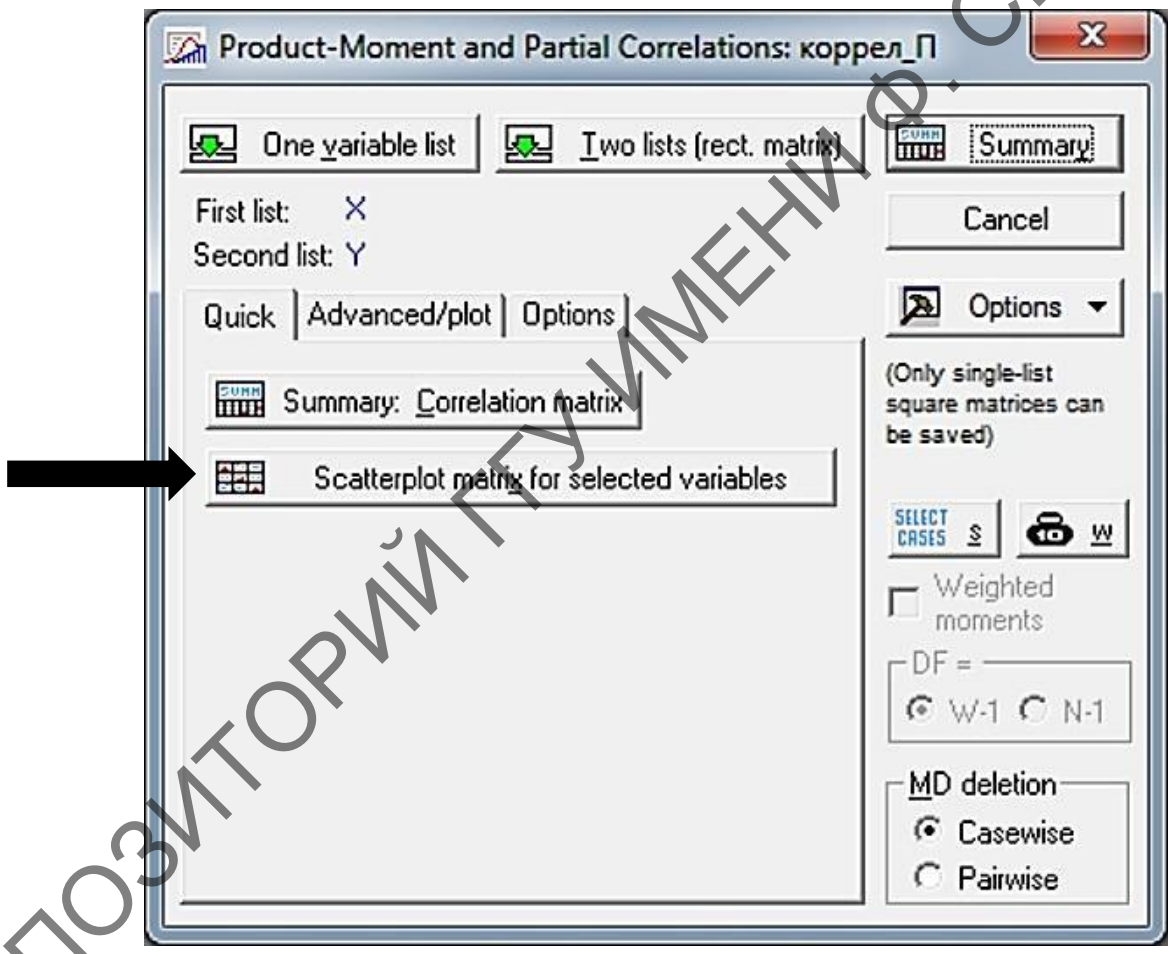


Рисунок 3.11 – Выбор опции **Scatterplot matrix for selected variables**

После чего нужно указать переменные для X и Y и нажать **ОК.** В появившемся окне программа отобразит точечный график, по осям которого отложены значения соответствующих переменных (рисунок 3.12).

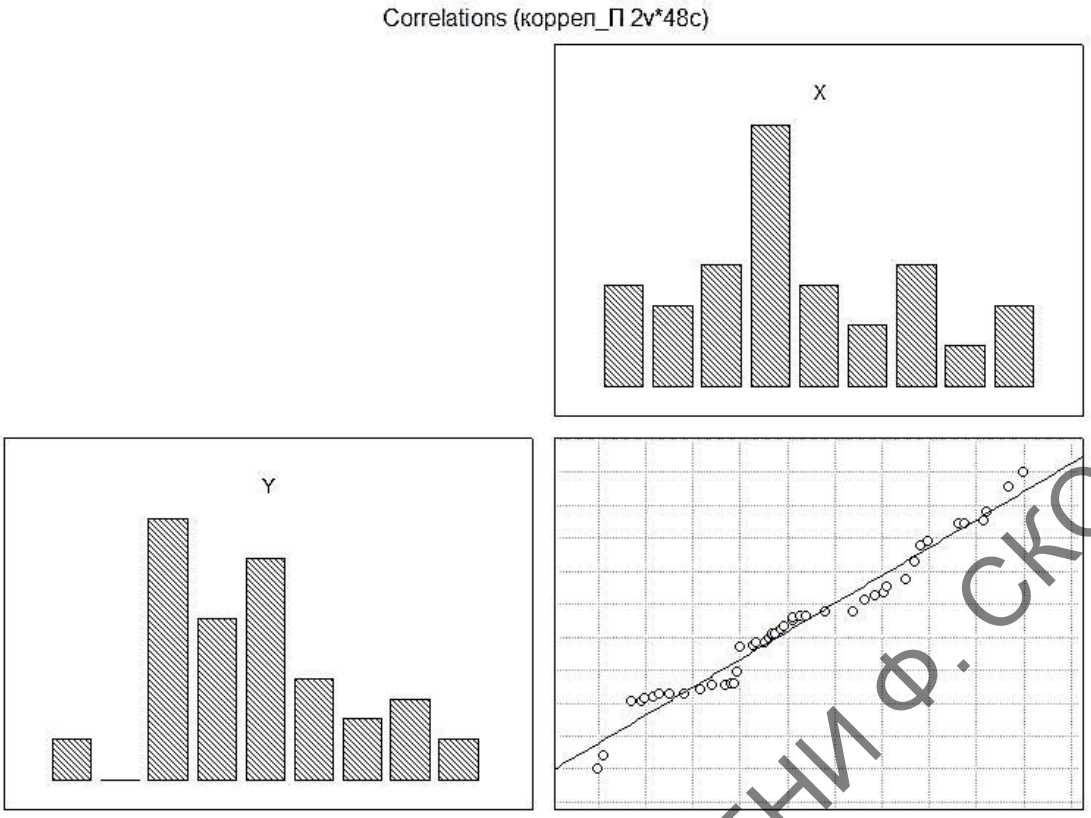


Рисунок 3.12 – График и гистограммы проверки условий расчёта коэффициента корреляции Пирсона

Сплошная диагональная линия на этом графике служит для оценки степени линейности связи между анализируемыми признаками. Если точки наблюдения укладываются вдоль этой линии на близком расстоянии (как в нашем случае), можно говорить о существовании прямолинейной зависимости. Наряду с рассмотренной нами диаграммой рассеяния программа также строит гистограммы распределения значений анализируемых признаков для проверки нормальности распределения (в нашем случае и первая, и вторая переменные распределены нормально).

В связи с тем, что наши данные удовлетворяют всем неоходимым условиям для проведения анализа, закрываем окно проверки и возвращаемся в диалоговое окно параметров анализа, нажав кнопку с называнием анализа в левом нижнем углу главного экрана программы (рисунок 3.13).

Рисунок 3.13 – Кнопка свёрнутого диалогового окна анализа Пирсона

Шаг 5. Проведение анализа.

Перед проведением анализа необходимо в диалоговом окне перейти на закладку **Options** (*Настройки*) и выполнить установки, показанные на рисунке 3.14, подсветив **Displey detaled table of results** (*Отобразить подробную таблицу с результатами*), и нажать кнопку **Summary** (*Результат*).

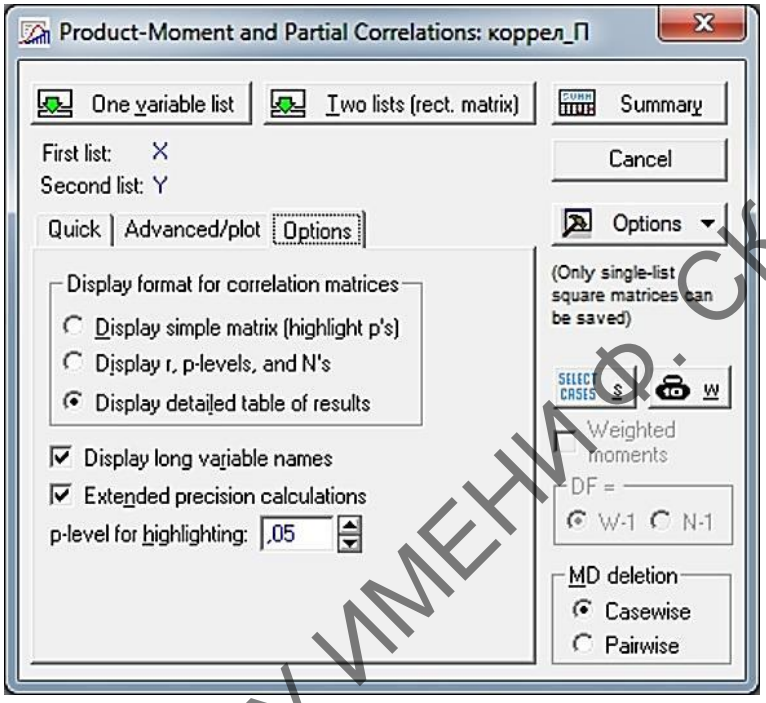


Рисунок 3.14 – Окончательная настройка опций корреляционного анализа

Шаг 6. Интерпретация результатов.

После нажатия кнопки **Summary** (*Результат*) программа произведет расчеты корреляции между *X* и *Y*, и через секунду на экране появится окно результатов (рисунок 3.15).

Correlations (коррел_П)
Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r?	t	p	N	Constant dep: Y	Slope dep: Y	Constant dep: X	Slope dep: X
X	9,68958	3,556549									
Y	20,56667	2,952484	0,983370	0,967017	36,72388	0,00	48	12,65659	0,816349	-14,6729	1,184563

Рисунок 3.15 – Результат расчёта коэффициента корреляции Пирсона

Полученная таблица отображает следующие параметры:

− **Mean** – среднее арифметическое каждой выборки;

− **Std.Dv.** – стандартное отклонение;

– **r(X,Y)** – значение коэффициента корреляции *r*;
– **r?** – значение коэффициента детерминации r^2;
– **t** – t–критерий Стьюдента;
– **p** – уровень значимости;
– **N** – число коррелируемых пар;
– **Constant dep:Y** – свободный член Y;
– **Slope dep: Y** – коэффициент при независимой переменной;
– **Constant dep:X** – свободный член X;
– **Slope dep:** X – коэффициент при зависимой переменной.

В нашем примере *r* = 0,98… при практически 100 % достоверности. Это очень хорошее значение (подсвечено красным цветом), показывающее, что построенная регрессия объясняет более 90 % разброса значений переменной X относительно среднего.

Шаг 7. Графическое отображение результатов анализа.

Для визуализации результатов анализа бывает полезным графическое отображение результатов корреляционного анализа Пирсона для включения в научную публикацию и наглядного подтверждения выводов. Поэтому необходимо вернуться в диалоговое окно анализа и в закладке **Advanced/plot** (*Расширенные настройки/графики*) нажать кнопку **2D scatterplots** (*Двухмерные графики*) (рисунок 3.16).

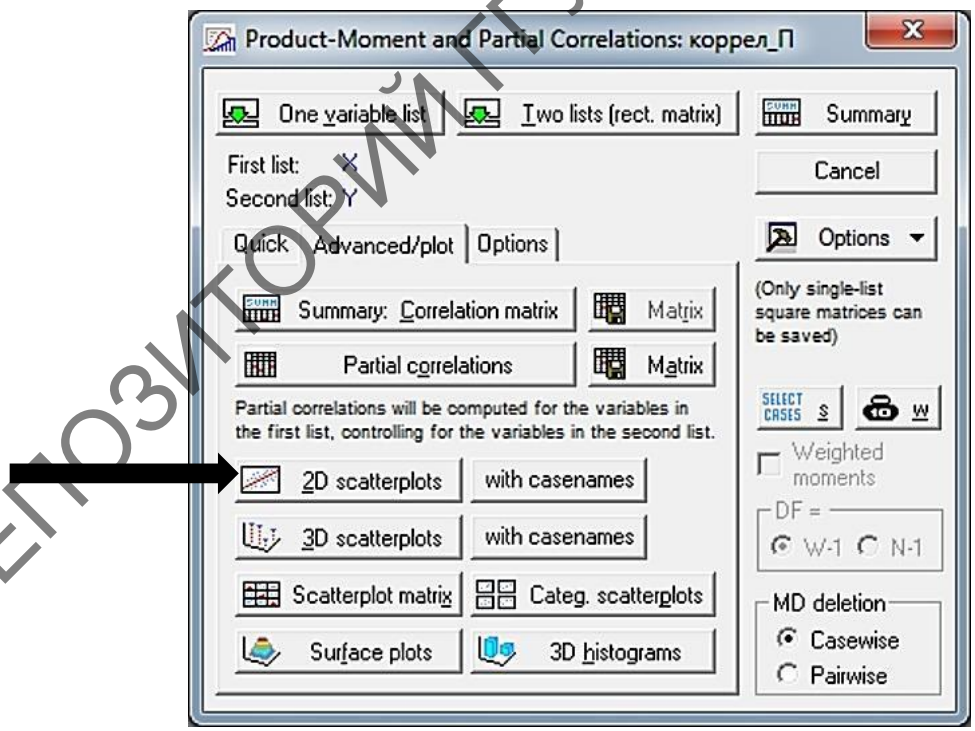


Рисунок 3.16 – Закладка **Advanced/plot**

В результате будет отражен график, на котором данные с подогнанной прямой имеют вид, представленный на рисунке 3.17.

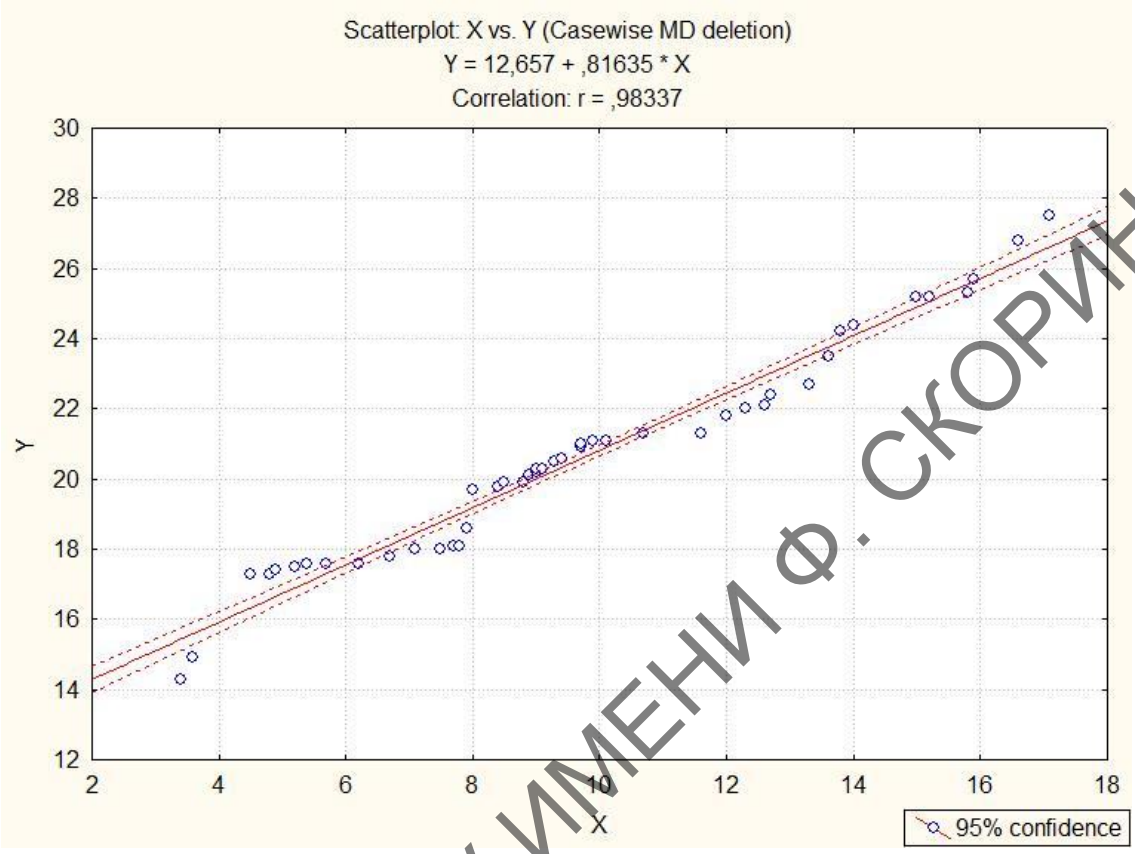


Рисунок 3.17 – Линейная регрессия для данных X и Y

3.3 Непараметрический корреляционный анализ Спирмена в STATISTICA

При расчете коэффициента корреляции Пирсона данных о весе (в г) левой камеры сердца (переменная *x*) и длине ядер (в μ) в мышцах сердца (переменная *y*) (таблица 3.3) исследователь столкнулся с наличием ненормального распределения у переменной *x* и сильным разбросом данных около центральной прямой (рисунок 3.18), что говорит о низкой линейной зависимости между признаками двух переменных.

Таблица 3.3 – Данные о весе (в г) левой камеры сердца (*x*) и длине ядер (в μ) в мышцах сердца (*y*)

X	207	221	256	262	273	289	291	292	304	328	372	397	460	632
Y	16,6	18	15,9	20,7	19,4	19,8	11,7	21	23	13,6	19,6	22,9	19,4	28,4

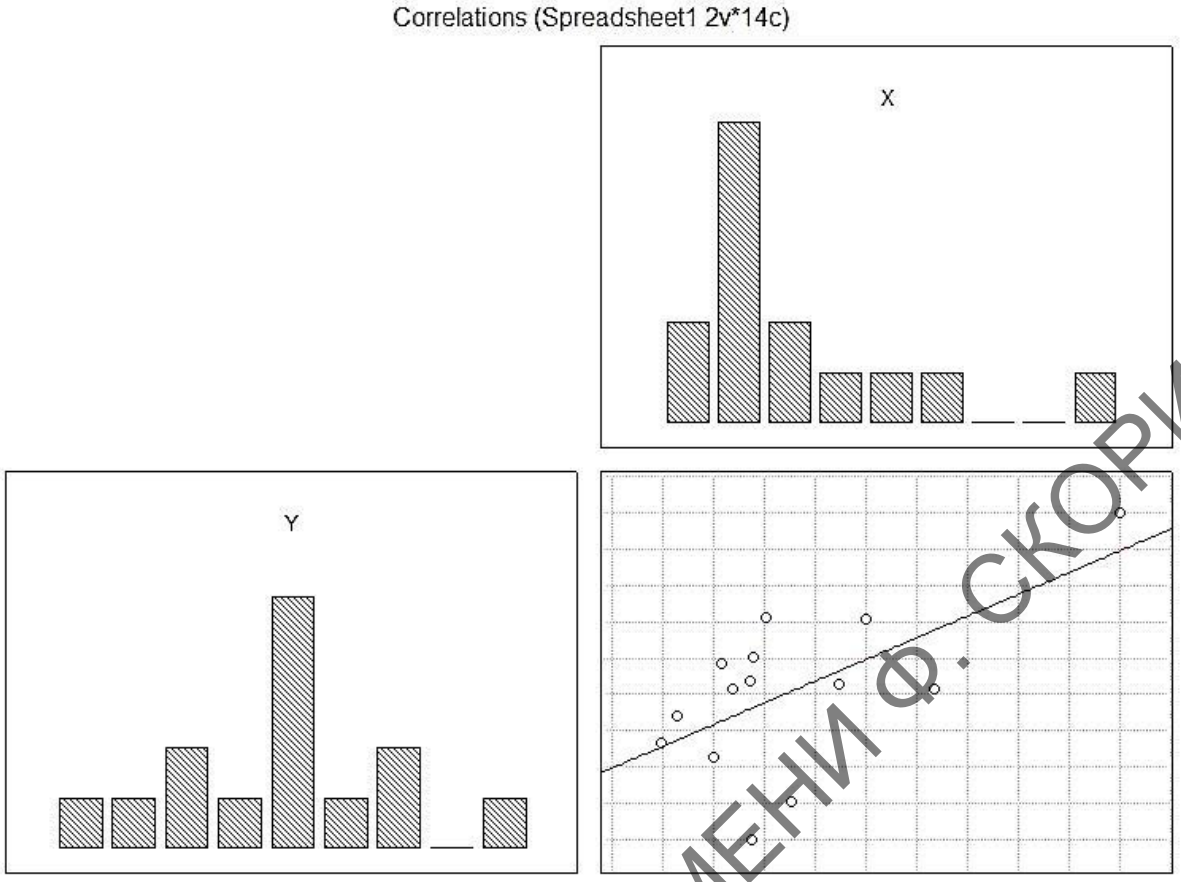


Рисунок 3.18 – График и гистограммы проверки условий расчёта коэффициента корреляции Пирсона в примере с параметрами сердца

Использование коэффициента корреляции Пирсона в данной ситуации приведёт к неверным выводам. Поэтому вместо него необходимо воспользоваться одним наиболее обычным непараметрическим коэффициентом корреляции – ранговым коэффициентом корреляции Спирмена.

Шаг 1. Создание электронной таблицы с данными.

Для проведения анализа необходимо предварительно составить таблицу с данными по признакам двух переменных – *X* и *Y*, представленных в виде двух столбцов (рисунок 3.19).

Шаг 2. Выбор анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **Nonparametrics** (*Непараметрические методы*) (рисунок 3.20) и далее **Correlations (Spearman, Kendall tau, gamma)** (*Корреляции* (*Спирмена, тау Кендалла, гамма*)) (рисунок 3.21).

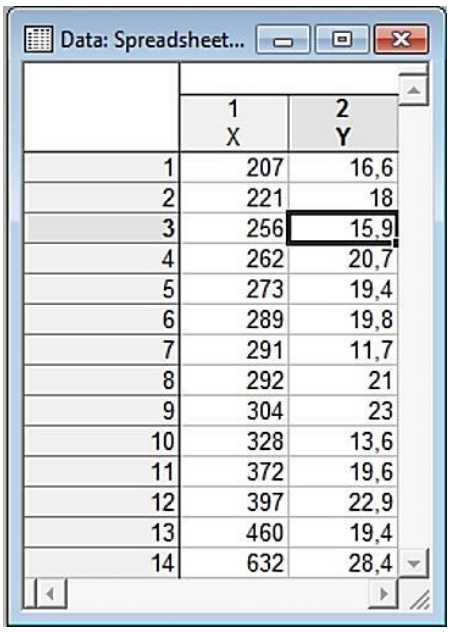


Рисунок 3.19 – Электронная таблица с данными

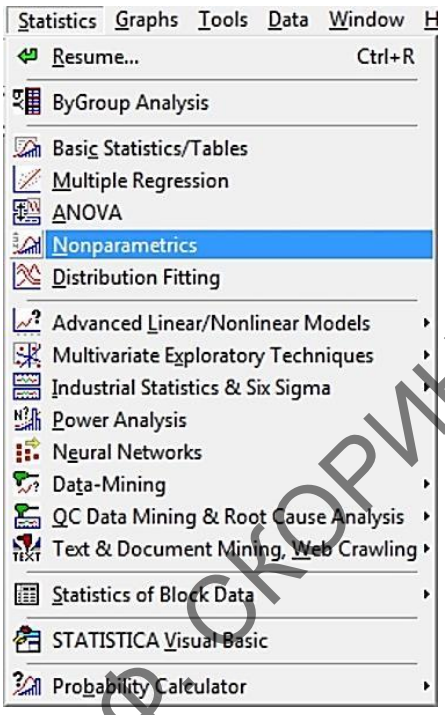


Рисунок 3.20 – Выбор опции **Nonparametrics**

Шаг 3. Выставление параметров.

В появившемся окне (рисунок 3.22) необходимо перейти на закладку **Advanced** (*Расширенные настройки*) и в разделе **Compute** (*Расчёт*) выбрать опцию **Detailed report** (*Подробный отчёт*) (рисунок 3.23), после чего следует нажать на кнопку **Variables** (*Переменные*) и выбрать столбцы, содержащие необходимые данные (рисунок 3.24). Затем нажать кнопку **ОК**.

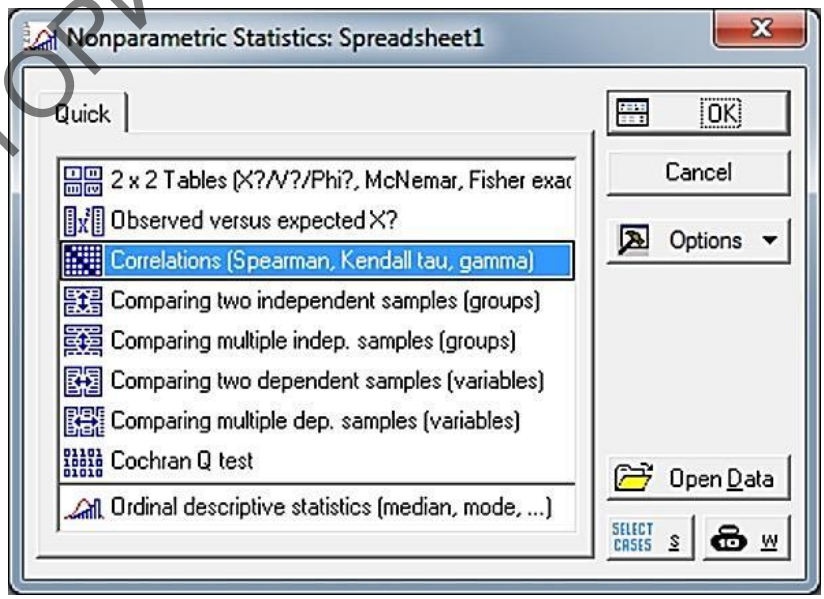


Рисунок 3.21 – Выбор опции **Correlations (Spearman, Kendall tau, gamma)**

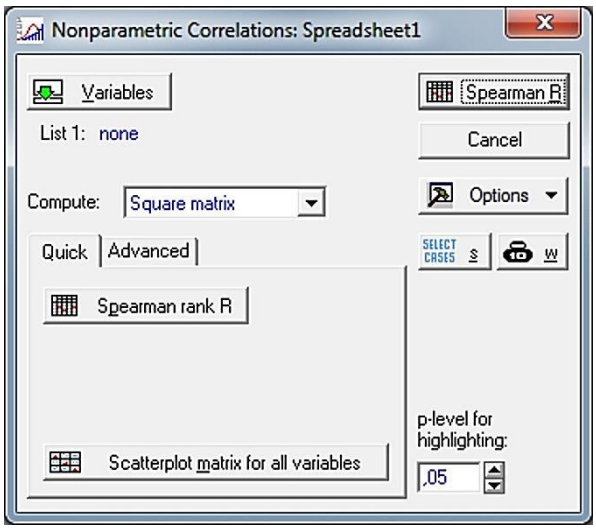


Рисунок 3.22 – Диалоговое окно **Nonparametric Correlations**

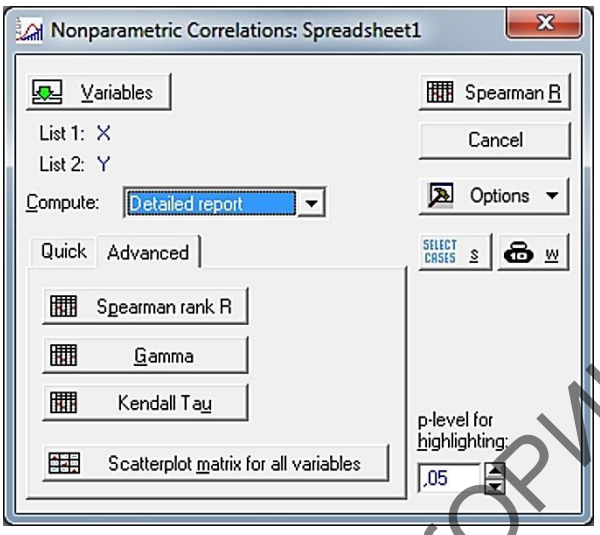


Рисунок 3.23 – Выставление параметров

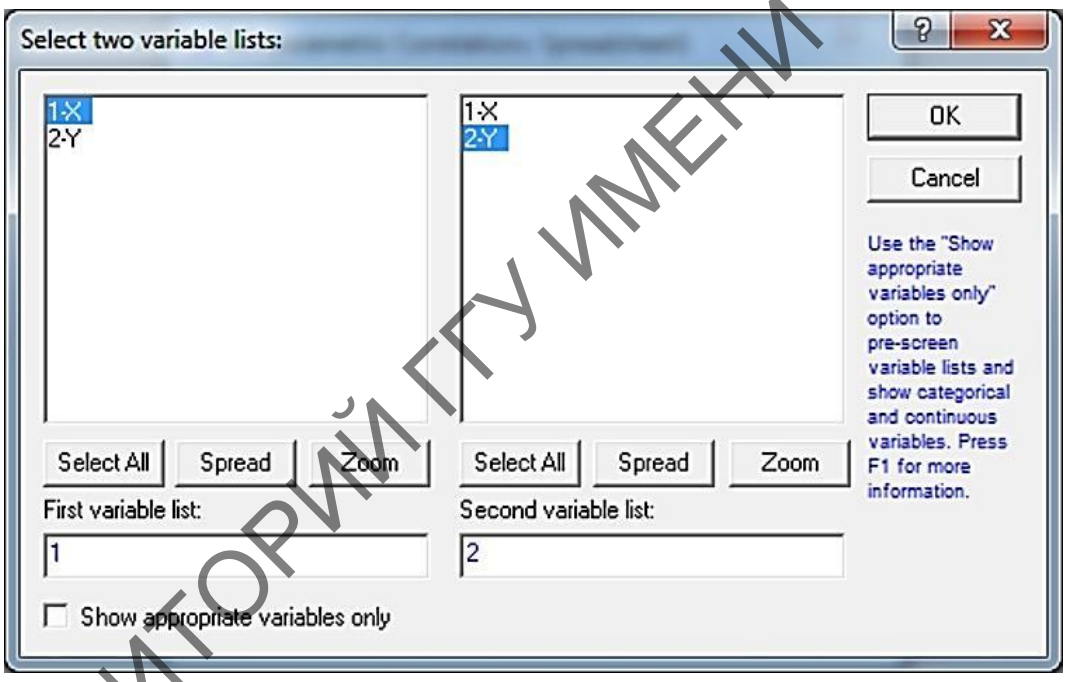


Рисунок 3.24 – Выбор переменных

Шаг 4. Проведение анализа.

Для окончания процедур расчёта в диалоговом окне **Nonparametric Correlations** (*Непараметрические корреляции*) после указания переменных нужно нажать кнопку **Spearman R** (*Коэффициент корреляции Спирмена*) или **Spearman rank R** (*Ранговый коэффициент корреляции Спирмена*). Появится таблица с результатами анализа (рисунок 3.25).

	Spearman Rank Order Correlations (Spreadsheet1) MD pairwise deleted Marked correlations are significant at p <,05000			
Pair of Variables	Valid N	Spearman R	t(N-2)	p-level
X & Y	14	0,468647	1,837750	0,090972

Рисунок 3.25 – Итоговая таблица корреляционного анализа Спирмена

Таблица содержит следующие параметры, отражённые в столбцах:

– **Valid N** – число наблюдений;

– **Spearman R** – коэффициент корреляции Спирмена;

– **t(N-2)** – значение критерия Стьюдента для числа степеней свободы n-2;

– **p-level** – уровень значимости (вероятность ошибки для нулевой гипотезы об отсутствии связи между признаками).

В нашем примере коэффициент корреляции Спирмена оказался равен 0,468647. При этом он не является статистически значимым ($p > 0,05$).

Задания

1) Надо установить, есть ли корреляция между высотой головы (x) и длиной 3-го членика усика (y) *Drosophila funebris*:

x	15	16	15	15	16	16	17	18	18	17	17	17	15	16	15	15	15	17
y	29	31	32	33	32	33	33	36	36	35	35	35	35	33	31	31	31	35
x	15	13	15	14	17	15	16	15	15	16	15	16	15	16	18	17	14	15
y	33	30	32	31	35	33	33	32	30	33	33	33	30	31	34	34	31	33
x	14	15	15	13	15	16	14	15	15	15	14	15	15	15	16	18	15	14
y	31	31	33	30	30	33	30	33	31	32	30	31	31	32	33	35	32	32

2) Получены следующие данные о продолжительности беременности у кроликов при различных размерах помета (число крольчат в помете (x) и длительность беременности в днях (y)):

x	1	8	3	5	7	8	4	8	3	4	4	8	8	5	7	6	6	5	6	6
y	33	30	31	31	31	32	31	31	32	33	32	31	31	31	31	30	31	32	32	31
x	6	5	7	8	10	6	7	6	7	6	5	10	7	8	8	6	5	6	5	4
y	32	32	31	32	31	31	30	31	31	32	31	30	32	32	31	31	31	32	30	31
x	6	8	6	5	8	7	6	5	9	5	3	4	7	8	9	5	6	2	2	4
y	31	31	32	32	31	30	32	31	31	31	32	32	31	31	31	31	31	32	33	33

Есть ли корреляция между длительностью плодоношения и размерами помета? Определите необходимый коэффициент корреляции и постройте линию регрессии. Сделайте вывод.

3) Были получены следующие данные о весе ягнят-баранчиков (одинцов) – y и весе баранов – их отцов – x (в кг).

x	76,6	72,2	67,0	66,5	63,3	65,4	63,9	63,1	63,0	62,5	62,2
y	4,56	4,79	4,49	4,32	4,59	4,32	4,67	4,29	4,57	4,20	4,12
x	61,0	60,2	60,0	59,6	59,5	58,9	58,0	57,8	57,6	57,0	
y	4,13	4,70	3,80	4,23	3,76	4,08	4,61	4,37	4,30	4,0	
x	56,8	55,4	55,0	53,8	53,7	52,0	51,4	51,0	50,9	48,5	
y	3,82	4,12	4,19	4,16	4,09	4,12	4,02	4,31	4,06	4,03	

Есть ли корреляция между весом баранчиков и весом их отцов? Определите необходимый коэффициент корреляции и постройте линию регрессии. Сделайте вывод.

4) У 40 серебристо-черных лисиц были измерены (в см) длина туловища x и длина хвоста y:

x	70	65	66	65	71	68	64	57	66	65	67	62	67	62	63	57	64	66	69	58
y	40	40	40	40	40	42	39	38	41	43	39	45	43	38	40	40	41	45	43	37
x	63	67	67	67	65	65	67	70	65	71	69	64	64	66	69	72	66	66	67	66
y	45	38	39	37	42	38	38	38	38	40	39	43	43	42	40	41	47	47	40	40

Определите необходимый коэффициент корреляции между длиной туловища и длиной хвоста серебристо-белых лисиц. Постройте линию регрессии, сделайте вывод.

5) Были получены следующие данные о весе x (в кг) и длине туловища y (в см) 100 серебристо-черных лисиц:

x	4,7	4,6	5,2	5,1	5,3	5,3	4,6	4,8	5,8	5,7	4,5	5,7	5,0	4,8	4,7	5,2	4,6
y	70	65	69	70	66	68	65	71	69	68	57	73	65	67	71	62	69
x	5,5	5,5	4,8	4,7	6,0	5,1	5,	4,5	5,0	5,0	4,9	5,5	5,2	5,6	5,2	5,7	5,3
y	62	63	67	64	64	66	68	69	58	63	67	74	67	67	70	65	71
x	5,4	5,3	4,6	5,6	5,1	4,9	5,2	5,3	5,0	5,3	5,6	5,0	5,1	5,5	5,6	5,2	5,0
y	63	64	64	66	63	69	62	72	66	66	67	67	66	63	67	62	71
x	5,5	5,6	5,0	6,7	4,7	5,3	5,0	5,1	5,0	5,1	4,8	5,0	6,0	5,5	4,6	4,5	4,5
y	67	66	66	69	64	69	70	62	68	68	72	68	67	66	69	65	65
x	5,4	5,0	4,9	5,0	5,7	5,9	5,6	5,1	5,1	4,6	4,9	6,2	5,6	5,2	5,1	4,5	
y	65	65	64	66	66	67	62	63	64	69	69	68	65	69	67	68	
x	4,8	5,5	6,0	5,3	4,8	5,3	5,1	5,4	4,7	5,0	5,9	5,0	5,2	5,6	5,2	5,1	
y	61	64	62	66	59	65	62	68	61	67	69	69	66	66	67	70	

Определите, есть ли корреляция между весом и длиной туловища у лисиц? Определите необходимый коэффициент корреляции и постройте линию регрессии. Сделайте обоснованный вывод.

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М. Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

4 Мастицкий, С. Э. Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований / С. Э. Мастицкий. – Минск : РУП «Институт рыбного хозяйства», 2009. – 76 с.

5 Рокицкий, П. Ф. Биологическая статистика / П. Ф. Рокицкий. – Минск : «Вышэйшая школа», 1973. – 320 с.

ТЕМА 4. РЕГРЕССИОННЫЙ АНАЛИЗ В EXCEL И STATISTICA 7.0

4.1 Проведение линейного регрессионного анализа в Excel.
4.2 Линейный регрессионный анализ в STATISTICA.

4.1 Проведение линейного регрессионного анализа в Excel

Несмотря на то, что расчёт корреляции характеризует направление и силу связи между двумя переменными, а коэффициент корреляции позволяет количественно охарактеризовать степень связи, с его помощью невозможно предсказать, чему в среднем будет равно значение одного признака при изменении другого на определённое значение.

Регрессионный анализ служит как раз для определения вида этой связи и даёт возможность прогнозировать значения одной (зависимой) переменной, отталкиваясь от значения другой (независимой) переменной.

Часто связь между двумя биологическими особенностями, выраженными какими-либо признаками, имеет линейный характер, и её можно выразить в виде линейного уравнения

$$y = a + bx,$$

где y и x – анализируемые признаки (y – значение функции, зависимая переменная; x – аргумент, независимая переменная);

a – свободный член уравнения (при $b = 0$ получаем $y = a$, т. е., другими словами, a – это точка, в которой линия регрессии пересекается с осью ординат; она называется **«Intercept»**);

b – коэффициент регрессии (отражает угол наклона линии регрессии – чем больше b отличается от 0, тем сильнее связь между анализируемыми признаками).

Данное уравнение можно использовать для описания связи между 2 признаками только при выполнении обязательных условий:

1) зависимость между признаками носит линейный характер;

2) оба признака распределены нормально.

Для примера проведения регрессионного анализа в электронных таблицах Excel используем данные 2 признаков, рассмотренные при изучении корреляционного анализа (см. таблицу 3.1 темы 3).

Перед тем, как приступить непосредственно к анализу, необходимо проверить, включён ли анализ данных (о том, как включить анализ данных в Excel – см. тему 2).

Шаг 1. Создание электронной таблицы с данными.

При создании таблицы с данными необходимо каждый из признаков (другими словами – отдельную переменную) разместить в отдельном столбце. То есть, в конечном итоге получим 2 столбца с данными (рисунок 4.1).

	A	B
1	X	Y
2	3,4	14,3
3	3,6	14,9
4	4,5	17,3
5	4,8	17,3
6	4,9	17,4
7	5,2	17,5
8	5,4	17,6
9	5,7	17,6
10	6,2	17,6
11	6,7	17,8
12	7,1	18
13	7,5	18
14	7,7	18,1
15	7,8	18,1
16	7,9	18,6
17	8	19,7
18	8,4	19,8

Рисунок 4.1 – Создание ряда данных в книге Excel

Шаг 2. Выбор анализа.

Для проведения регрессионного анализа необходимо перейти в пункт главного меню **Данные,** а затем открыть модуль **Анализ данных**, выбрать там из списка опцию **Регрессия** (рисунок 4.2) после чего щелкнуть мышкой **OK**.

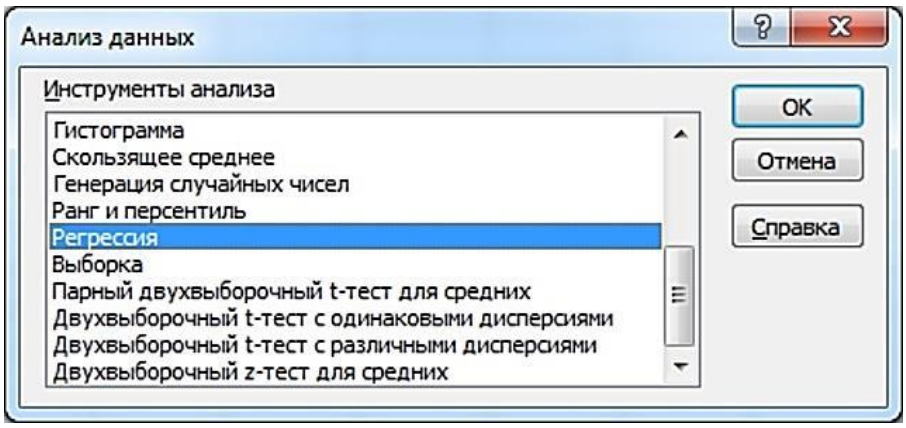


Рисунок 4.2 – Выбор опции **Регрессия**
в диалоговом окне **Анализ данных** в Excel

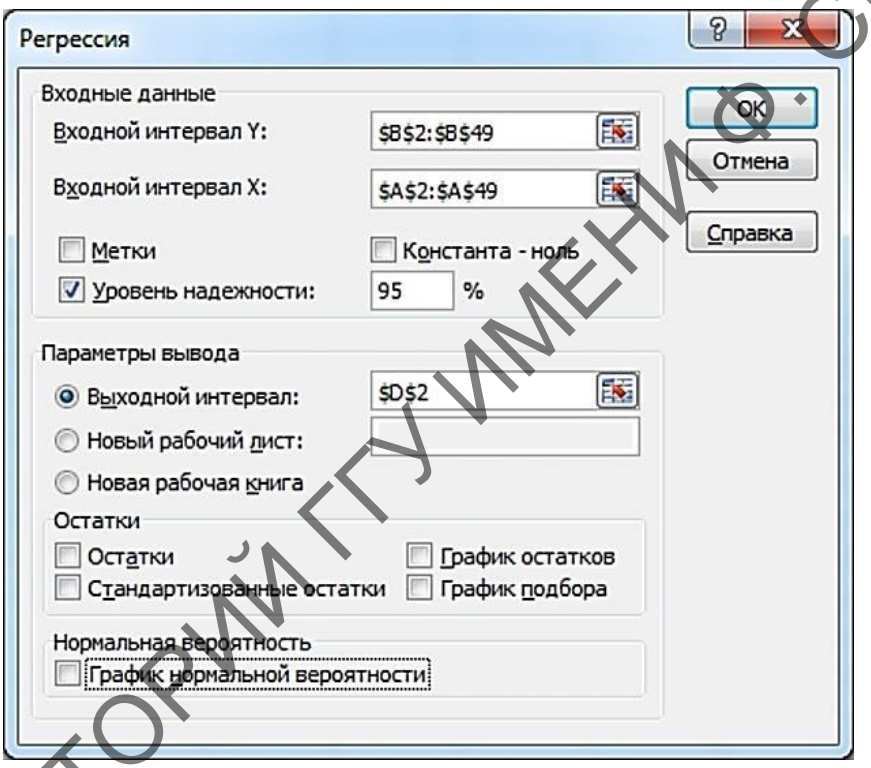


Рисунок 4.3 – Выставление параметров
в диалоговом окне **Регрессия** в Excel

Шаг 3. Выставление параметров.

В появившемся диалоговом окне регрессионного анализа необходимо указать диапазон данных зависимой переменной (y), диапазон данных независимой переменной (x), указать наличие заголовка в случае его использования в диапазоне данных и выставить уровень надёжности с выводом данных в текущем листе *Excel* (рисунок 4.3), затем щелкнуть мышкой **OK**.

Результат обработки появится в указанном поле (рисунок 4.4).

D	E	F	G	H	I	J	K	L
ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,983370049							
R-квадрат	0,967016654							
Нормированный R-квадрат	0,966299624							
Стандартная ошибка	0,542007076							
Наблюдения	48							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	1	396,1931698	396,1931698	1348,643212	9,90728E-36			
Остаток	46	13,51349686	0,293771671					
Итого	47	409,7066667						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	12,65658596	0,229160549	55,23021313	1,05952E-43	12,1953097	13,11786222	12,1953097	13,11786222
Переменная X 1	0,816348901	0,022229376	36,72387796	9,90728E-36	0,771603487	0,861094315	0,771603487	0,861094315

Рисунок 4.4 – Итоги регрессионного анализа в Excel

Для нашего массива данных получена очень надежная регрессия с высоким коэффициентом корреляции: Y = 12,6+0,81*X; r = 0,983370049.

Шаг 4. Графическое отображение результатов.

Для наглядного отображения как регрессии, так и корреляции в *Excel* можно воспользоваться мастером диаграмм. Для этого выделяем наш диапазон данных без заголовка (!), в главном меню Excel выбираем опцию **Вставка**, а затем заходим в блок меню **Диаграммы**, после чего выбираем тип диаграммы **Точечная** (рисунок 4.5).

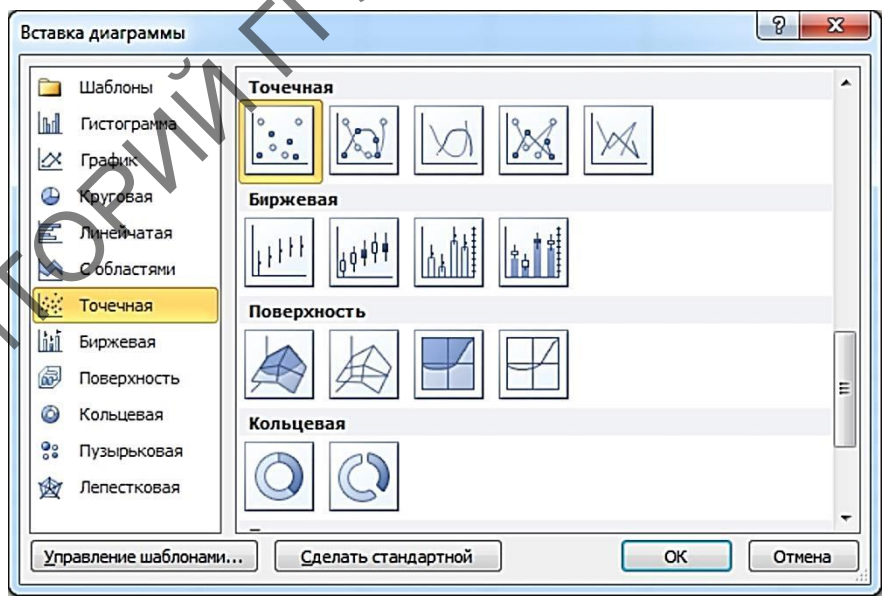


Рисунок 4.5 – Вставка диаграммы в Excel

После выбора диаграммы необходимо щелкнуть мышкой **OK**. В результате получим график, показанный на рисунке 4.6.

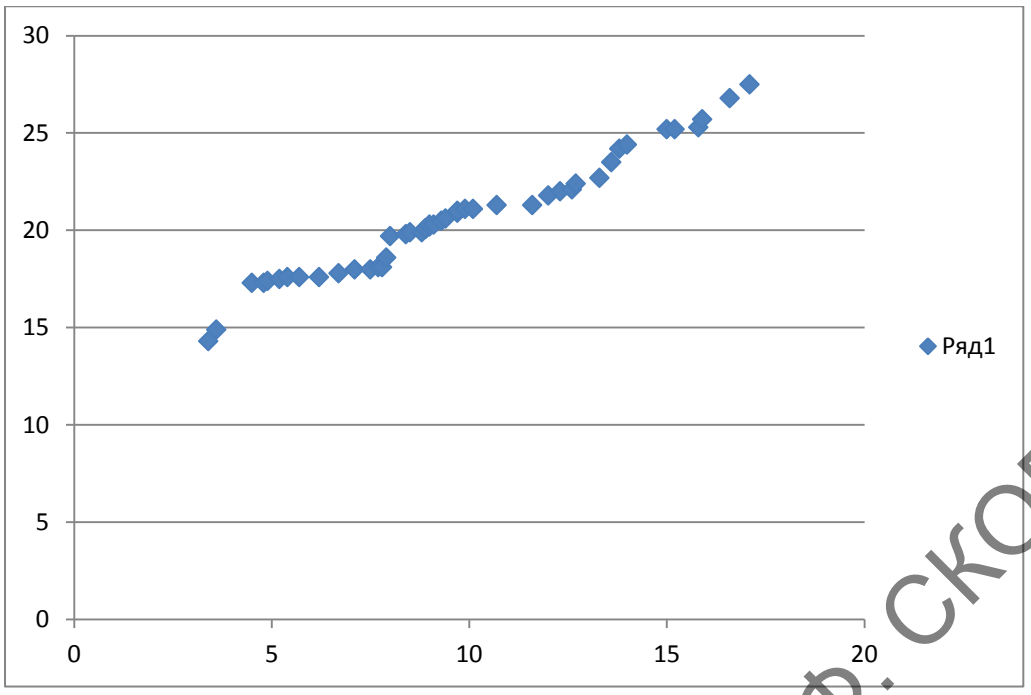


Рисунок 4.6 – Начальный график результатов регрессионного анализа в Excel

Далее необходимо щёлкнуть правой кнопкой по любой из точек и в контекстном меню выбрать опцию **Добавить линию тренда** (рисунок 4.7).

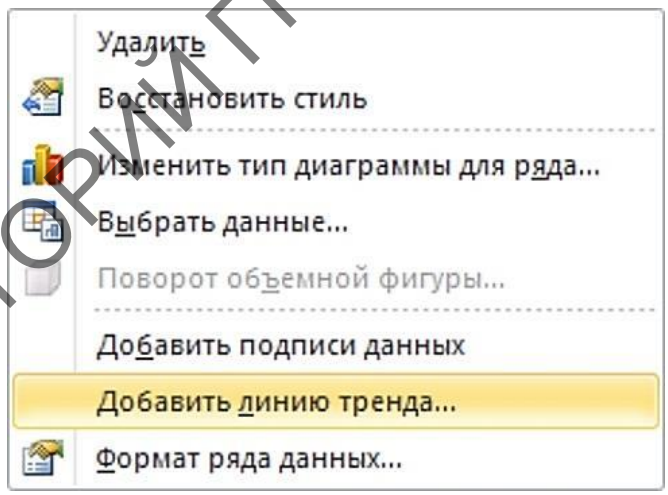


Рисунок 4.7 – Контекстное меню

После этого в диалоговом окне **Формат линии тренда**, в правой его части нужно выбрать опцию – **Линейная**, а остальные установки – как показано на рисунке 4.8, и нажать кнопку **Закрыть**.

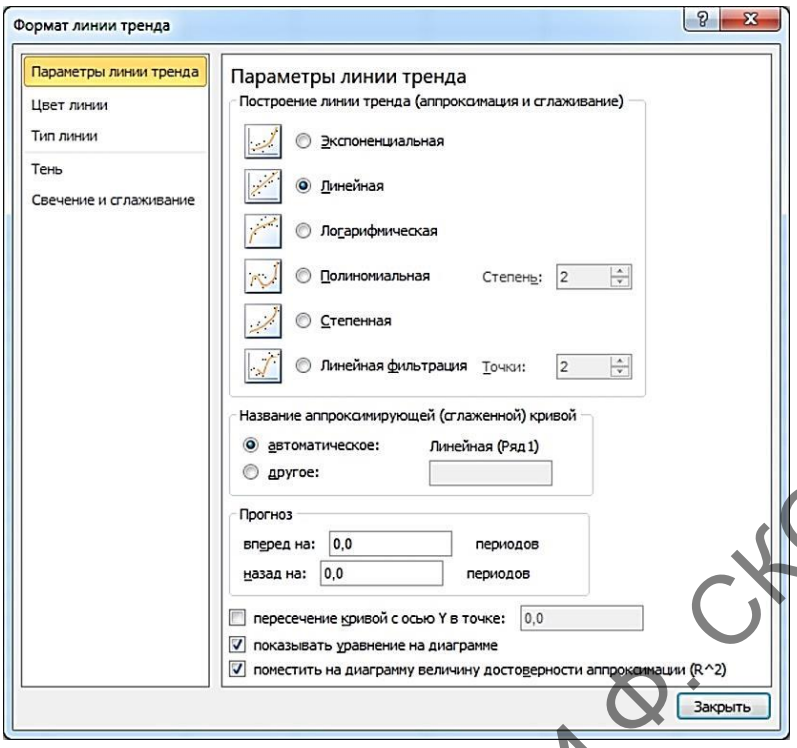


Рисунок 4.8 – Настройки линии тренда в Excel

После появления линии тренда необходимо щёлкнуть правой клавишей мыши по полю с легендой и удалить её. Окончательный результат отобразится на рисунке 4.9.

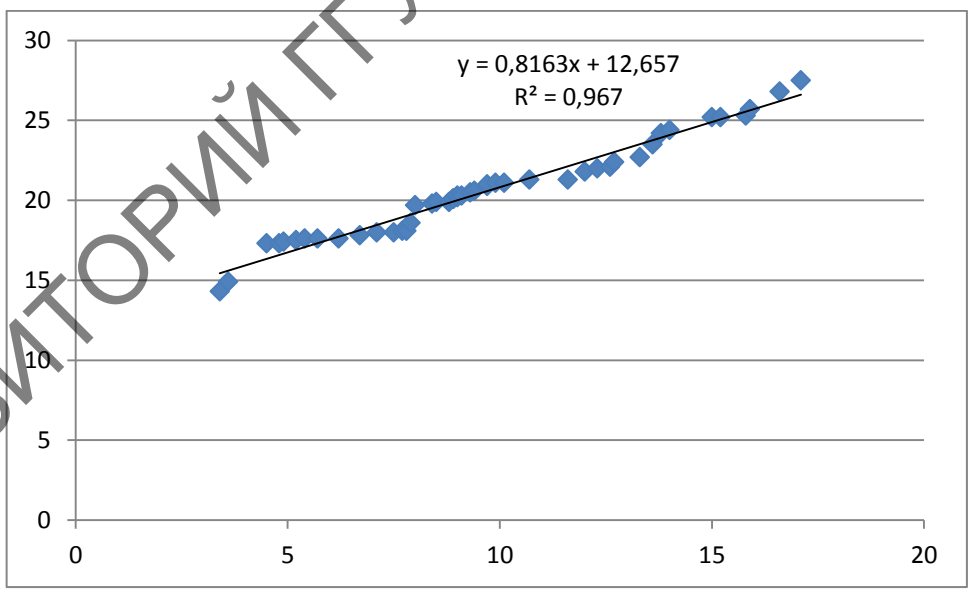


Рисунок 4.9 – Отредактированный график результатов регрессионного анализа в Excel

Уравнение регрессии и R^2 находятся в поле графика. Как видно, они такие же, как и при выполнении регрессионного анализа.

4.2 Линейный регрессионный анализ в STATISTICA

Для примера расчёта воспользуемся данными, отражающими изменения в приросте цыплят в граммах (переменная *y*, зависимая) в зависимости от увеличения в корме комплексной кормовой добавки в граммах (переменная *x*, независимая) (таблица 4.1).

Таблица 4.1 – Изменение веса цыплят при использовании комплексной кормовой добавки

В граммах

X	0,6	0,8	0,9	1,5	1,9	3,6	4,6	5,1	6,1
Y	5,1	5,1	5,7	14,4	16,2	16,3	21,3	22,5	28,2

Шаг 1. Создание электронной таблицы с данными.

Для проведения анализа необходимо предварительно составить таблицу с данными по признакам двух переменных – X и Y, представленных в виде двух столбцов (рисунок 4.10).

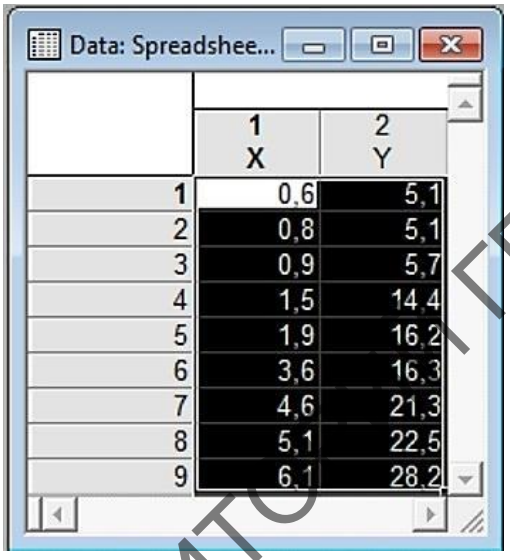


Рисунок 4.10 – Электронная таблица данных для расчёта регрессии

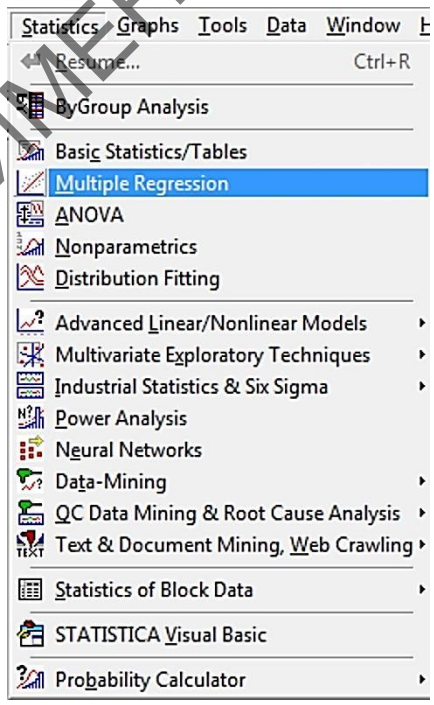


Рисунок 4.11 – Выбор модуля **Correlation Matrices**

Шаг 2. Выбор анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **Multiple Regression** (*Множественная регрессия*) (рисунок 4.11) и нажать **ОК**.

Шаг 3. Указание переменных.

В появившемся диалоговом окне (рисунок 4.12) необходимо выбрать переменные, которые необходимо включить в анализ.

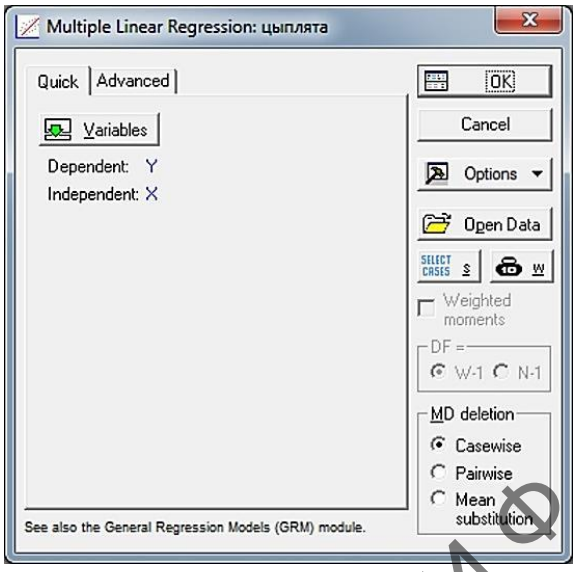


Рисунок 4.12 – Диалоговое окно **Multiple Linear Regression**

Для этого необходимо нажать кнопку **Variables** (*Переменные*), указать зависимую (**Dependent variable**) и независимую (**Independent variable**) переменные (рисунок 4.13) и нажать **ОК**.

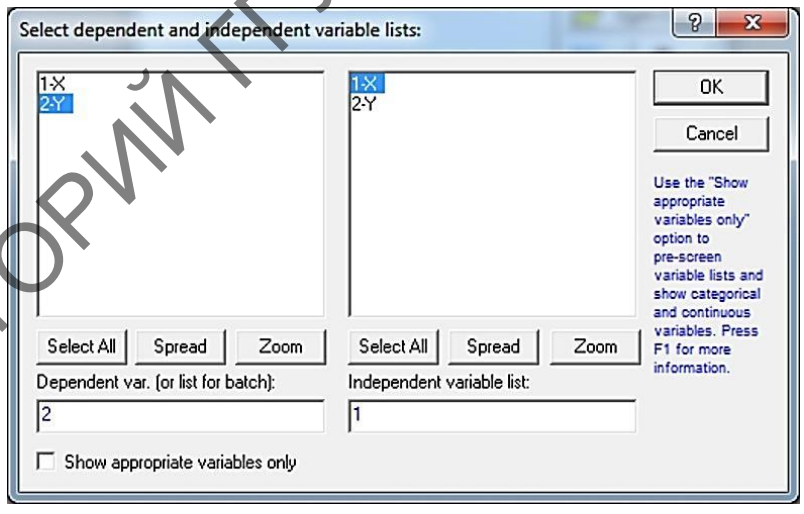


Рисунок 4.13 – Указание зависимой и независимой переменных

Шаг 4. Выставление параметров.

В появившемся диалоговом окне необходимо перейти на закладку **Advanced** (*Расширенные настройки*), указать в окне выбора **Input file** (*Способ ввода файла с данными*) ввод данных в рядах (**Raw data**),

поставить «галочку» в строке **Advanced options (stepwise or ridge regression)** (*Продвинутые настройки* (*ступенчатая или сглаженная регрессия*)) (рисунок 4.14) и щелкнуть мышкой **OK**.

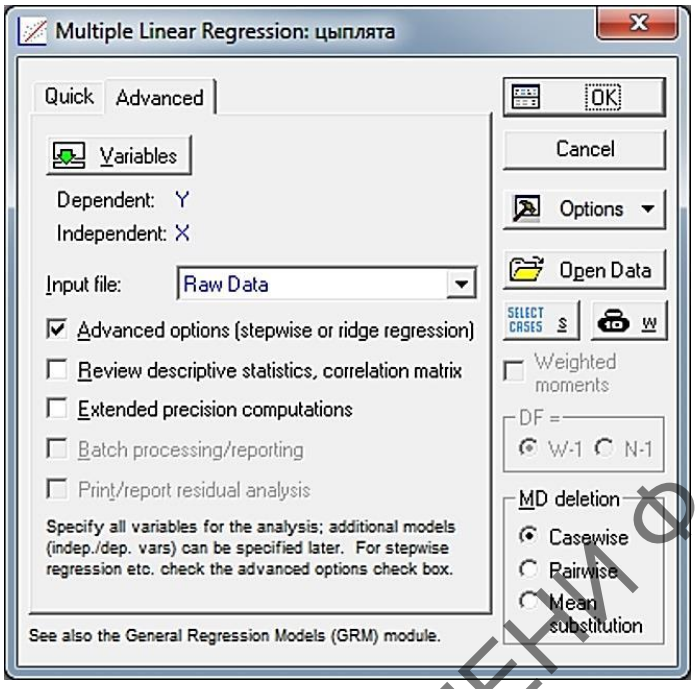


Рисунок 4.14 – Стартовая панель модуля **Multiple regression**

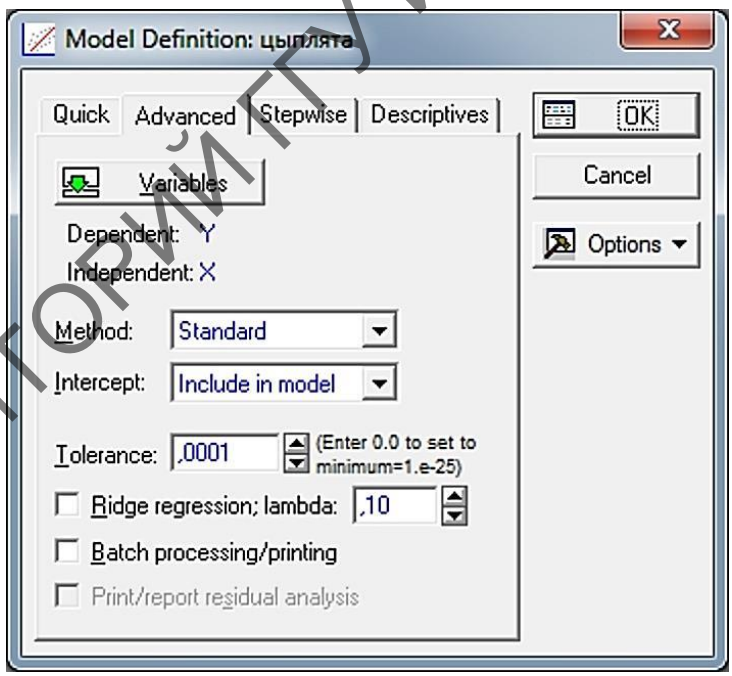


Рисунок 4.15 – Окно построения модели в модуле **Multiple regression**

В появившемся окне подробных настроек необходимо выбрать стандартный метод оценивания – в опции **Method** (*Метод*): **Standard** (*Стандартный*) (рисунок 4.15) и далее нажать кнопку **OK**.

Шаг 5. Проведение анализа.

Программа произведет оценивание параметров модели и на экране появится диалоговое окно оценивания параметров (рисунок 4.16):

– **Dependent**: имя зависимой переменной (в нашем случае – *Y*);

– **No. of cases**: число наблюдений (в нашем случае – 9);

– **Intercept**: значение свободного члена (*a*) регрессионного (линейного) уравнения;

– **Std. error:** стандартная ошибка свободного члена (*a*) регрессионного (линейного) уравнения;

– **Multiple R**: коэффициент множественной корреляции;

– **R?**: коэффициент детерминации R^2 (квадрат коэффициента множественной корреляции). Очень важный показатель регрессионного анализа, изменяется в пределах от 0 до 1 и отражает своеобразное «качество» рассчитанной регрессии, показывая долю (в процентах) общего разброса выборочных точек, которая объясняется построенной регрессией (например, при $R^2 = 0{,}897$, как в нашем случае, следует вывод о том, что практически 90 % дисперсии зависимой переменной *Y* объясняется вариацией независимой переменной *X*);

– **Adjusted R?**: скорректированный на число степеней свободы коэффициент детерминации;

– **Standard error of estimate**: стандартная ошибка оценки (мера рассеяния наблюдаемых значений относительно регрессионной прямой;

– **F, df и p**: F-критерий, число степеней свободы, принятое при его расчете, и вероятность ошибки для нулевой гипотезы F-теста (в регрессионном анализе именно F-тест применяется для оценки статистической значимости модели). При $p < 0{,}05$ считается, что рассчитанная регрессия в достаточной мере достоверно описывает связь между исследуемыми признаками;

– **t(df) и p**: критерий Стьюдента *t* (используется для проверки нулевой гипотезы о равенстве 0 – свободного члена регрессионного уравнения), p – вероятность ошибки для этой нулевой гипотезы;

– **X beta**: стандартизованный коэффициент регрессии, т. е. коэффициент регрессии, который получается в случае предварительной стандартизации обеих переменных. Стандартизация переменных – такое преобразование, когда их средние значения стали бы равны 0, а стандартные отклонения – 1). Расчет **beta** позволяет оценить, в какой степени значения зависимой переменной определяются значениями независимой переменной. При наличии одной независимой переменной коэффициент **beta** идентичен **Multiple R**.

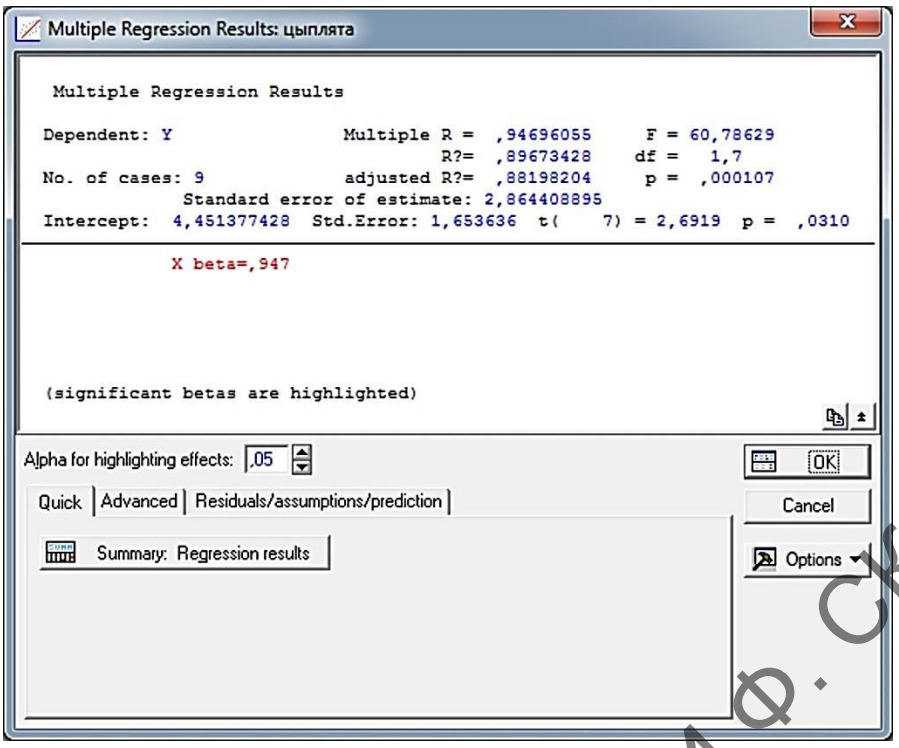


Рисунок 4.16 – Окно оценивания параметров

Шаг 6. Итоговые результаты.

Нажмите кнопку **Regression summary** (*Итоговый результат регрессии*). На экране появится электронная таблица вывода, в которой представлены итоговые результаты оценивания регрессионной модели (рисунок 4.17):

– **Beta**: стандартизованный коэффициент регрессии;

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

– **B**: свободный член регрессионного уравнения (в строке *Intercept*) и коэффициента регрессии (нижняя строка таблицы);

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

– **t(df)**: значения t-критерия Стьюдента (используется для проверки гипотезы о равенстве обоих коэффициентов уравнения нулю);

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

Regression Summary for Dependent Variable: Y (цыплята)
R= ,94696055 R?= ,89673428 Adjusted R?= ,88198204
F(1,7)=60,786 p<,00011 Std.Error of estimate: 2,8644

N=9	Beta	Std.Err. of Beta	B	Std.Err. of B	t(7)	p-level
Intercept			4,451377	1,653636	2,691873	0,031002
X	0,946961	0,121459	3,774406	0,484112	7,796556	0,000107

Рисунок 4.17 – Итоговая таблица регрессии

Из результатов, отражённых на рисунке 4.17, видно, что оба коэффициента регрессии статистически значимо отличаются от 0 ($p < 0,05$ и $p << 0,001$; подсвечено красным). То есть в целом построенная регрессионная модель отлично описывает связь между привесом цыплят и кормлением их комплексной кормовой добавкой.

Шаг 7. Проверка остатков.

После проведённого регрессионного анализа нелишним будет провести так называемый «анализ остатков» (остатки – это разности между наблюдаемыми значениями зависимой переменной и теми её значениями, которые предсказываются регрессионной моделью).

Для этого необходимо вернуться в окно анализа регрессии, нажав на кнопку в левом нижнем углу экрана, а затем в открывшемся диалоговом окне перейти на закладку **Residuals/Assumptions/ Predictions** (*Остатки/Условия/Предсказания*) и нажать кнопку **Perform residual analysis** (*Выполнить анализ остатков*) (рисунок 4.18).

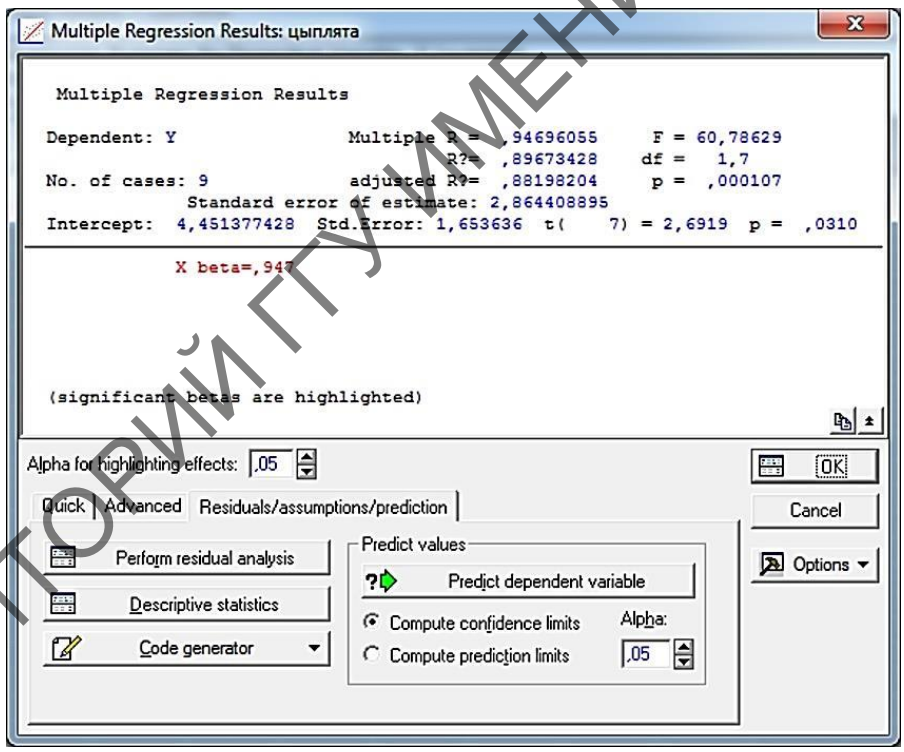


Рисунок 4.18 – Закладка **Residuals/Assumptions/ Predictions**

Анализ остатков проводится по следующей схеме:

1) Нормальность распределения остатков.

Для этого на закладке **Quick** (*Быстрый расчёт*) подмодуля анализа остатков (рисунок 4.19) нужно нажать кнопку **Normal plot of residuals** (*Нормальный график остатков*), чтобы построить график

нормальных вероятностей. Если точки на этом графике достаточно тесно укладываются вдоль теоретически ожидаемой прямой, можно заключить, что остатки распределяются нормально (рисунок 4.20). В противном случае линейная регрессионная модель для анализируемых переменных будет неприменима.

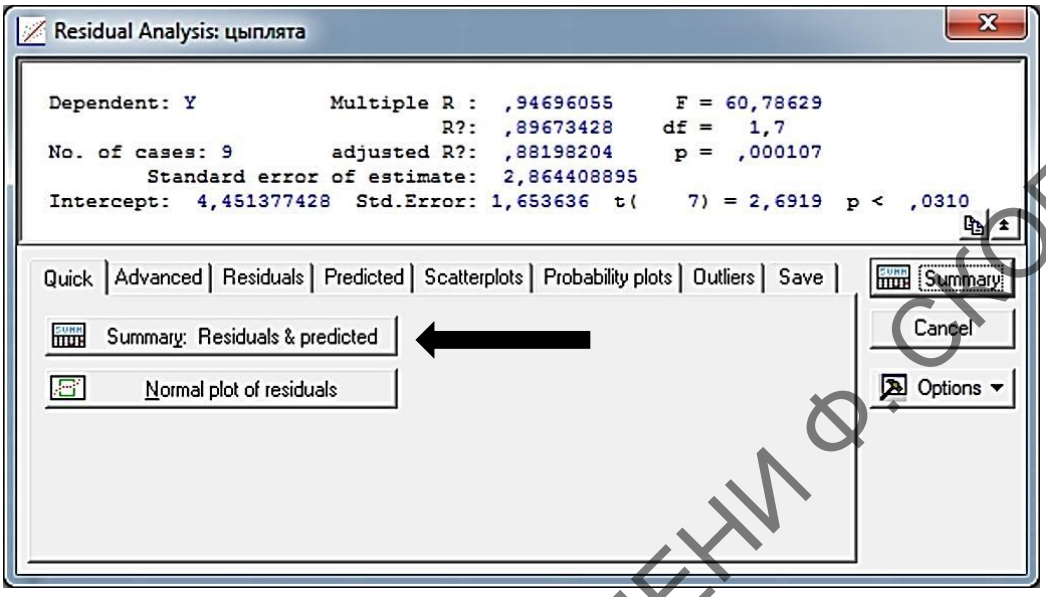


Рисунок 4.19 – Подмодуль анализа остатков модуля регрессионного анализа

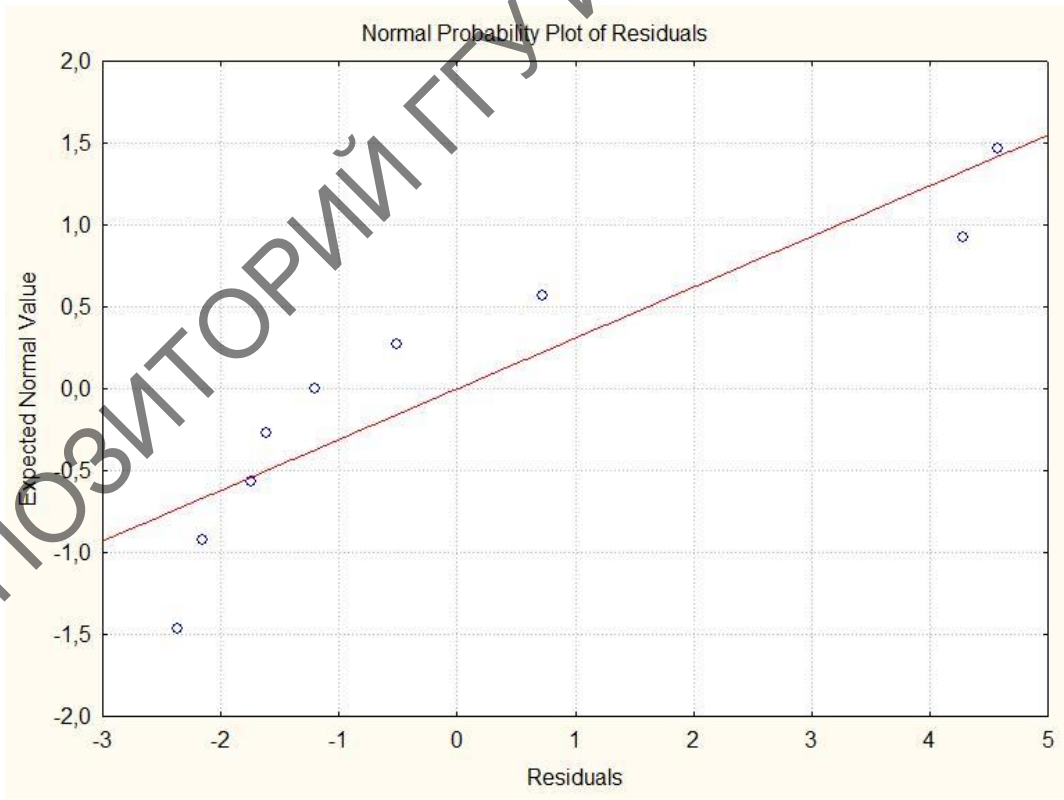


Рисунок 4.20 – Итог проверки на нормальность остатков

В приведённом анализе наблюдается довольно значительный разброс значений от прямой, поэтому можно сказать, что анализируемые данные не подчиняются закону нормального распределения.

2) Неизменность дисперсии.

Для проверки этого условия необходимо в подмодуле оценки остатков перейти на закладку **Scatterplots** (*Диаграммы рассеяния*) и нажать кнопку **Predicted vs. Residuals** (*Предсказанные против остатков*), чтобы построить график зависимости значений остатков от предсказываемых моделью значений зависимой переменной (рисунок 4.21).

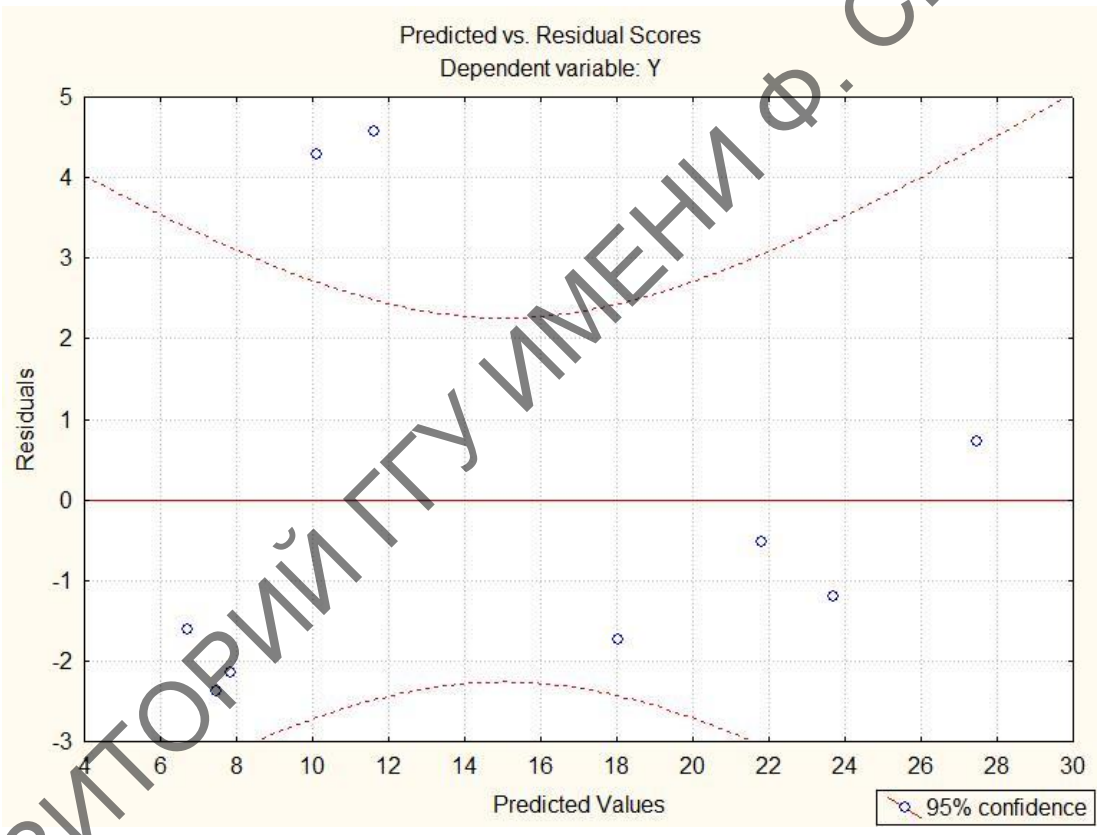


Рисунок 4.21 – Результат проверки однородности дисперсии

Проверяемое условие выполняется в том случае, если точки на данном графике будут располагаться хаотично и не проявлять закономерностей. В противном случае (в расположении точек имеется тенденция – разброс увеличивается слева направо, точки тесно укладываются вдоль прямой и др.), линейный регрессионный анализ неприменим. В этом примере график демонстрирует именно такую зависимость.

Задания

1) У 20 взрослых мужчин были измерены высота (длина тела) x (в см) и вес y (в кг):

x	165	176	175	168	167	172	175	180	179	173
y	56	75	70	61	61	63	72	80	76	68
x	166	178	169	169	170	176	180	169	177	176
y	58	76	60	64	63	71	78	63	75	71

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

2) Предполагается, что между количеством настриженной шерсти y и живым весом овец x имеется зависимость. Для 10 овец были получены следующие данные (в кг):

x	50	55	60	50	65	60	50	55	50	65
y	4,0	4,2	4,1	4,2	4,5	4,3	4,1	4,4	4,0	4,2

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

3) Были получены следующие данные о длине грудного x и брюшного у плавника у окуня озера Баторино:

x	38	31	36	43	29	33	28	25	36	26	21	30
y	40	34	38	42	26	33	29	26	36	27	22	32
x	27	27	28	26	26	25	24	28	28	27	33	27
y	28	26	32	26	28	27	25	28	30	26	32	27
x	26	23	22	25	24	29	25	25	30	23	24	32
y	29	23	24	30	26	30	27	28	32	23	24	32
x	24	25	30	25	26	30	29	22	29	28	26	28
y	25	27	33	27	27	32	28	24	31	32	27	30
x	25	31	25	32	27	31	28	29	26	32	27	31
y	25	34	26	32	29	30	29	29	26	35	26	33
x	28	28	26	33	30	27	21	28	26	30	23	27
y	29	31	29	33	31	31	23	30	27	29	24	28

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

4) Для 10 петушков леггорнов 15-дневного возраста были получены следующие данные о весе их тела x (в г) и весе гребня y (в мг):

x	83	72	69	90	90	95	95	91	75	70
y	56	42	18	84	56	107	90	68	31	48

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

5) Путем еженедельного взятия проб с поля было изучено изменение высоты растений сои y (в см) с возрастом x (в неделях):

x	1	2	3	4	5	6	7
y	5	13	16	23	33	38	40

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

6) Для установления связи между содержанием фосфора в почве x и содержанием фосфора в злаковых растениях y было проведено 9 анализов со следующими результатами:

x	1	4	5	9	13	11	23	23	28
y	64	71	54	81	93	76	77	95	109

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

7) Было проведено сравнение удоев первой лактации x с третьей y по 33 коровам холмогорской помеси (в л):

x	1522	239	1521	2700	1789	2496	1197	1105	1701	2218	1790
y	3693	4453	1446	2134	2940	4353	2066	2152	2396	2435	3140
x	2964	1287	1756	1406	1810	1299	2609	2519	1927	1655	1320
y	4700	2113	2513	3249	2553	2320	4612	3201	3173	3326	1639
x	2586	1928	3884	2968	2200	1753	1508	1803	1811	2300	1697
y	4562	3482	4257	3465	2448	3435	3747	2112	3061	2985	2721

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

8) На белых крысах была показана следующая зависимость между температурой внешней среды x (в град.) и количеством поглощенного кислорода y (в мл/г веса):

x	0	5	10	15	20	25	28	29
y	3,83	3,35	2,60	2,02	1,69	1,42	1,39	1,38
x	30	31	32	33	34	35	40	
y	1,29	1,39	1,39	1,45	1,65	1,61	2,40	

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

9) Вычислите коэффициент регрессии по следующему ряду данных (в мм) о длине хвоста x и общей длине тела y самок королевской змеи *Lampropeltis polyzona*:

x	37	49	50	51	53	54	68	86	93	106
y	284	375	353	366	418	408	510	627	683	820
x	130	137	142	142	146	149	155	156	187	
y	1056	986	1086	1086	1078	1122	1254	1202	1387	

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

10) Между возрастом овцематок x (в годах) и длительностью плодоношения ягнят y (в днях) оказалась следующая зависимость:

x	2	3	4	5	6	7
y	149,5	149,3	150,0	150,9	150,5	151,4

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

11) Фактическая урожайность зерновых культур (в ц/га) в одном совхозе по годам была следующей:

Годы	1953	1954	1955	1956	1957	1958	1959
Урожайность	7,8	7,7	8,5	10,0	8,4	8,3	10,5

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

12) У 10 телят по глубине груди x (в см) и живому весу y (в кг) были получены следующие данные:

x	91	86	94	95	104	92	98	84	96	99
y	62	43	60	73	87	65	79	52	65	68

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

13) Известны данные для 10 бычков о весе при рождении x (в кг) и суточном привесе y (в г):

x	38,5	46,0	43,0	43,0	40,5	44,0	38,0	35,0	40,5	54,0
y	694	901	736	1005	841	743	896	863	855	830

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

14) Были получены следующие данные о потреблении кислорода y пиявками (в г на кг/час) в зависимости от температуры x (в градусах):

x	5,5	5,6	6,2	8,4	9,0	10,5	16,1
y	16,1	14,9	18,8	32,5	32,1	37,1	88,5
x	16,6	17,1	18,8	19,8	20,0	20,7	26,5
y	91,0	94,0	122,0	162,0	167,0	187,0	436,0

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

15) Под влиянием облучения рентгеновыми лучами наблюдалось следующее замедление размножения вируса мозаики Аукуба y (в тыс.) в зависимости от длительности облучения x (в мин):

x	0	3	7,5	15	30	45	60
y	271	226	209	108	59	29	12

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

16) Было учтено среднесуточное количество перевариваемых веществ корма x (в кг), съеденного коровой за 12 месяцев лактации:

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
x	14,3	12,8	13,2	13,6	13,4	13,2	12,9	12,8	12,5	12,2	11,9	11,5

Определите коэффициент регрессии, постройте график и уравнение линейной регрессии, проверьте остатки.

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

4 Мастицкий, С. Э. Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований / С. Э. Мастицкий. – Минск : РУП «Институт рыбного хозяйства», 2009. – 76 с.

5 Рокицкий, П. Ф. Биологическая статистика / П. Ф. Рокицкий. – Минск : «Вышэйшая школа», 1973. – 320 с.

ТЕМА 5. КРИВОЛИНЕЙНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ В СРЕДЕ EXCEL И STATISTICA 7.0

5.1 Криволинейная корреляция и регрессия в среде MS Excel.
5.2 Криволинейная корреляция и регрессия в среде STATISTICA 7.0.

5.1 Криволинейная корреляция и регрессия в среде MS Excel

Если связь между изучаемыми явлениями существенно отклоняется от пропорциональной (это легко установить по графику), то коэффициент линейной корреляции непригоден в качестве меры связи. Он может указать на отсутствие сопряженности там, где налицо сильная криволинейная зависимость. Поэтому необходим новый показатель, который правильно измерял бы степень криволинейной зависимости. Таким показателем является корреляционное отношение, обозначаемое греческой буквой η (эта). Оно измеряет степень корреляции при любой ее форме.

Криволинейная связь между признаками – это такая связь, при которой равномерным изменениям первого признака соответствуют неравномерные изменения второго, причем эта неравномерность имеет определенный закономерный характер.

При графическом изображении криволинейных связей, когда по оси абсцисс откладывают значения первого признака (аргумент, независимая переменная), а по оси ординат – значения второго признака (функция, зависимая переменная) и полученные точки соединяют, получают изогнутые линии. Характер изогнутости зависит от природы коррелируемых признаков.

Корреляционное отношение измеряет степень корреляции при любой ее форме.

В отличие от коэффициента линейной корреляции, который дает одинаковую меру связи признаков (первого со вторым и второго с первым), корреляционное отношение второго признака по первому обычно не бывает равно корреляционному отношению первого признака по второму. Поэтому крайне важно определить, какая выборка является аргументом, а какая функцией.

По виду линии на графике можно определить характер связи (прямолинейная или криволинейная), также тип аппроксимации.

Задачей исследователя является подобрать вид функции, которая бы наиболее четко ложилась на поле регрессии.

Рассмотрим криволинейную корреляцию и регрессию на следующем примере.

Исследовалось влияние вводимого перорально с кормом сорбента (x) в граммах молочным коровам на снижение содержания ^{137}Cs в молоке (y) в процентах (таблица 5.1).

Таблица 5.1 – Дозы вводимого сорбента (x) и снижение содержания ^{137}Cs в молоке (y)

x	14,5	5,1	6,1	6,5	6,7	8,2	7,4	10,7	14,3	8,1	8,1	4,6	8,4
y	49,8	22,5	28,2	31,2	33,5	41,8	34,9	46,9	48,8	40,3	40,3	21,3	42,2
x	9,6	8,8	9,1	9,6	7,4	8,5	11,2	12,7	13,4	13,8	7,2	8,0	
y	44,9	43,9	44,3	44,8	36,2	43,9	47,1	47,9	48,2	48,7	34,1	40,3	

Шаг 1. Создание электронной таблицы с данными.

При создании таблицы с данными необходимо каждый из признаков (другими словами – отдельную переменную) разместить в отдельном столбце, т. е., в конечном итоге получим 2 столбца с данными (рисунок 5.1).

	A	B
1	x	y
2	14,5	49,8
3	5,1	22,5
4	6,1	28,2
5	6,5	31,2
6	6,7	33,5
7	8,2	41,8
8	7,4	34,9
9	10,7	46,9
10	14,3	48,8
11	8,1	40,3
12	8,1	40,3
13	4,6	21,3
14	8,4	42,2
15	9,6	44,9
16	8,8	43,9
17	9,1	44,3
18	9,6	44,8
19	7,4	36,2
20	8,5	43,9
21	11,2	47,1
22	12,7	47,9
23	13,4	48,2
24	13,8	48,7
25	7,2	34,1
26	8	40,3

Рисунок 5.1 – Создание ряда данных в книге Excel

Шаг 2. Предварительный график.

Необходимо выделить наш диапазон данных без заголовка (!), а затем в главном меню Excel выбрать опцию **Вставка,** перейти в блок меню **Диаграммы**, после чего выбрать тип диаграммы **Точечная** (рисунок 5.2).

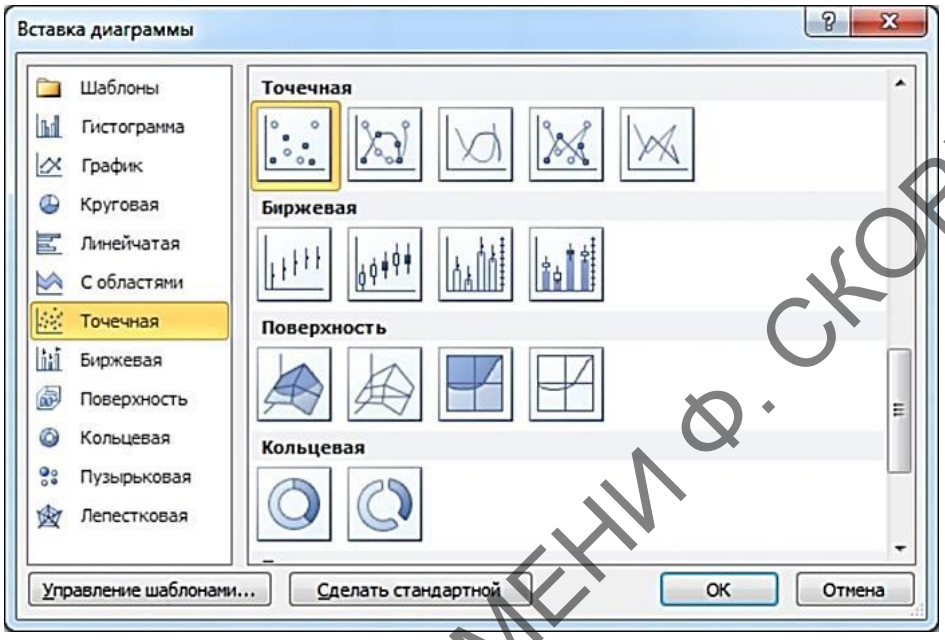


Рисунок 5.2 – Вставка диаграммы в Excel

После выбора диаграммы необходимо щелкнуть мышкой **OK**. В результате получим график, показанный на рисунке 5.3.

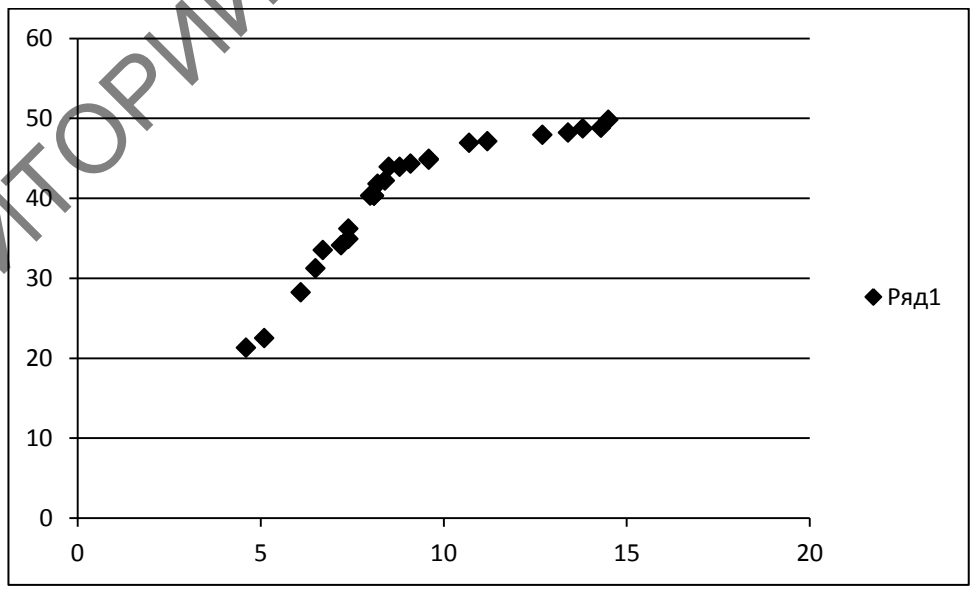


Рисунок 5.3 – Начальный график результатов криволинейной корреляции и регрессии в Excel

Шаг 3. Настройка графика.

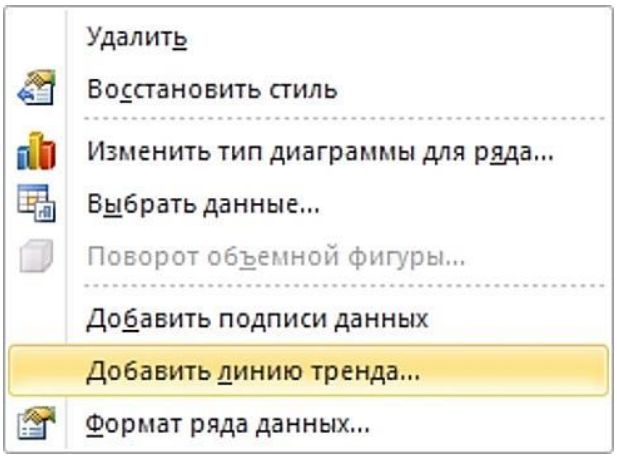


Рисунок 5.4 – Контекстное меню

Далее необходимо щёлкнуть правой кнопкой по любой из точек и в контекстном меню выбрать опцию **Добавить линию тренда** (рисунок 5.4).

После этого в диалоговом окне **Формат линии тренда**, в правой его части, нужно выбрать тип аппроксимации – **Полиноминальная 2 степени**, а остальные установки – как показано на рисунке 5.5 и нажать кнопку **Закрыть**.

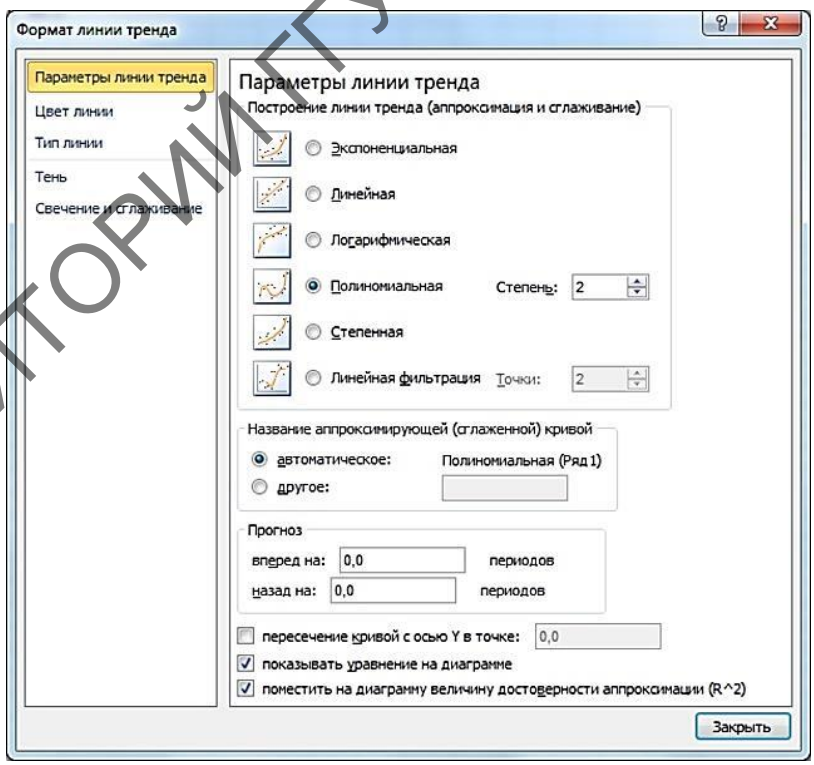


Рисунок 5.5 – Настройки линии тренда в Excel

После появления линии тренда необходимо щёлкнуть правой клавишей мыши по полю с легендой и удалить её. Окончательный результат отобразится на рисунке 5.6.

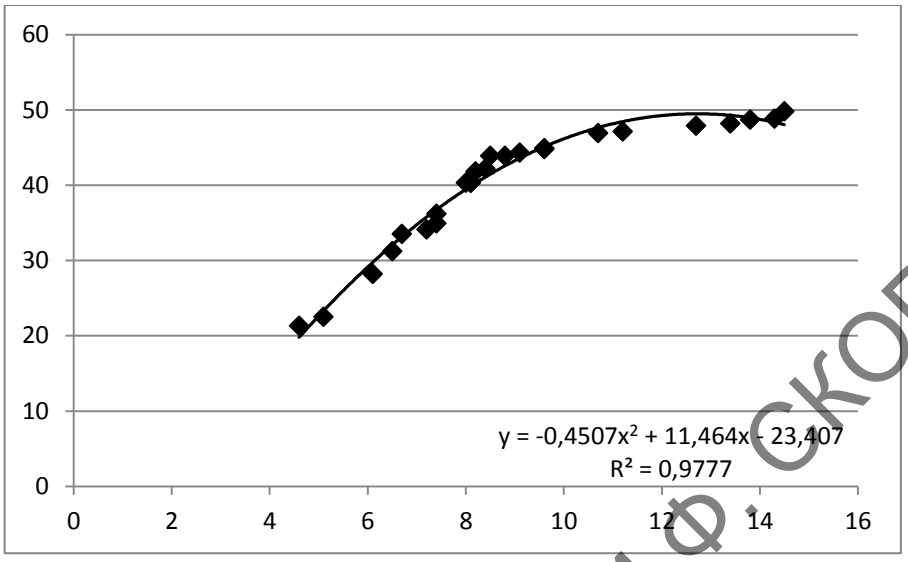


Рисунок 5.6 – Отредактированная диаграмма

Уравнение регрессии и квадрат корреляционного отношения находятся в правом нижнем углу диаграммы.

Результаты анализа показали, что корреляционная связь между массой вводимого сорбента и уровнем снижения содержания ^{137}Cs в молоке велика $r = 0{,}99$ и описывается полиномом 2 степени:

$$y = -0{,}4507 \cdot x^2 + 11{,}464 \cdot x - 23{,}407.$$

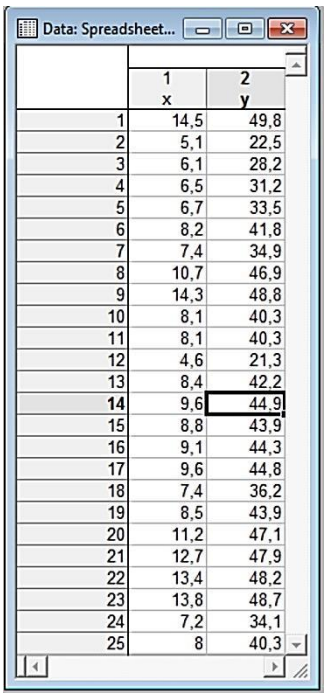
На графике (рисунок 5.6) видно, что при массе вводимого сорбента больше 12 г наблюдается насыщение. Поэтому оптимальное значение массы сорбента следует рекомендовать в пределах 12–14 г.

5.2 Криволинейная корреляция и регрессия в среде STATISTICA 7.0

Расчёт криволинейной корреляции и регрессии в среде STATISTICA 7.0 рассмотрим на предыдущем примере о сорбенте.

Шаг 1. Создание электронной таблицы с данными.

Для проведения анализа необходимо предварительно составить таблицу с данными по признакам двух переменных – *x* и *y*, представленных в виде двух столбцов (рисунок 5.7). При этом следует иметь в виду, что *x* – это независимая переменная, а *y* – зависимая переменная.



	1 x	2 y
1	14,5	49,8
2	5,1	22,5
3	6,1	28,2
4	6,5	31,2
5	6,7	33,5
6	8,2	41,8
7	7,4	34,9
8	10,7	46,9
9	14,3	48,8
10	8,1	40,3
11	8,1	40,3
12	4,6	21,3
13	8,4	42,2
14	9,6	44,9
15	8,8	43,9
16	9,1	44,3
17	9,6	44,8
18	7,4	36,2
19	8,5	43,9
20	11,2	47,1
21	12,7	47,9
22	13,4	48,2
23	13,8	48,7
24	7,2	34,1
25	8	40,3

Рисунок 5.7 – Электронная таблица данных для расчёта

Шаг 2. Выбор анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*) и в нём – модуль **Advanced Linear/ Nonlinear Models** (*Расширенные линейные/нелинейные модели*), а затем опцию **Nonlinear estimation** (*Нелинейная оценка*) (рисунок 5.8) и нажать **ОК**.

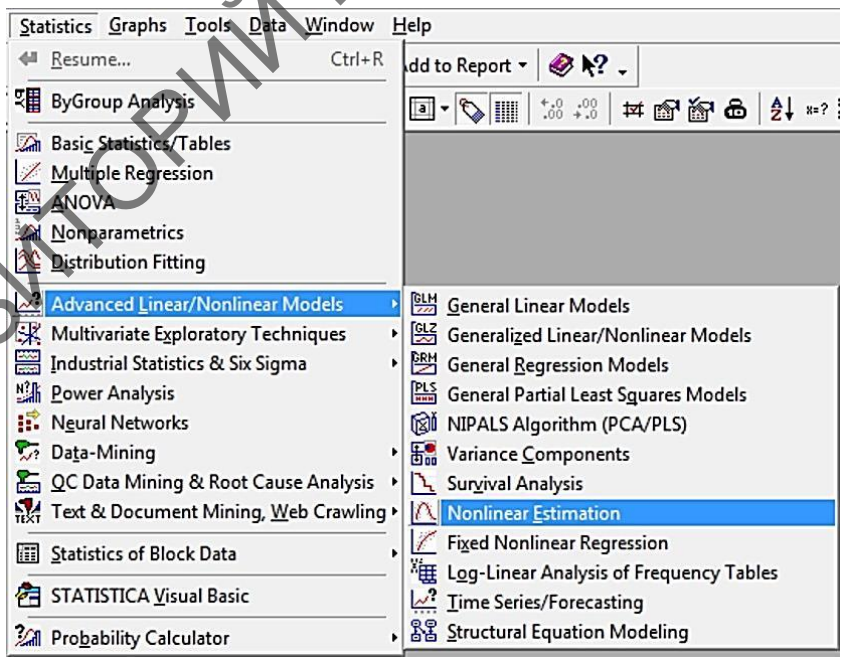


Рисунок 5.8 – Запуск модуля **Nonlinear estimation**

Шаг 3. Выбор модуля.

В появившемся на экране диалоговом окне необходимо выбрать опцию **User specified regression, least squares** (*Регрессия, определяемая пользователем, метод наименьших квадратов*) и далее щелкните мышью **ОК** (рисунок 5.9).

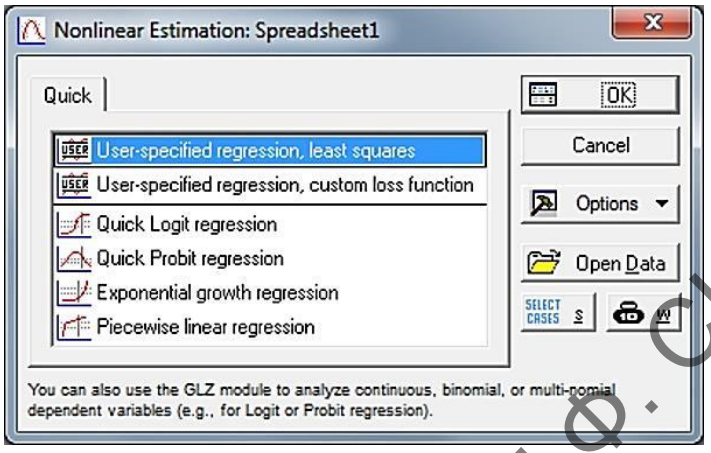


Рисунок 5.9 – Диалоговое окно модуля **Nonlinear estimation**

Шаг 4. Ввод функции.

В появившемся окне нужно щелкнуть мышью по кнопке **Function of estimated** (*Предполагаемая функция*) (рисунок 5.10).

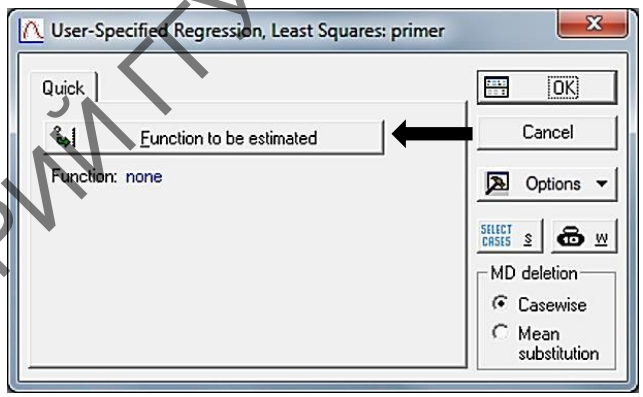


Рисунок 5.10 – Панель ввода функции

В предлагаемом поле нового окна ввода функции с клавиатуры введите предполагаемую функцию. В отличие от подобной операции в табличном редакторе MS Excel в STATISTICA можно ввести любую формулу, связывающую зависимую и независимую переменные.

В данном примере предполагается, что наиболее подходящей функцией является полином второй степени типа:

$$y = a + bx + cx^2$$

или, в конкретном случае, в соответствии с таблицей исходных данных: *v*2 = *a*+*b***v*1+*c***v*1**2, где *v*1 – независимая переменная *x*, а *v*2 – зависимая переменная *y* (рисунок 5.11).

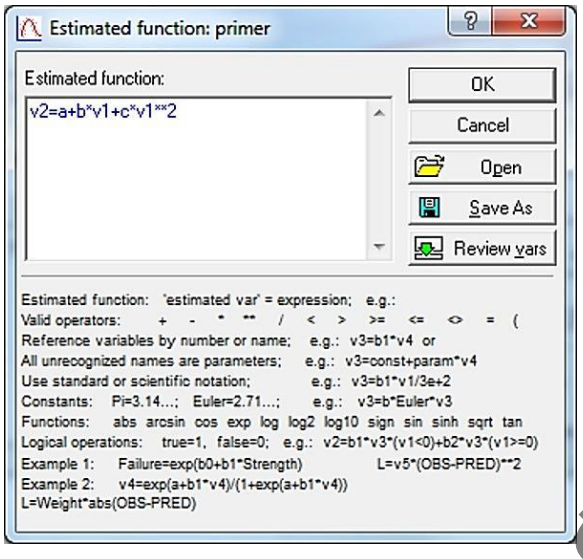


Рисунок 5.11 – Ввод функции

В нижней части рисунка приведен перечень алгебраических и функциональных символов, которые воспринимаются программой. После ввода функции необходимо нажать **ОК**, программа перейдёт в предыдущее окно, и затем нужно ещё раз нажать **ОК** для вывода промежуточных результатов (рисунок 5.12).

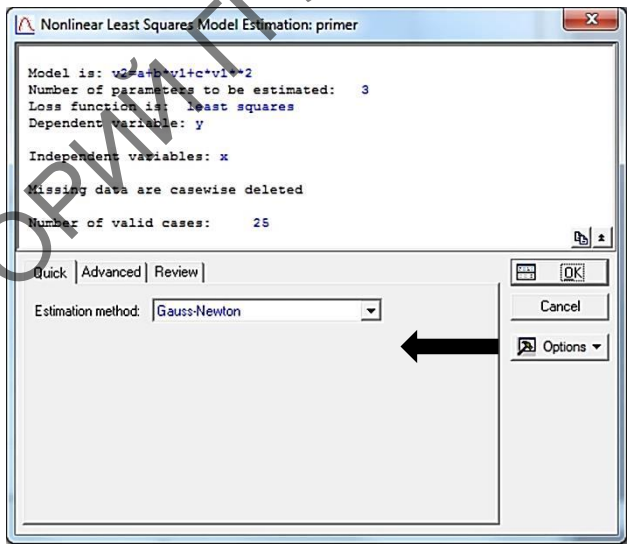


Рисунок 5.12 – Диалоговое окно запуска аппроксимации

Шаг 5. Расчёт параметров.

Верхняя часть окна информирует о следующем:

– **Model is:** *модель* (в нашем случае *v*2 = *a*+*b***v*1+*c***v*1**2, т. е. $y = a+bx+cx^2$);

– **Number of parameters to be estimated:** *количество оцениваемых параметров* (в нашем случае – 3);

– **Loss function is:** *используемая в модели функция* (в нашем случае – *наименьших квадратов*);

– **Dependent variable:** *зависимая переменная* (в нашем случае – *y*);

– **Independent variables:** *независимые переменные* (в нашем случае – *x*);

– **Missing data are casewise deleted:** *пропущенные данные удаляются*;

– **Number of valid cases:** *количество действительных случаев* (в нашем случае – анализируемых пар, т. е. 25).

В середине окна, в поле с надписью **Estimation method** (*Метод оценки*), необходимо раскрыть список и выбрать метод аппроксимации, например: **Gauss–Newton** (рисунок 5.12), и нажать **OK**.

Шаг 6. Просмотр результатов.

В верхней части появившегося окна результатов (рисунок 5.13) в разделе **Proportion of variance accounted for** (*Соотношения переменных рассчитаны для*) показаны значения корреляционного отношения (0,98879305) и его квадрата (0,97771169). Это указывает на сильную корреляционную связь между переменными.

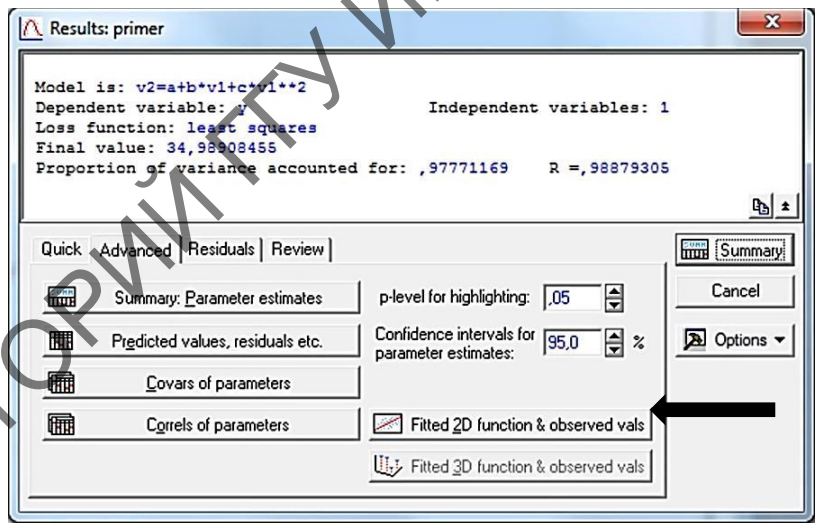


Рисунок 5.13 – Окно результатов

Шаг 7. Оценка результатов.

Для просмотра результатов для каждого элемента уравнения $y = a+bx+cx^2$ (полинома второй степени, который был заложен первоначально как исходная функция) необходимо щелкнуть мышью по кнопке **Summary: Parameters & standard errors** (*Итоговые параметры и стандартные ошибки*). Полученные результаты

(рисунок 5.14) подкрашены красным цветом, что свидетельствует о достоверности аппроксимации функцией: $y = –0{,}45x^2 + 11{,}46x – 23{,}41$.

Model is: y=a+b*x+c*x**2 (primer)
Dep. Var. : y
Level of confidence: 95.0% (alpha=0.050)

	Estimate	Standard error	t-value df = 22	p-level	Lo. Conf Limit	Up. Conf Limit
a	-23,4071	3,010054	-7,7763	0,000000	-29,6496	-17,1646
b	11,4636	0,641511	17,8697	0,000000	10,1332	12,7940
c	-0,4507	0,032066	-14,0558	0,000000	-0,5172	-0,3842

Рисунок 5.14 – Результаты аппроксимации

Столбцы имеют следующие обозначения:
– **Estimate** (*Оценка*);
– **Standard Error** (*Стандартная ошибка*);
– **t-value df=22** (*t-критерий при 22 степенях свободы*);
– **p-level** (*уровень значимости меньше 0,05*);
– **Lo. Conf. Limit** (*Верхний предел достоверности*);
– **Up. Conf. Limit** (*Нижний предел достоверности*).

Шаг 8. Графическое отображение результатов.

Для графического отображения результатов необходимо кликнуть левой клавишей мыши по кнопке **Fitted 2D function & observed vals** (*Подогнанная функция*) (рисунок 5.13). На рисунке 5.15 отражена получившаяся графическая интерпретация корреляционной связи исходных массивов в виде заданной функции: $y = –0{,}45x^2 + 11{,}46x – 23{,}41$.

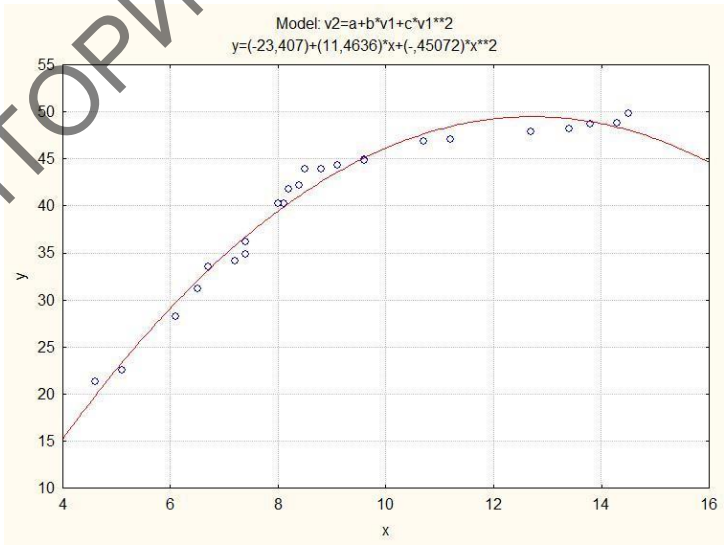


Рисунок 5.15 – Графическая интерпретация криволинейной корреляционной связи

Шаг 9. Анализ достоверности регрессии.

В окне результатов (рисунок 5.13) необходимо перейти на закладку **Quick** и нажать кнопку **Analysis of Variance** (*Анализ вариантов*) (рисунок 5.16).

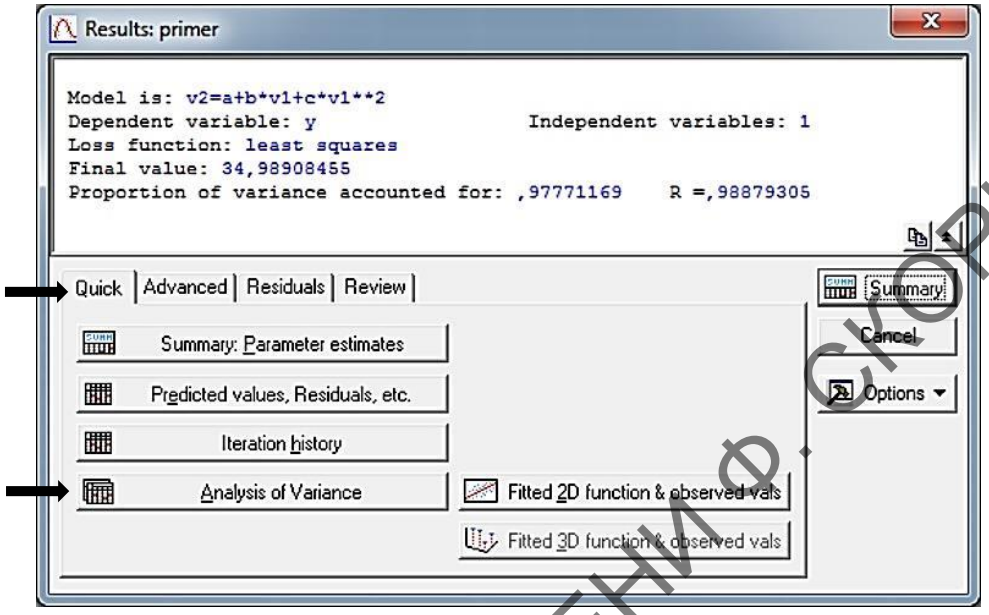


Рисунок 5.16 – Выбор опции **Analysis of Variance**

Результат выполненной операции представлен на рисунке 5.17, который свидетельствует о достоверности регрессии ($F = 8806{,}2$ при $p < 0{,}00$).

Effect	Model is: v2=a+b*v1+c*v1**2 (primer) Dep. Var. : y				
	1 Sum of Squares	2 DF	3 Mean Squares	4 F-value	5 p-value
Regression	42016,29	3,00000	14005,43	8806,160	0,00
Residual	34,99	22,00000	1,59		
Total	42051,28	25,00000			
Corrected Total	1569,84	24,00000			
Regression vs.Corrected Total	42016,29	3,00000	14005,43	214,118	0,00

Рисунок 5.17 – Результат анализа вариантов

Шаг 10. Просмотр результатов остатков.

Для просмотра наблюдаемых результатов в сравнении с предсказанными моделями необходимо в окне результатов (рисунок 5.13) перейти на закладку **Residuals** (*Остатки*) и нажать кнопку **Observed, predicted, residual vals** (*Наблюдаемый, предсказанный, остаточный*) (рисунок 5.18).

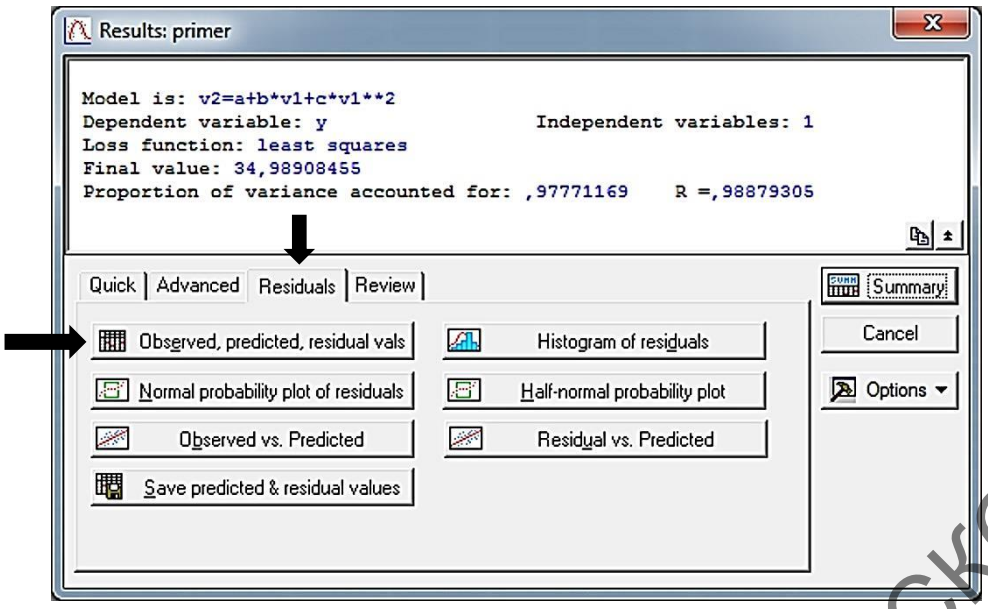


Рисунок 5.18 – Выбор опции **Observed, predicted, residual vals**

Результаты выполненной операции представлены на рисунке 5.19.

	Model is: v2=a+b*v1+c*v1**2 (prime Dep. Var. : y		
	Observed	Predicted	Residuals
1	49,80000	48,05252	1,74748
2	22,50000	23,33429	-0,83429
3	28,20000	29,74990	-1,54990
4	31,20000	32,06375	-0,86375
5	33,50000	33,16658	0,33342
6	41,80000	40,28853	1,51147
7	34,90000	36,74256	-1,84256
8	46,90000	47,65128	-0,75128
9	48,80000	48,35592	0,44408
10	40,30000	39,87683	0,42317
11	40,30000	39,87683	0,42317
12	21,30000	19,78845	1,51155
13	42,20000	41,08488	1,11512
14	44,90000	45,10577	-0,20577
15	43,90000	42,56941	1,33059
16	44,30000	43,58815	0,71185
17	44,80000	45,10577	-0,30577
18	36,20000	36,74256	-0,54256
19	43,90000	41,46953	2,43047
20	47,10000	48,44776	-1,34776
21	47,90000	49,48504	-1,58504
22	48,20000	49,27500	-1,07500
23	48,70000	48,95666	-0,25666
24	34,10000	35,76592	-1,66592
25	40,30000	39,45612	0,84388

Рисунок 5.19 – Наблюдаемые и аппроксимированные значения функции

Шаг 11. Оценка адекватности модели.

Для оценки адекватности полученной модели необходимо в закладке **Residuals** *(Остатки)* щелкнуть левой клавишей мыши по кнопке **Observed vs. Predicted** (*Наблюдаемые против предсказанных*)

(рисунок 5.20), после чего на экран будет выведен график массива наблюдаемых и предсказанных значений.

Исходя из визуальных результатов, отображённых на полученном графике (рисунок 5.21), видно, что, массивы наблюдаемых и предсказанных значений описываются линейной функцией $y = k \cdot x$. При этом $k = 1$ и коэффициент парной корреляции близок к 1.

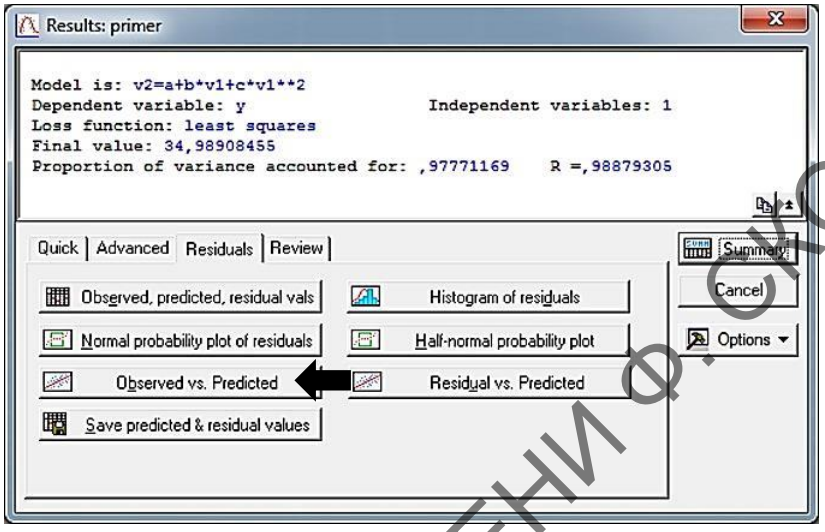


Рисунок 5.20 – Выбор опции **Observed vs. Predicted**

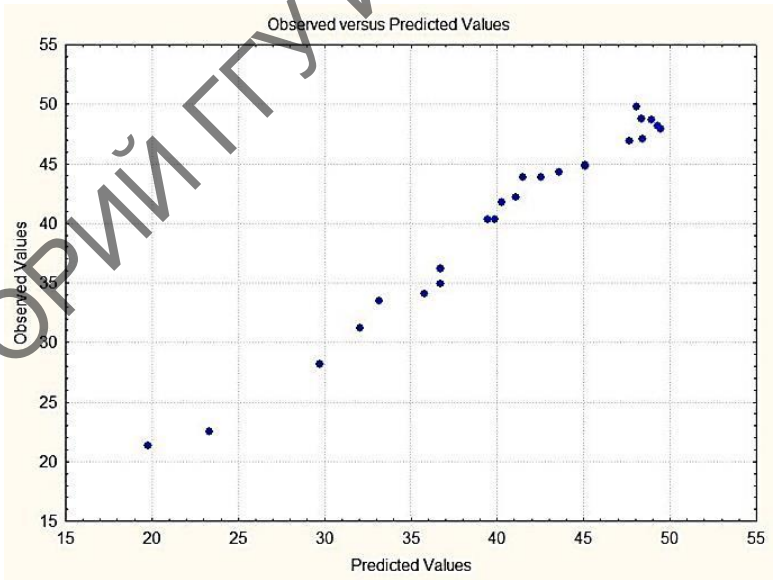


Рисунок 5.21 – Визуализация результатов анализа

В заключение следует добавить, что корреляция и регрессия достоверны, так как F = 8806,2 (рисунок 5.17) и t-value a ≈ 7,8; t-value b ≈ 17,9 и t-value c ≈ 14,1 (рисунок 5.14), что существенно выше критических значений при p < 0,00.

Задание

Выполните расчетные процедуры в соответствии с порядком операций для расчёта криволинейной корреляции и регрессии. Варианты 1–3 рассчитайте при помощи электронных таблиц Excel, варианты 4–7 – с использованием программного пакета STATISTICA 7.0. Объясните полученный результат.

Варианты заданий													
1		2		3		4		5		6		7	
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
14,5	49,8	33,3	0,03	9,06	2,13	0,9	66,9	0,97	0,4	1,5	0,7	0,87	0,15
5,1	22,5	7,1	0,2	8,95	2,12	0,01	0,01	1,21	2,79	1,47	0,69	0,98	0,11
6,1	28,2	11,2	0,13	5,58	1,61	-0,8	-46,4	0,83	0,09	1,58	0,73	1,32	0,04
6,5	31,2	14,3	0,08	11,26	2,38	1,2	165,1	1,16	1,83	1,57	0,73	1,22	0,05
6,7	33,5	17,2	0,06	9,14	2,14	1,5	314,8	0,95	0,31	1,45	0,68	0,81	0,18
8,2	41,8	22,2	0,05	2,87	0,83	1,1	141,1	1,53	22,9	1,45	0,69	1,47	0,02
7,4	34,9	34,7	0,03	12,01	2,42	-0,1	-0,1	1,09	1,13	1,02	0,51	0,96	0,11
10,7	46,9	11,2	0,1	6,86	1,81	0,4	5,8	0,48	0	1,18	0,58	0,77	0,2
14,3	48,8	3,7	0,37	7	1,89	0,8	60,5	1,1	1,15	0,92	0,46	0,93	0,12
8,1	40,3	33,8	0,03	2,86	0,99	-0,1	-0,3	1,15	1,73	0,96	0,48	0,88	0,14
8,1	40,3	10,9	0,12	7,59	1,97	1,2	175	1,04	0,73	1,78	0,79	0,96	0,11
4,6	21,3	13,1	0,08	10,47	2,25	2,6	1845,6	1,23	3,15	0,55	0,2	0,73	0,23
8,4	42,2	18,1	0,07	6,78	1,77	1,6	421,1	0,75	0,04	1,21	0,6	1,56	0,02
9,6	44,9	20,1	0,06	2,57	0,69	1,2	192,9	1,02	0,62	0,86	0,43	1,25	0,05
8,8	43,9	23,4	0,05	7,59	1,96	0,7	42,4	1,11	1,3	0,61	0,26	1,27	0,04
9,1	44,3	30,1	0,04	9,59	2,17	-0,01	-0,01	1,14	1,63	1,09	0,54	0,68	0,26
9,6	44,8	27	0,04	11,08	2,35	0,02	0,02	1,35	7,22	0,86	0,43	1,12	0,07
7,4	36,2	25,4	0,04	12,33	2,43	0,2	1,1	1,11	1,31	1,12	0,56	1,53	0,02
8,5	43,9	25,6	0,04	11,76	2,42	1,6	443,8	1,78	88,7	0,53	0,19	0,72	0,23
11,2	47,1	30,1	0,04	9,49	2,2	0,4	8,4	0,79	0,06	1,16	0,58	1,13	0,07
12,7	47,9	27,7	0,04	12,2	2,48	-0,3	-2,8	0,84	0,1	1,05	0,52	1,12	0,07
13,4	48,2	26	0,05	9,62	2,22	-0,5	-13,1	0,83	0,1	0,57	0,22	0,94	0,12
13,8	48,7	23,2	0,05	12,53	2,48	0,8	50	1,21	2,76	0,9	0,45	1,32	0,04
7,2	34,1	19	0,06	3,51	1,06	0,3	4,3	1,2	2,61	0,75	0,36	0,95	0,12
8	40,3	32,1	0,03	4,76	1,46	0,6	19,8	1,31	5,75	0,86	0,43	1,2	0,06
		16,8	0,07	10,41	2,3	0,2	1,1	1,67	49,21	0,77	0,37	0,72	0,23
		12,1	0,09	8,71	2,09	0,6	23,2	1,18	2,29	1,61	0,74	0,69	0,25
		9,3	0,17	8,7	2,12	1,4	255,9	0,65	0,01	0,68	0,31	0,96	0,11
		6,7	0,26	9,68	2,17	0,8	44,8	0,89	0,17	1,01	0,51	0,9	0,14
		21,5	0,05	10,04	2,24	1,5	326,4	0,68	0,01	1,36	0,65	1,67	0,01

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

4 Рокицкий, П. Ф. Биологическая статистика / П. Ф. Рокицкий. – Минск : «Вышэйшая школа», 1973. – 320 с.

ТЕМА 6. ДИСПЕРСИОННЫЙ АНАЛИЗ В EXCEL И STATISTICA 7.0

6.1 Однофакторный дисперсионный анализ в Excel и STATISTICA 7.0.
6.2 Двухфакторный дисперсионный анализ в Excel и STATISTICA 7.0.
6.3 Непараметрический дисперсионный анализ в STATISTICA 7.0.

6.1.1 Однофакторный дисперсионный анализ в Excel и STATISTICA 7.0

1.1 Однофакторный дисперсионный анализ в Excel

Дисперсионный анализ (лат. «*dispersio*» – рассеивание; в английском варианте – *analysis of variance* (ANOVA) – анализ переменных) применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну зависимую количественную переменную.

В основе дисперсионного анализа лежит предположение о том, что одни переменные могут рассматриваться как причины (факторы, независимые переменные), а другие – как следствия (зависимые переменные).

Цель дисперсионного анализа – исследование значимости различия между средними с помощью сравнения (анализа) дисперсий. Необходимо иметь в виду, что в случае простого сравнения средних двух выборок, дисперсионный анализ даст тот же результат, что и обычный *t*-тест.

По количеству выявляемых регулируемых факторов дисперсионный анализ может быть:

– *однофакторным* – изучается влияние одного фактора на результаты эксперимента;

– *двухфакторным* – изучается влияние двух факторов;

– *многофакторным* – позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие.

Дисперсионный анализ относится к группе параметрических методов, и поэтому его следует применять только тогда, когда доказано, что распределение подчиняется закону нормального распределения.

Для рассмотрения работы однофакторного дисперсионного анализа воспользуемся данными по оценке влияния на урожайность ржи внесения различных смесей удобрений (в т/га) (таблица 6.1).

Таблица 6.1 – Урожайность ржи после внесение различных смесей удобрений

в т/га

Повторности	Удобрения			
	Контроль	Смесь 1	Смесь 2	Смесь 3
1	24,2	26,6	27	32,2
2	31	25,6	33	31
3	34,2	31	32,2	30

Шаг 1. Создание электронной таблицы с данными.

При создании таблицы с данными необходимо каждый из признаков (другими словами – отдельную переменную) разместить в отдельном столбце. То есть, в конечном итоге получим 4 столбца с данными (рисунок 6.1).

Контроль	Смесь 1	Смесь 2	Смесь 3
24,2	26,6	27	32,2
31	25,6	33	31
34,2	31	32,2	30

Рисунок 6.1 – Создание файла данных в книге Excel

Шаг 2. Выбор анализа.

Для проведения регрессионного анализа необходимо перейти в пункт главного меню **Данные,** а затем открыть модуль **Анализ данных**, выбрать там из списка опцию **Однофакторный дисперсионный анализ** (рисунок 6.2), после чего щелкнуть мышкой **ОК**.

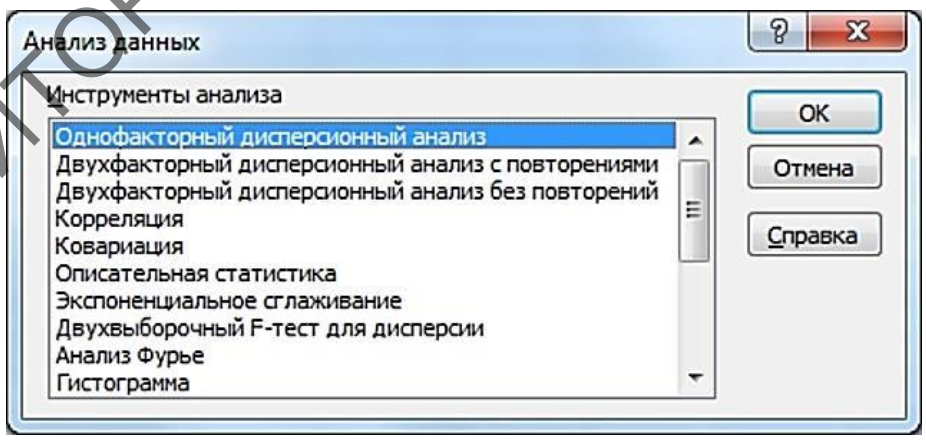


Рисунок 6.2 – Выбор опции **Однофакторный дисперсионный анализ** в диалоговом окне **Анализ данных** в Excel

Шаг 3. Выставление параметров.

Появившееся диалоговое окно (рисунок 6.3) имеет следующие элементы:

1) группа **Входные данные**:

– **Входной интервал** – *здесь указывается диапазон адресов ячеек с данными* (*можно и с заголовками*);

– **Группирование (по столбцам / по строкам)** – *избранный пользователем способ группировки признаков переменных*;

– **Метки в первой строке** – *указание программе на наличие заголовков в названии переменных*;

– **Альфа** – *указание уровня значимости p*;

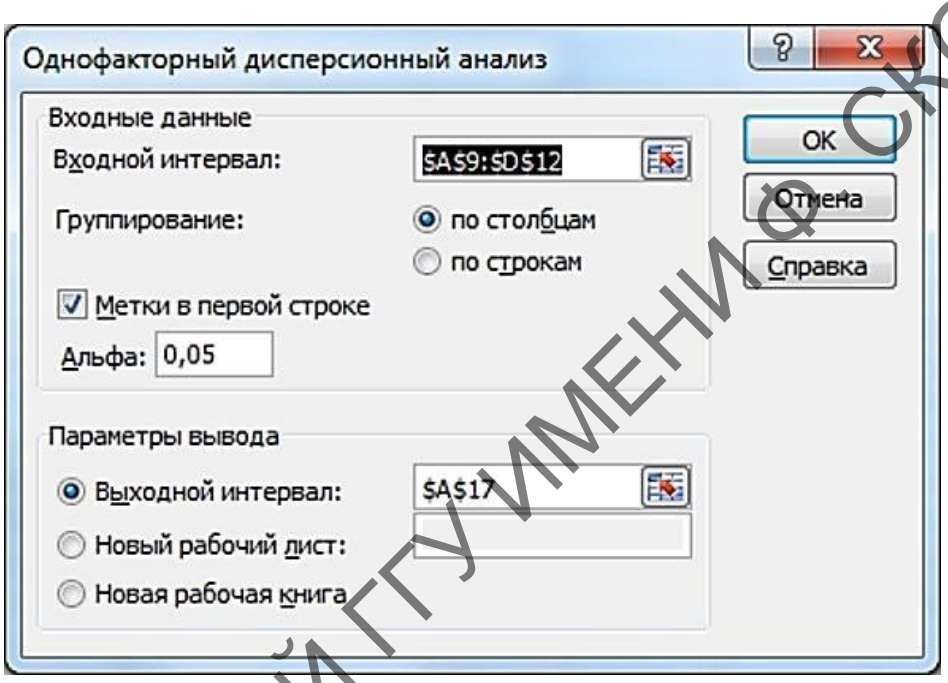


Рисунок 6.3 – Выставление параметров в диалоговом окне **Однофакторный дисперсионный анализ** в Excel

2) группа **Параметры вывода**:

– **Выходной интервал** – *адрес ячейки текущей книги* Excel, где будет отображён результат анализа;

– **Новый рабочий лист** – *вывод результатов анализа на указанный новый лист рабочей книги* Excel;

– **Новая рабочая книга** – *вывод результатов анализа в новый файл* Excel.

В качестве входного интервала необходимо указать весь спектр данных, включая заголовки, в связи с чем необходимо отметить поле **Метка** (рисунок 6.3), в качестве выходного интервала нужно указать любой адрес ячейки, например $A$17 (рисунок 6.3), уровень

значимости поставить 0,05 (95 % точности) и нажать кнопку **ОК**. Результат обработки появится в указанном поле (рисунок 6.4).

17	Однофакторный дисперсионный анализ						
18							
19	ИТОГИ						
20	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
21	Контроль	3	89,4	29,8	26,08		
22	Смесь 1	3	83,2	27,7333333	8,253333333		
23	Смесь 2	3	92,2	30,7333333	10,61333333		
24	Смесь 3	3	93,2	31,0666667	1,213333333		
25							
26							
27	Дисперсионный анализ						
28	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
29	Между группами	20,226667	3	6,74222222	0,584248026	0,641961164	4,066180551
30	Внутри групп	92,32	8	11,54			
31							
32	Итого	112,54667	11				

Рисунок 6.4 – Итоги однофакторного дисперсионного анализа в Excel

Результаты дисперсионного анализа будут состоять из 2 таблиц. В первой таблице для каждого столбца исходной таблицы, в которых располагаются анализируемые группы, приведены числовые параметры: количество чисел (счет), суммы по столбцам, средние и дисперсии по столбцам.

Во второй части результатов отражены собственно результаты дисперсионного анализа и используются следующие обозначения (рисунок 6.4):

– **SS** – *сумма квадратов*;

– **df** – *число степеней свободы*;

– **MS** – *средний квадрат* (*дисперсия*);

– **F** – *эмпирическое значение критерия Фишера* (*фактическое значение*);

– **p-значение** – *уровень значимости результатов дисперсионного анализа данных, расположенных по столбцам*;

– **F-критическое** – *табличное значение критерия Фишера при заданном ранее уровне значимости* (в данном случае $p = 0,05$).

Таким образом, сумма квадратов, обусловленная влиянием исследуемого фактора (межгрупповая сумма), равна 20,23. Остаточная сумма квадратов (внутригрупповая) равна 92,32. Соответствующие дисперсии: межгрупповая (для исследуемого фактора) – 6,74, остаточная, внутригрупповая – 11,54.

В результате проведённого анализа можно сказать, что, так как эмпирическое значение критерия Фишера меньше табличного (т. е. кри-

тического), то нет оснований отвергать нулевую гипотезу о том, что данные удобрения не влияют на урожайность ржи.

6.1.2 Однофакторный дисперсионный анализ в STATISTICA 7.0

Шаг 1. Создание электронной таблицы с данными.

Для проведения анализа необходимо предварительно составить таблицу с данными. Следует обратить внимание на то, что в среде STATISTICA 7.0 таблица с данными для дисперсионного анализа составляется по иному принципу, нежели в Excel. Для однофакторного дисперсионного анализа используется 2 переменные – одна из них является описанием фактора, а вторая – зависимая, т. е. показывает значения фактора (рисунок 6.5).

Data: Spreadsheet1* (2v b...

	1 Фактор	2 Урожайность
1	Контроль	24,2
2	Контроль	31
3	Контроль	34,2
4	Смесь 1	26,6
5	Смесь 1	25,6
6	Смесь 1	31
7	Смесь 2	27
8	Смесь 2	33
9	Смесь 2	32,2
10	Смесь 3	32,2
11	Смесь 3	31
12	Смесь 3	30

Рисунок 6.5 – Электронная таблица данных для расчёта однофакторного дисперсионного анализа. Первая переменная («Фактор») – описание фактора, вторая переменная («Урожайность») – значение фактора

Шаг 2. Выбор анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **ANOVA** (*Анализ переменных*) (рисунок 6.6) и нажать **ОК**.

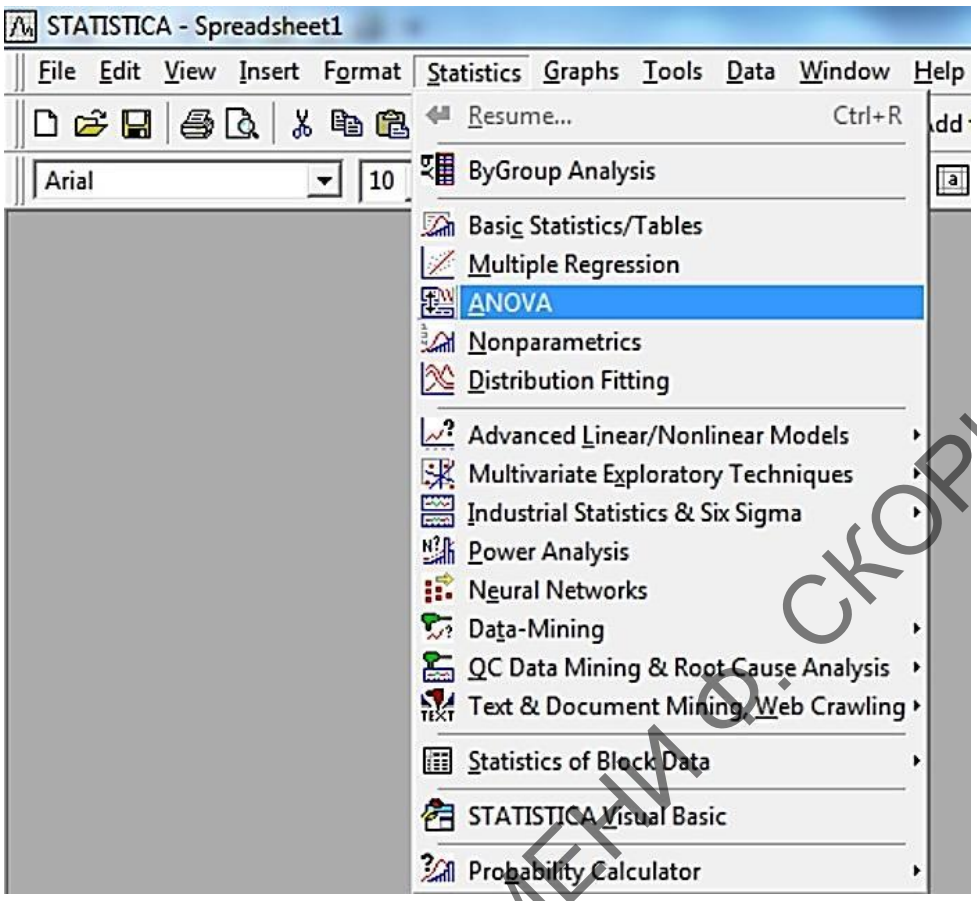


Рисунок 6.6 – Запуск модуля **ANOVA**

Шаг 3. Выбор модуля.

В появившемся на экране диалоговом окне необходимо выбрать опцию **One way ANOVA** (*Однофакторный анализ вариант*) (рисунок 6.7) и нажать **OK**.

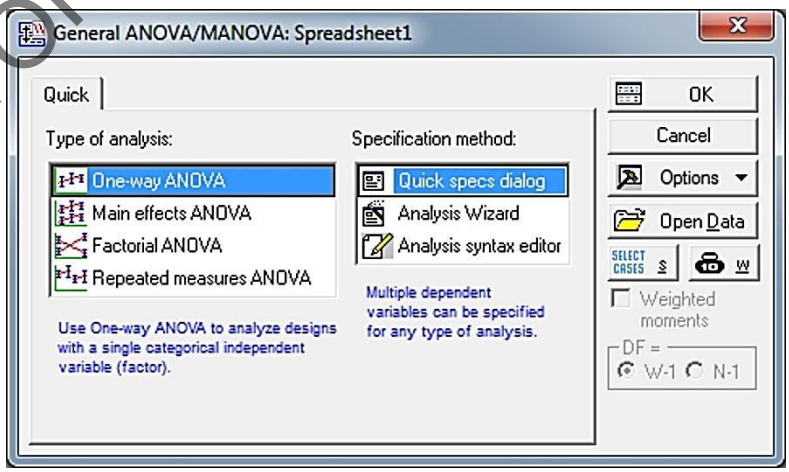


Рисунок 6.7 – Диалоговое окно модуля **ANOVA**

Шаг 4. Выбор переменных.

В появившемся на экране диалоговом окне однофакторного анализа (рисунок 6.8) в закладке **Quick** (*Быстрый анализ*) необходимо нажать кнопку **Variables** (*Переменные*) и в диалоговом окне выбора переменных указать слева зависимую переменную, а справа – категориальное описание, т. е. переменную, описываемую как фактор (рисунок 6.9) и далее щелкнуть левой клавишей мыши на **ОК**.

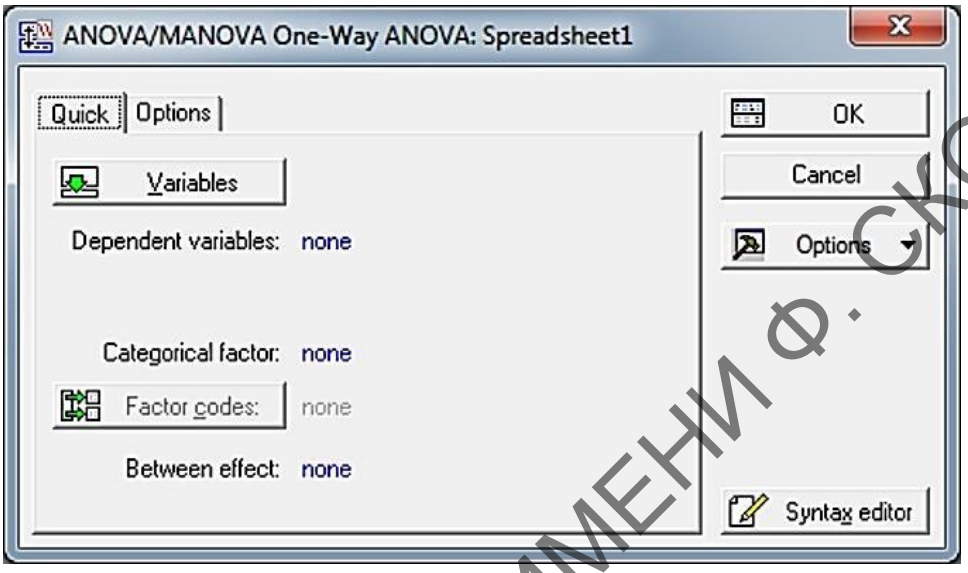


Рисунок 6.8 – Диалоговое окно модуля **One way ANOVA**

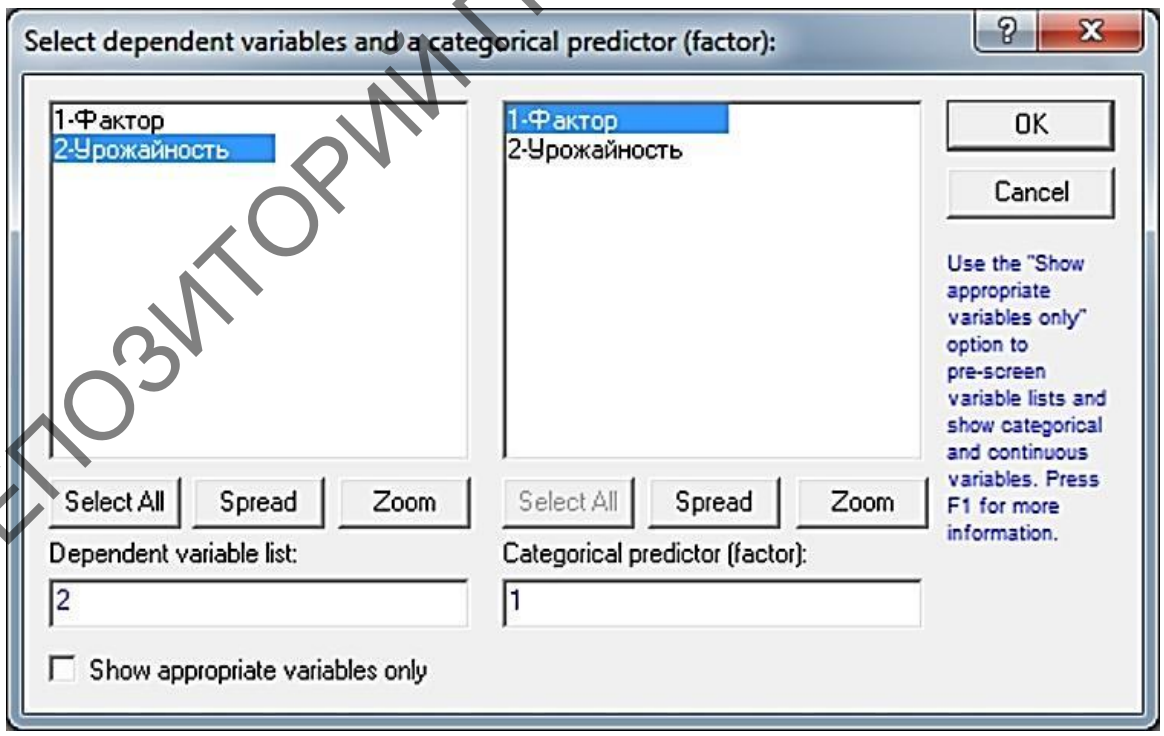


Рисунок 6.9 – Выбор переменных для анализа

Шаг 5. Выбор фактора.

После выбора переменных и нажатия кнопки **ОК** программа возвращается к первоначальному окну модуля **One way ANOVA** (*Однофакторный анализ вариант*) (рисунок 6.8).

Для определения нужных нам факторов для расчёта в этом же окне необходимо нажать кнопку **Categorical factor** (*Определяемый фактор*) и в появившемся окне (рисунок 6.10) нужно указать название того фактора, перечисленного в первой переменной, влияние которого мы пытаемся учитывать.

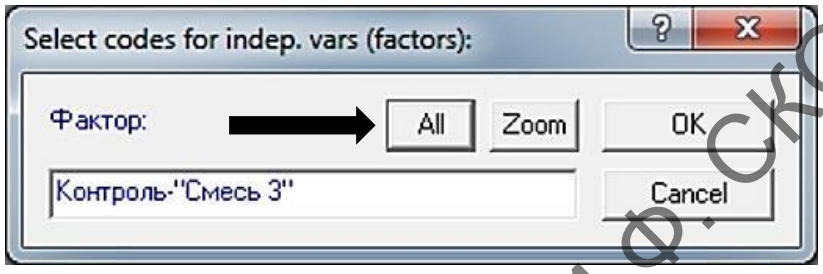


Рисунок 6.10 – Выбор фактора

В нашем случае учитывается влияние всех факторов, поэтому нужно нажать кнопку **All** (*Все*) и далее щелкнуть левой клавишей мыши на **ОК**. Программа вернётся в окно модуля **One way ANOVA** (*Однофакторный анализ вариант*), которое принимает окончательный вид перед проведением анализа (рисунок 6.11). После необходимо кликнуть левой клавишей мыши на **ОК**.

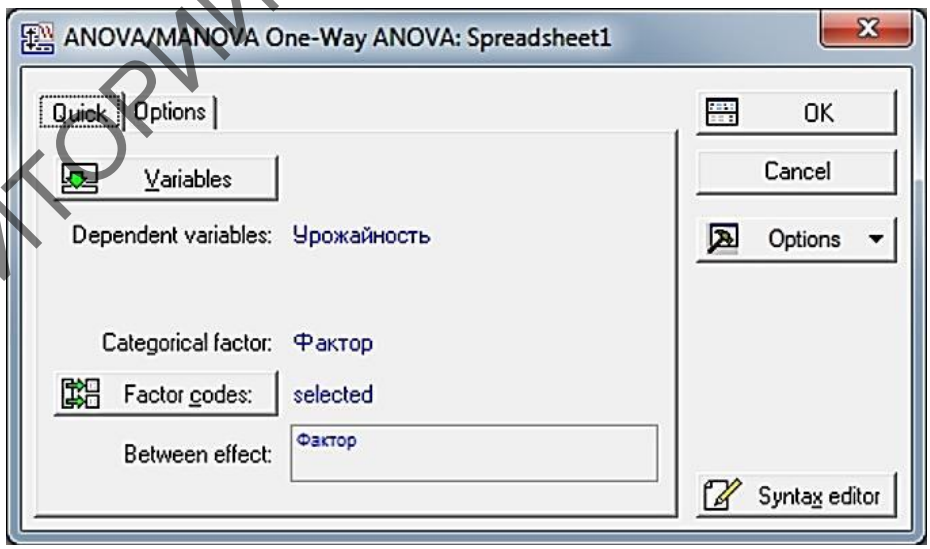


Рисунок 6.11 – Диалоговое окно модуля **One way ANOVA**, подготовленное для анализа

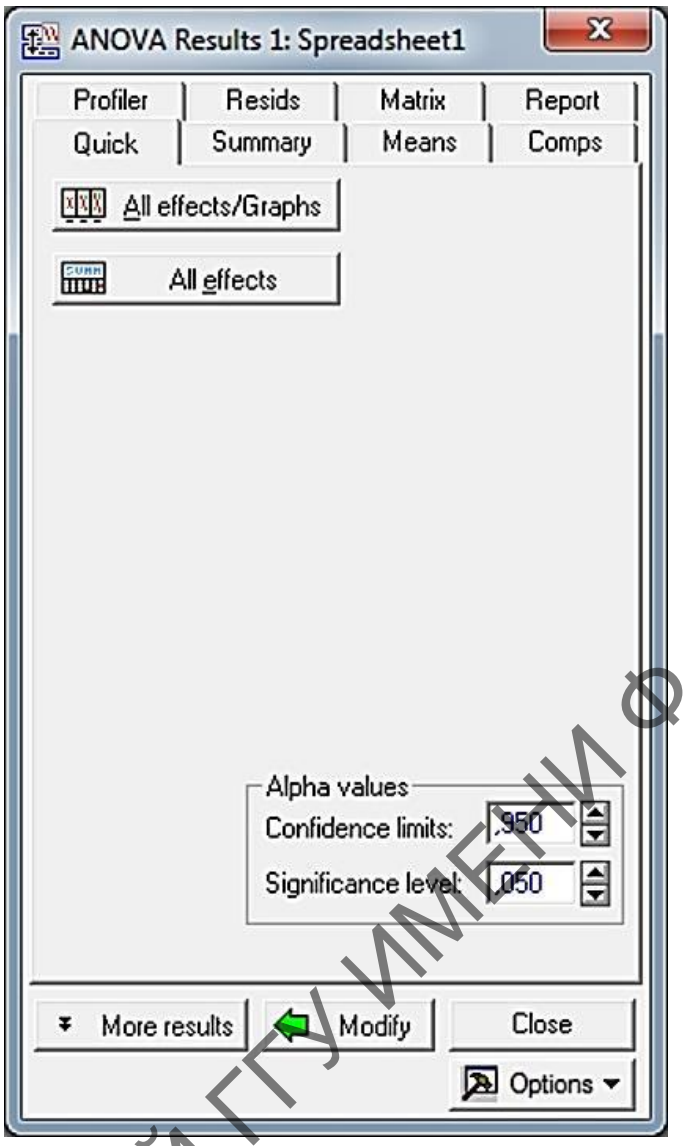


Рисунок 6.12 – Диалоговое окно модуля результатов однофакторного **ANOVA**

Шаг 6. Просмотр результатов.

После нажатия на кнопки **ОК** на предыдущем шаге выполнения анализа на экране появляется диалоговое окно результатов (рисунок 6.12).

Для простого просмотра результатов анализа достаточно нажать кнопку **All effects** (*Все эффекты*). Результаты будут отображены в специальной таблице (рисунок 6.13). Она имеет следующие столбцы:

– **SS** – *сумма квадратов*;

– **Degr. of Freedom** – *степени свободы*;

– **MS** – *средний квадрат* (*дисперсия*);

– **F** – *значение критерия Фишера*;

– **p** – *уровень значимости результатов дисперсионного анализа данных*.

Effect	Univariate Tests of Significance for Урожайность (урожайность) Sigma-restricted parameterization Effective hypothesis decomposition SS	Degr. of Freedom	MS	F	p
Intercept	10680,33	1	10680,33	925,5055	0,000000
Фактор	20,23	3	6,74	0,5842	0,641961
Error	92,32	8	11,54		

Рисунок 6.13 – Таблица результатов однофакторного **ANOVA**

В рассмотренном примере F-критерий показывает, что различие между средними (строка «Фактор») статистически незначимо (значимо на уровне 0,64, то есть больше, чем критическое значение 0,05). Поскольку различие между средними значениями незначимо, нулевая гипотеза не отвергается (удобрения не влияют на урожайность ржи).

Шаг 7. Дополнительные результаты.

Для просмотра дополнительных итогов необходимо в диалоговом окне модуля итогов анализа (рисунок 6.12) нажать кнопку **More results** (*Больше результатов*) и перейти в расширенное окно результатов анализа (рисунок 6.14).

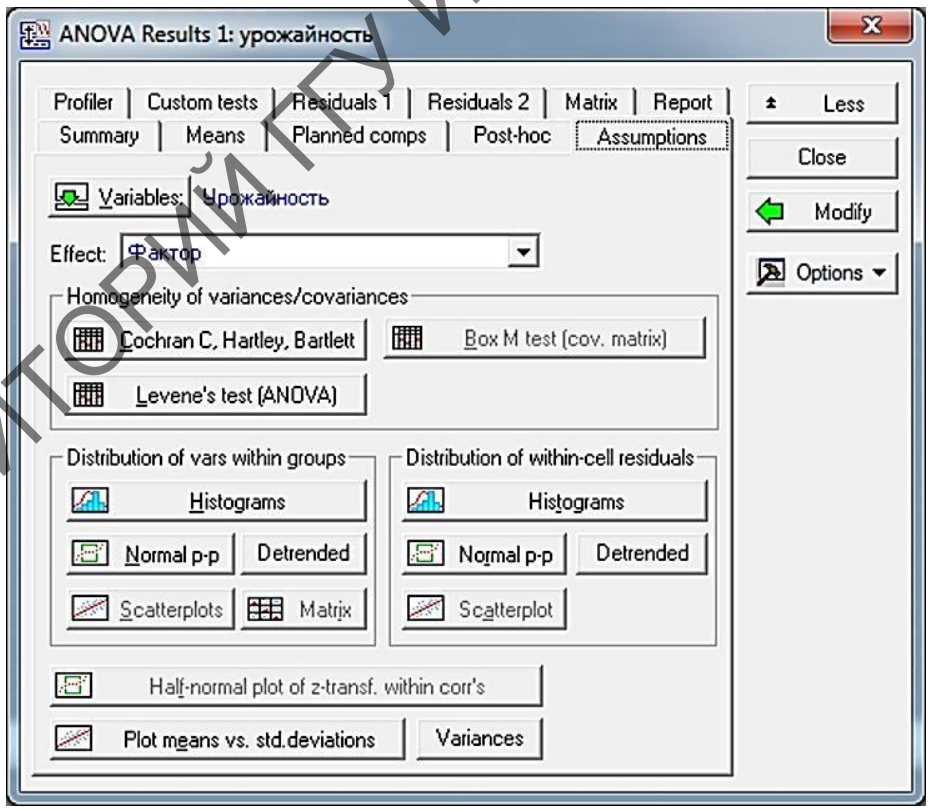


Рисунок 6.14 – Таблица результатов однофакторного **ANOVA**

Сначала необходимо проверить, насколько данный анализ соответствует критериям нормальности распределения и однородности дисперсий. Для этого следует перейти на закладку **Assumptions** (*Допущения*) (рисунок 6.13).

Шаг 8. Проверка однородности групповых дисперсий.

Для данной проверки необходимо в поле **Homogeneity of variances/covariances** (*Однородность вариант/ковариант*) диалогового окна нажать на кнопку **Levene's test (ANOVA)** (*тест Левена* (*дисперсионный анализ*)). В итоге появится таблица с результатами (рисунок 6.15).

Если результат этого теста указывает на отсутствие различий между дисперсиями ($p > 0,05$), то применение параметрического варианта дисперсионного анализа является обоснованным.

Levene's Test for Homogeneity of Variances (урожайность)
Effect: Фактор
Degrees of freedom for all F's: 3, 8

	MS Effect	MS Error	F	p
Урожайность	4,489877	1,997778	2,247435	0,160068

Рисунок 6.15 – Таблица результатов теста Левена

В нашем случае различий нет ($p = 0,16$).

Шаг 9. Проверка нормальности распределения.

Для проведения данной операции необходимо обратить внимание на поле **Distribution of variables within groups** (*Распределение переменных внутри групп*) (рисунок 6.14). Метод, которым необходимо будет воспользоваться, зависит от объёма сравниваемых выборок:

1) Объём выборок менее 30 значений. В таких случаях чаще (и лучше) использовать график нормальных вероятностей (кнопка **Normal p-p**). После этого программа спросит, какие группы нужно проверить. Необходимо выбрать **All Groups** (*Все группы*) и нажать **ОК.** Это подходит и к нашему случаю. В итоге график имеет вид, отображённый на рисунке 6.16.

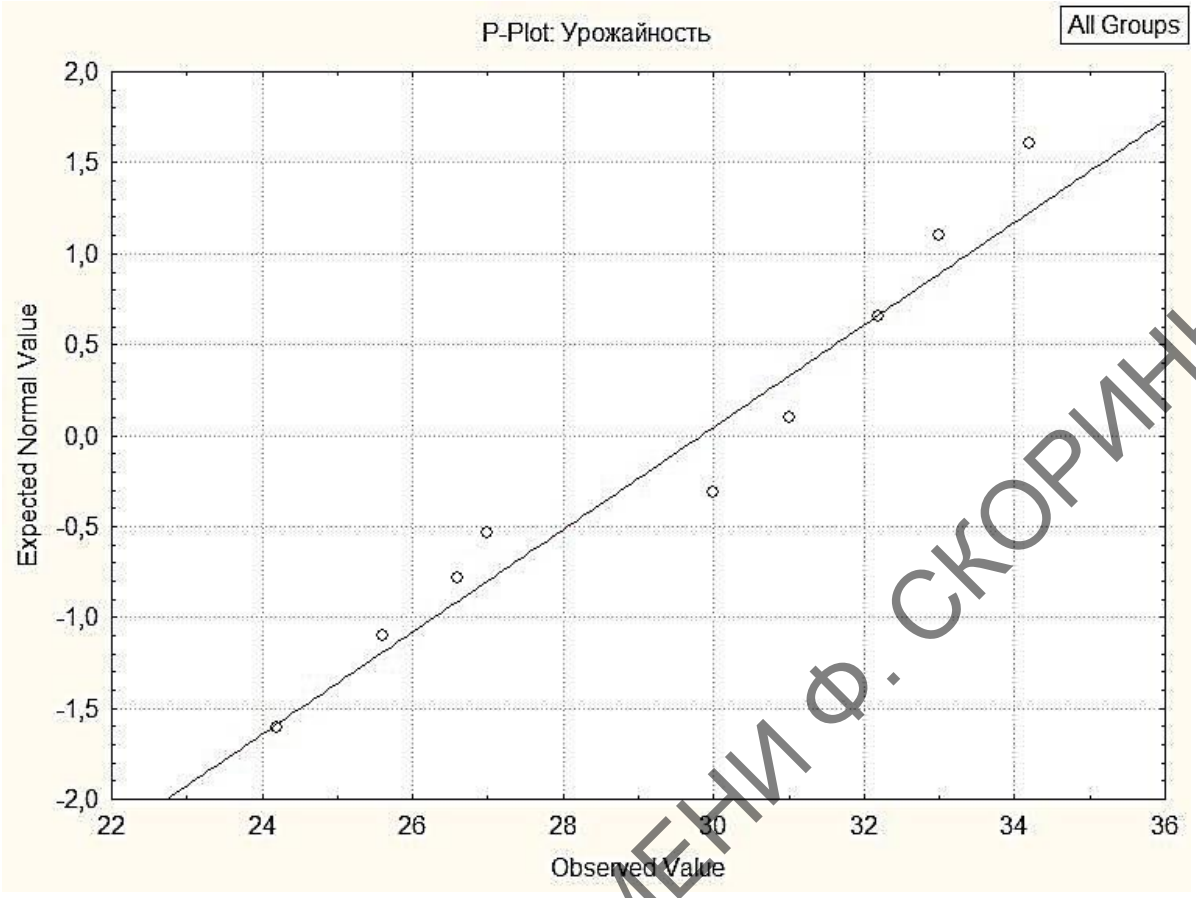


Рисунок 6.16 – График нормальных вероятностей

Так как точки наблюдения укладываются вдоль теоретически ожидаемой прямой, можно говорить о том, что подтверждает нормальность распределения.

2) Объём выборки больше 30 значений. При данной ситуации строят гистограмму (кнопка **Histograms**). После чего программа спросит, какие группы нужно проверить. Необходимо выбрать **All Groups** (*Все группы*) и нажать **ОК.**

Шаг 10. Графическое отображение результатов.

Для графического отображения полученных результатов анализа можно поступить двояко. Во-первых, воспользоваться возможностями самого модуля. Для этого в итоговом окне расширенных результатов (рисунок 6.14) необходимо перейти на закладку **Summary** (*Итоги*) и нажать кнопку **All effects**/**Grafs** (*Все эффекты/Графики*), затем в появившемся окне нужно нажать **OK**. Результат представлен на рисунке 6.17.

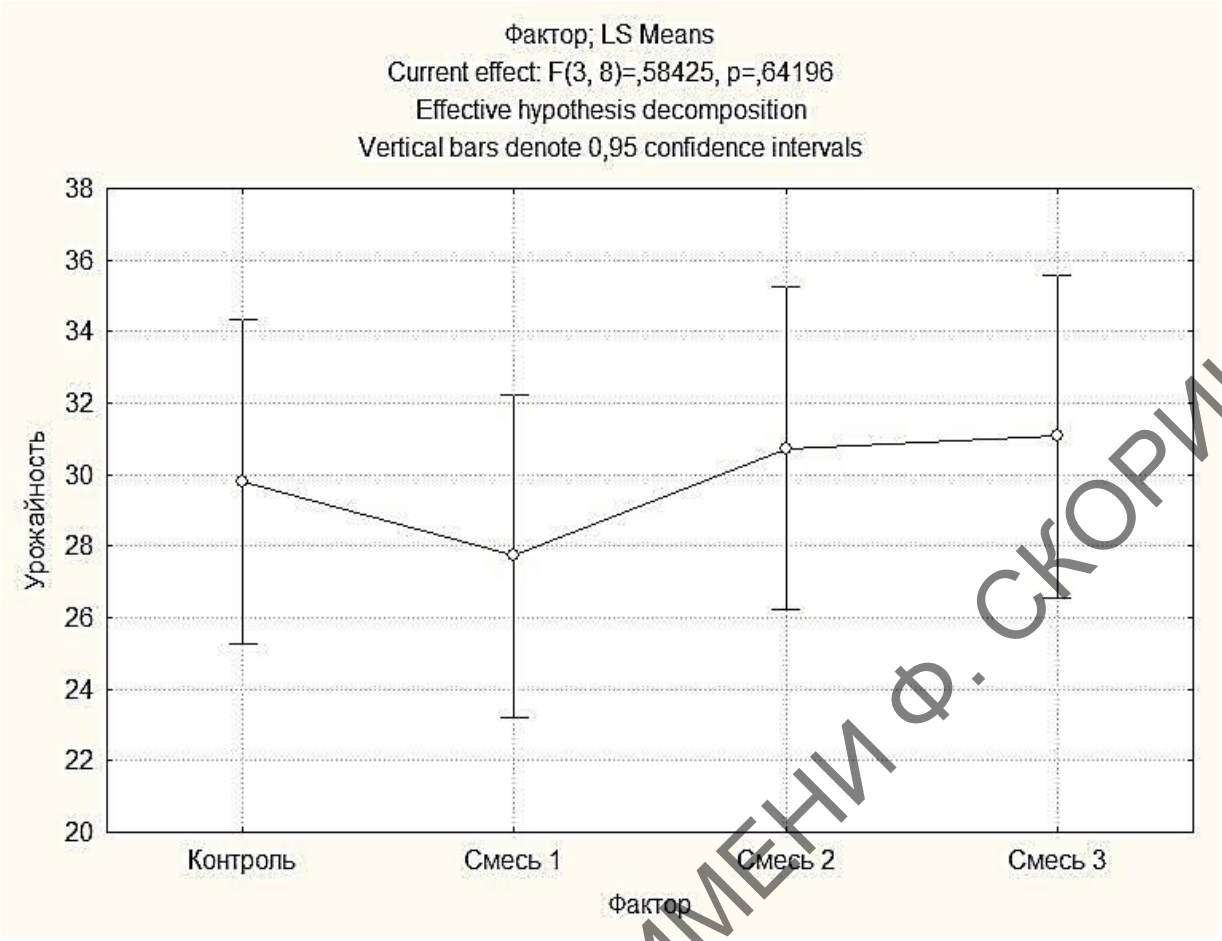


Рисунок 6.17 – График результатов дисперсионного анализа

Второй способ связан с возможностью раздела меню **Graphs** (*Графики*). Для этого необходимо в главном меню выбрать раздел **Graphs** (*Графики*) и опцию **2D Graphs** (*Двухмерные графики*), затем – **Box plots** (*Графики в виде ящиков*) (см. Тема 1, пункт 1.2.3). В появившемся диалоговом окне нужно перейти на закладку **Advanced** (*Расширенные настройки*), указать программе необходимые переменные, нажав кнопку **Variables** (*Переменные*), и выполнить установки, как это показано на рисунке 6.18.

После выставления указанных параметров необходимо нажать кнопку **ОК**. Итоговый график изображён на рисунке 6.19. В итоге мы получаем визуализацию статистических параметров при воздействии факторов, для каждого из которых показаны:

– **Mean** – *среднее*;
– **SD** – *стандартное отклонение*;
– **SE** – *стандартная ошибка*.

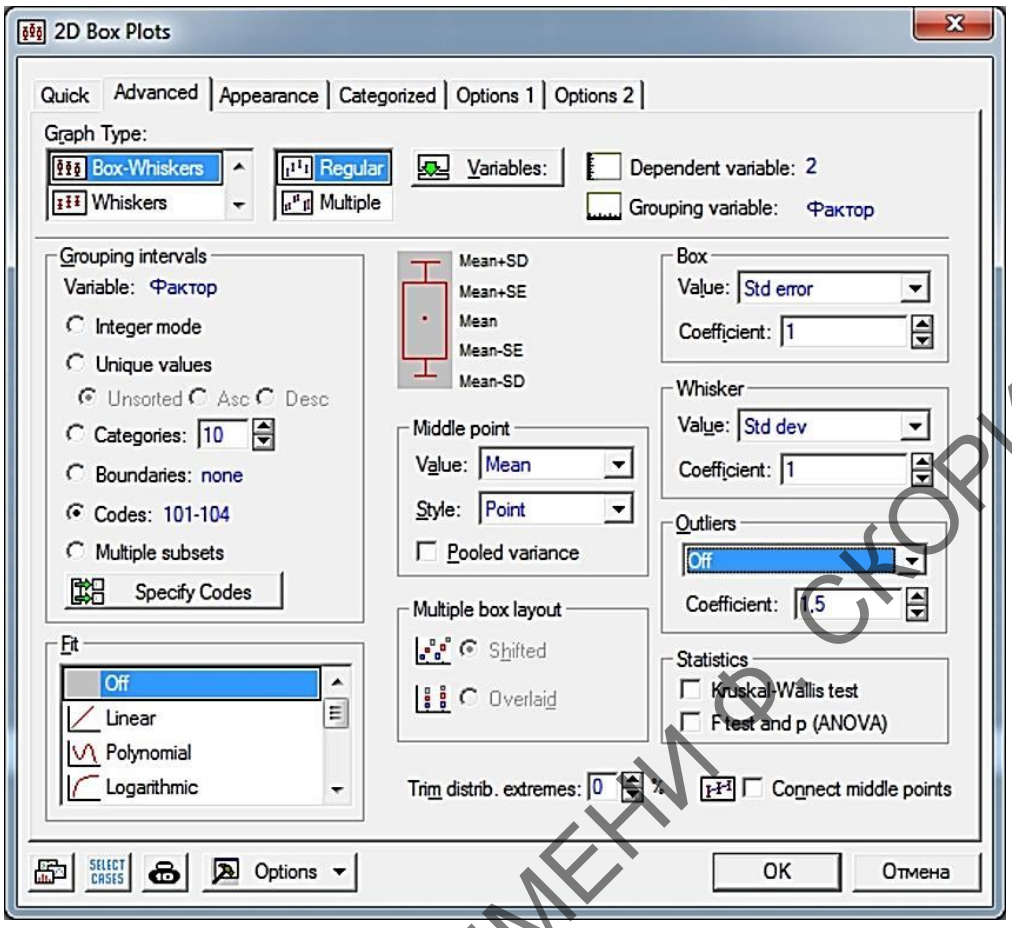


Рисунок 6.18 – Диалоговое окно **2D Box plots**

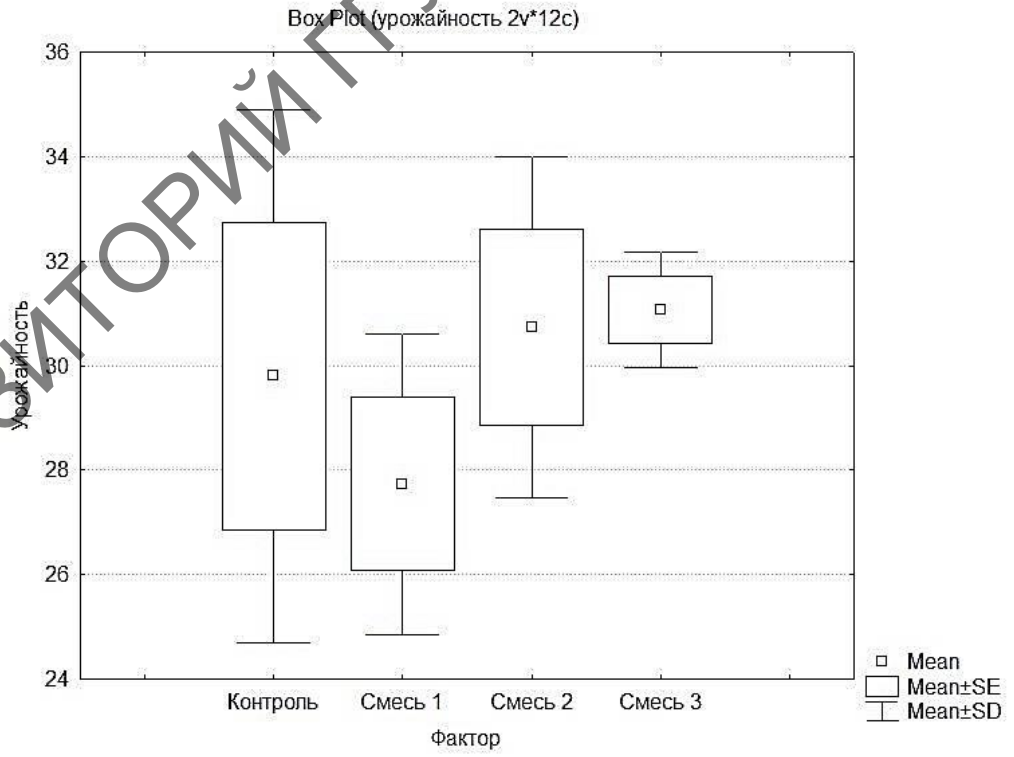


Рисунок 6.19 – График диапазонов **2D Box plots**

Результаты, полученные в пакете Statistica 7.0, естественно, точно такие же, как и в электронных таблицах MS Excel. Поэтому вывод об воздействии удобрений на урожайность аналогичен. Однако в пакете Statistica 7.0 имеется возможность визуализации результатов, что, несомненно, его делает более привлекательным при представлении результатов исследований.

6.2. Двухфакторный дисперсионный анализ в Excel и STATISTICA 7.0

6.2.1 Двухфакторный дисперсионный анализ в Excel

Двухфакторный дисперсионный анализ предполагает наличие влияния на признак не одного фактора, а двух. В качестве примера рассмотрим уже известную нам по первому разделу урожайность ржи, но учтём тот факт, что повторности – это разные сорта этого растения (таблица 6.2), в связи с чем будем оценивать влияние на урожайность не только вида удобрения, но и сорта.

Таблица 6.2 – Урожайность ржи после внесение различных смесей удобрений

в т/га

Сорт	Удобрения			
	Контроль	Смесь 1	Смесь 2	Смесь 3
Рожь 01	24,2	26,6	27	32,2
Рожь 02	31	25,6	33	31
Рожь 03	34,2	31	32,2	30

Шаг 1. Создание электронной таблицы с данными.

При создании таблицы с данными необходимо каждый из признаков (другими словами – отдельную переменную) разместить в отдельном столбце. Первый фактор имеет 3 уровня (3 строки), второй фактор имеет 4 уровня (4 столбца) (рисунок 6.20).

	A	B	C	D	E
1		Удобрения			
2	Сорт	Контроль	Смесь 1	Смесь 2	Смесь 3
3	Рожь 01	24,2	26,6	27	32,2
4	Рожь 02	31	25,6	33	31
5	Рожь 03	34,2	31	32,2	30

Рисунок 6.20 – Создание файла данных в книге Excel

Шаг 2. Выбор анализа.

Для проведения регрессионного анализа необходимо перейти в пункт главного меню **Данные,** а затем открыть модуль **Анализ данных**, выбрать там из списка опцию **Двухфакторный дисперсионный анализ без повторений** (рисунок 6.21), после чего щелкнуть мышкой **ОК**.

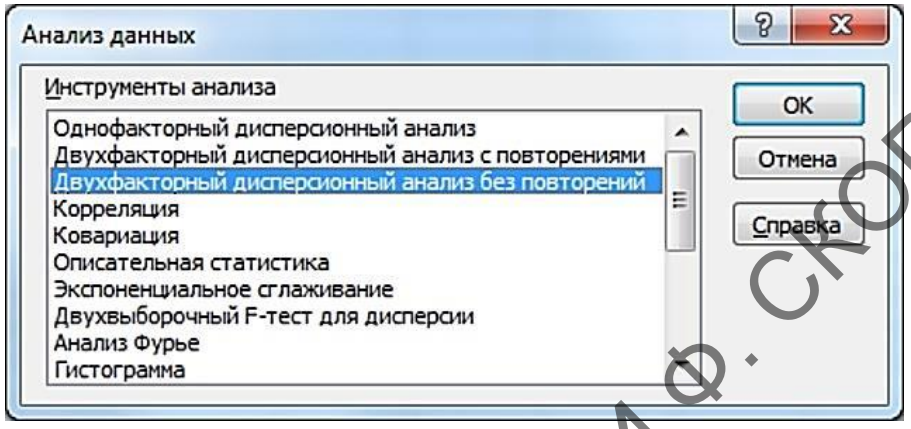


Рисунок 6.21 – Выбор опции **Двухфакторный дисперсионный анализ без повторений** в диалоговом окне **Анализ данных** в Excel

Шаг 3. Выставление параметров.

Появившееся диалоговое окно (рисунок 6.22) имеет элементы, идентичные рассмотренным в пункте 6.1.1, в связи с чем действия, которые необходимо будет провести, также идентичны (пример представлен на рисунке 6.22).

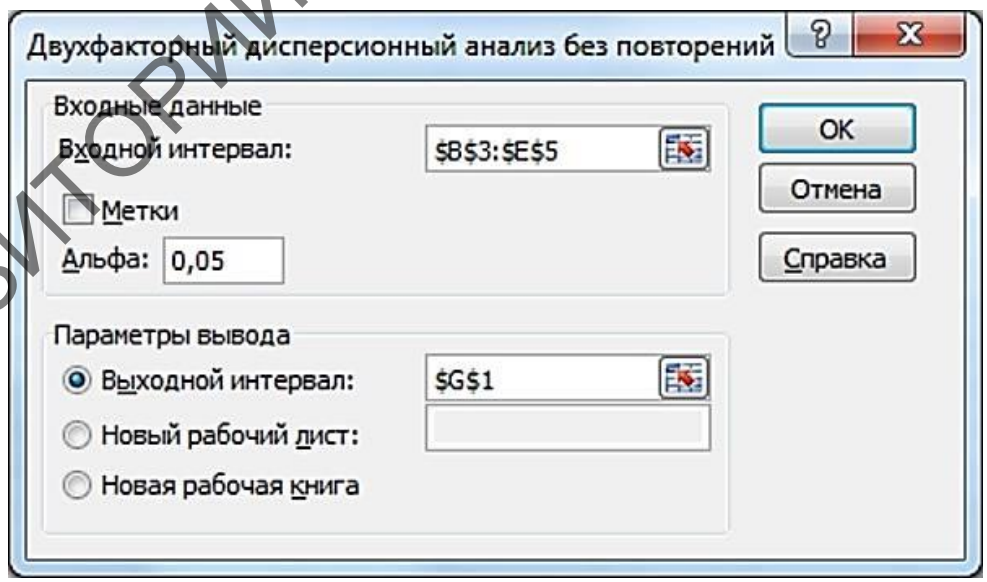


Рисунок 6.22 – Выставление параметров в диалоговом окне **Двухфакторный дисперсионный анализ без повторений** в Excel

Результат обработки появится в указанном поле (рисунок 6.23).

G	H	I	J	K	L	M
Двухфакторный дисперсионный анализ без повторений						
ИТОГИ	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
Строка 1	4	110	27,5	11,34666667		
Строка 2	4	120,6	30,15	10,09		
Строка 3	4	127,4	31,85	3,263333333		
Столбец 1	3	89,4	29,8	26,08		
Столбец 2	3	83,2	27,73333	8,253333333		
Столбец 3	3	92,2	30,73333	10,61333333		
Столбец 4	3	93,2	31,06667	1,213333333		
Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Строки	38,44667	2	19,22333	2,140947902	0,198716604	5,14325285
Столбцы	20,22667	3	6,742222	0,750897166	0,560517731	4,757062663
Погрешность	53,87333	6	8,978889			
Итого	112,5467	11				

Рисунок 6.23 – Итоги двухфакторного дисперсионного анализа без повторений в Excel

Результаты дисперсионного анализа будут состоять из 2 таблиц. В первой таблице для каждой строки и каждого столбца исходной таблицы, в которых располагаются анализируемые группы, приведены числовые параметры: количество чисел (счет), суммы, средние и дисперсии.

Во второй части результатов отражены собственно результаты дисперсионного анализа и используются следующие обозначения (рисунок 6.23):

– **SS** – *сумма квадратов*;

– **df** – *число степеней свободы*;

– **MS** – *средний квадрат* (*дисперсия*);

– **F** – *эмпирическое значение критерия Фишера* (*фактическое значение*);

– **p-значение** – *уровень значимости результатов дисперсионного анализа данных, расположенных по столбцам*;

– **F-критическое** – *табличное значение критерия Фишера при заданном ранее уровне значимости* (в данном случае $p = 0,05$).

Сумма квадратов, обусловленная влиянием первого фактора (сорт ржи), равна 38,45. Сумма квадратов, обусловленная влиянием второго фактора (удобрение), равна 20,23, а остаточная, внутригрупповая сумма квадратов (погрешность) равна 53,87. Остаточная, внутригрупповая дисперсия (погрешность) равна 8,98.

В результате проведённого анализа можно сказать, что, так как в обоих случаях эмпирическое значение критерия Фишера меньше табличного (т. е. критического) при уровне значимости значительно

больше 0,05, то нет оснований отвергать нулевую гипотезу о том, что данные удобрения и сорт культуры не влияют на урожайность ржи.

6.2.2 Двухфакторный дисперсионный анализ в STATISTICA 7.0

Шаг 1. Создание электронной таблицы с данными.

Для проведения анализа необходимо предварительно составить таблицу с данными. В связи с тем, что анализ у нас двухфакторный, то нужно добавить вторую группирующую переменную (с примером об урожайности ржи – это сорт растения). Для расчёта в STATISTICA 7.0 данных, представленных ранее в таблице, недостаточно для параметрического расчёта критерия Фишера, поэтому следует увеличить выборку. Можно составить новую таблицу данных, а можно использовать предыдущую, добавив в неё новую переменную и значения переменных. О том, как это делается в STATISTICA 7.0, мы рассматривали при изучении темы 1. Готовая таблица изучения влияния вносимого удобрения и сорта растения на урожайность ржи представлена на рисунке 6.24.

Data: урожайность_2ф (3v by 24c)

	1 Фактор	2 Сорт	3 Урожайность
1	Контроль	Рожь 01	24,2
2	Контроль	Рожь 02	31
3	Контроль	Рожь 03	34,2
4	Контроль	Рожь 01	22,2
5	Контроль	Рожь 02	31,3
6	Контроль	Рожь 03	34,4
7	Смесь 1	Рожь 01	25,1
8	Смесь 1	Рожь 02	26,4
9	Смесь 1	Рожь 03	29,8
10	Смесь 1	Рожь 01	26,6
11	Смесь 1	Рожь 02	25,6
12	Смесь 1	Рожь 03	31
13	Смесь 2	Рожь 01	27
14	Смесь 2	Рожь 02	33
15	Смесь 2	Рожь 03	32,2
16	Смесь 2	Рожь 01	32,3
17	Смесь 2	Рожь 02	33,1
18	Смесь 2	Рожь 03	31,9
19	Смесь 3	Рожь 01	31
20	Смесь 3	Рожь 02	32
21	Смесь 3	Рожь 03	30
22	Смесь 3	Рожь 01	32,2
23	Смесь 3	Рожь 02	31
24	Смесь 3	Рожь 03	30

Рисунок 6.24 – Электронная таблица данных для расчёта двухфакторного дисперсионного анализа. Первая переменная («Фактор») – описание фактора, вторая переменная («Сорт») – описание сорта ржи, третья переменная («Урожайность») – значение фактора

Шаг 2. Выбор модуля.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **ANOVA** (*Анализ переменных*) (рисунок 6.6) и нажать **ОК**. В появившемся диалоговом окне выбрать модуль **Factorial ANOVA** (*Факторный дисперсионный анализ*) (рисунок 6.25) и нажать **ОК**.

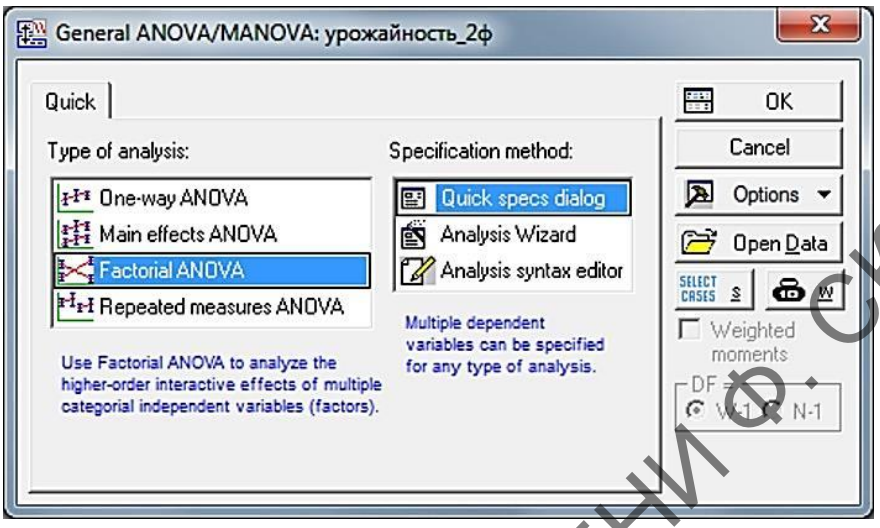


Рисунок 6.25 – Диалоговое окно модуля **ANOVA**

Шаг 3. Выбор переменных.

В появившемся на экране диалоговом окне двухфакторного анализа (рисунок 6.26) в закладке **Quick** (*Быстрый анализ*) необходимо нажать кнопку **Variables** (*Переменные*) и в диалоговом окне выбора переменных указать слева зависимую переменную («Урожайность»), а справа – группирующие переменные («Фактор» и «Сорт») (рисунок 6.27).

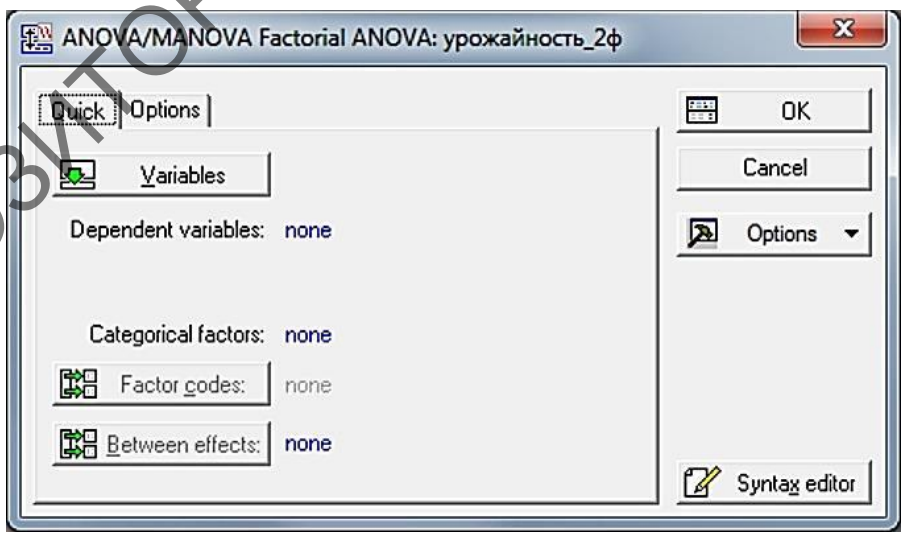


Рисунок 6.26 – Диалоговое окно модуля **Factorial ANOVA**

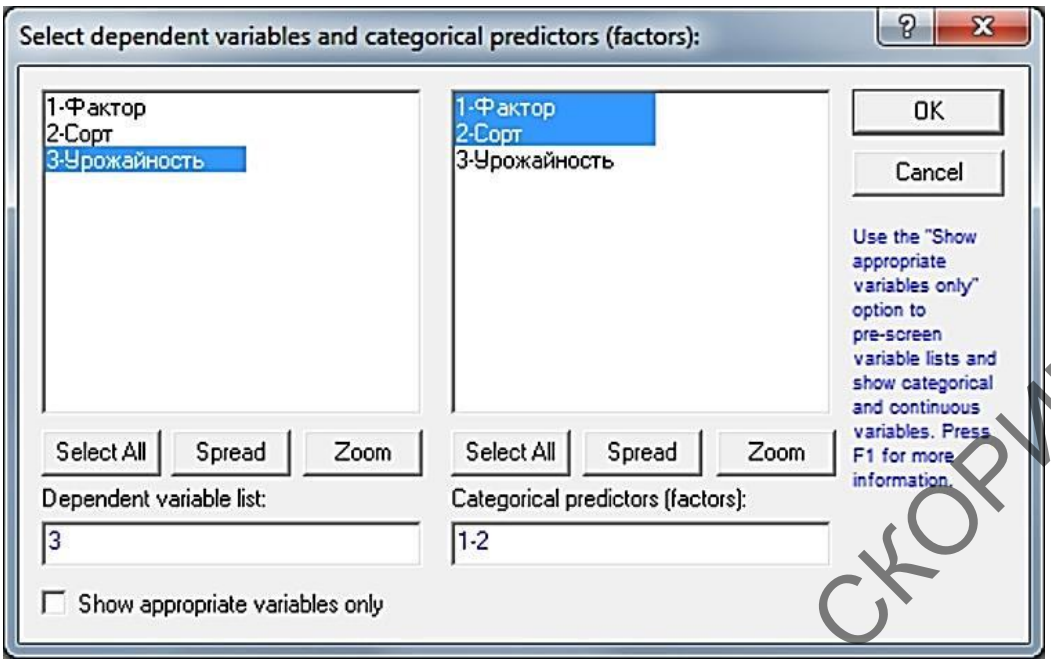


Рисунок 6.27 – Выбор переменных для анализа

Шаг 4. Выбор фактора.

После выбора переменных и нажатия кнопки **ОК** программа возвращается к первоначальному окну модуля **Factorial ANOVA** (*Факторный дисперсионный анализ*) (рисунок 6.26).

Для определения нужных нам факторов в этом же окне необходимо нажать кнопку **Factor codes** (*Коды фактора*) и в появившемся окне (рисунок 6.28) нужно указать название фактора, перечисленного в первой и второй переменных, влияние которого мы пытаемся учитывать.

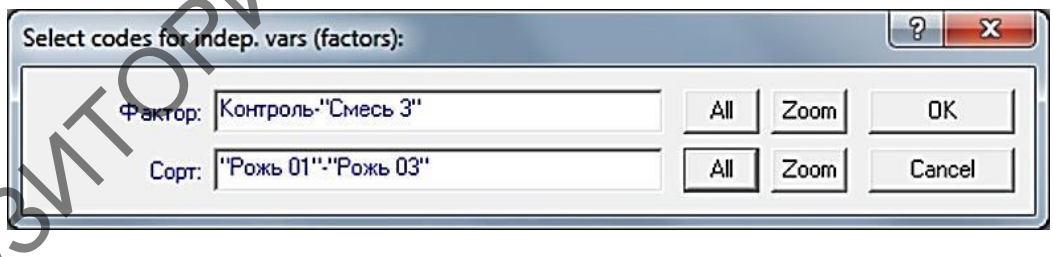


Рисунок 6.28 – Выбор кодов фактора

В нашем случае учитывается влияние всех факторов, поэтому нужно нажать кнопку **All** (*Все*) для каждого из группирующих факторов и далее щелкнуть левой клавишей мыши на **ОК**. Программа вернётся в окно модуля **Factorial ANOVA** (*Факторный дисперсионный анализ*), которое принимает окончательный вид перед проведением анализа (рисунок 6.29). После необходимо кликнуть левой клавишей мыши на **ОК**.

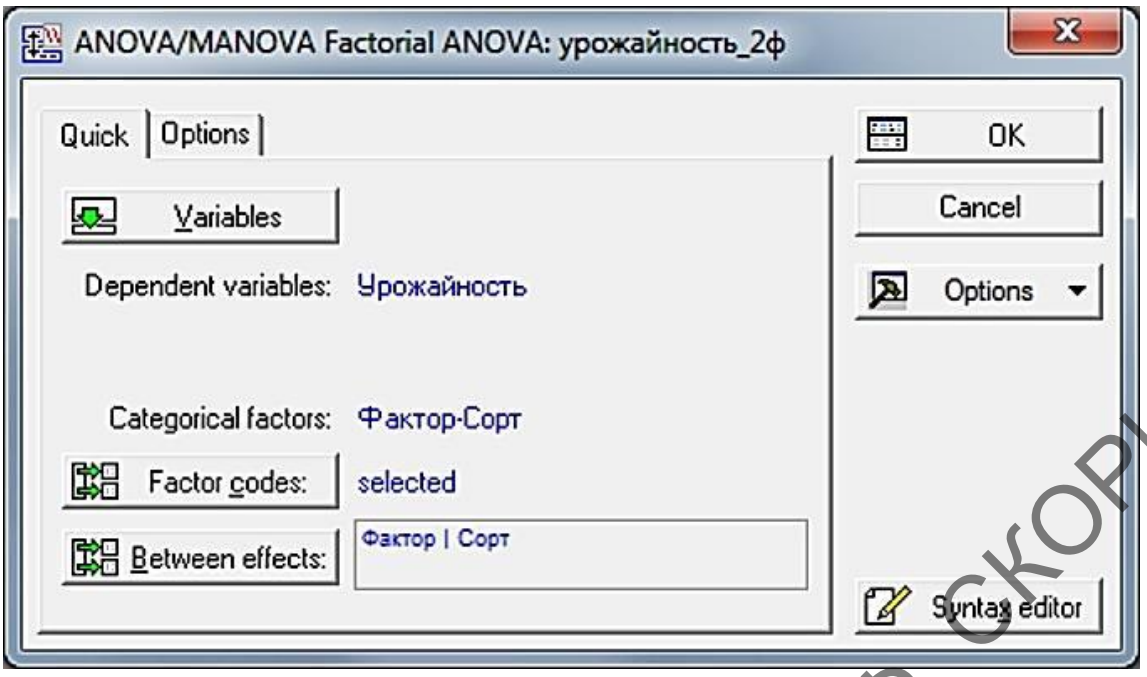


Рисунок 6.29 – Диалоговое окно модуля **Factorial ANOVA**, подготовленное для анализа

Шаг 5. Просмотр результатов.

После нажатия на кнопки **ОК** на предыдущем шаге выполнения анализа на экране появляется диалоговое окно результатов (рисунок 6.12).

Для простого просмотра результатов анализа достаточно нажать кнопку **All effects** (*Все эффекты*). Результаты будут отображены в специальной таблице (рисунок 6.30).

Univariate Tests of Significance for Урожайность (урожайность_2ф)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	21450.26	1	21450,26	13169,77	0,000000
Фактор	62,44	3	20,81	12,78	0,000480
Сорт	71,01	2	35,51	21,80	0,000101
Фактор*Сорт	102,03	6	17,00	10,44	0,000360
Error	19,54	12	1,63		

Рисунок 6.30 – Таблица результатов двухфакторного **ANOVA**

Обозначения в таблице такие же, как и при однофакторном анализе.

При анализе результатов следует обратить внимание на вторую и третью строки: «Фактор» и «Сорт» соответственно. В конце этих строк приведены вероятности ошибок для нулевых гипотез об отсутствии влияния вида удобрения и сорта ржи на её урожайность. Видно, что в обоих случаях $p < 0{,}05$. Это свидетельствует о значительном влиянии указанных факторов на урожайность ржи.

Четвёртая строка («Фактор*Сорт») отражает степень взаимного влияния исследуемых факторов на урожайность. Исходя из показателей *p*, можно отметить, что взаимное их воздействие также влияет на урожайность ржи. Аналогичным образом в программе STATISTICA можно выполнить дисперсионный анализ и с большим количеством факторов.

6.3 Непараметрический дисперсионный анализ в STATISTICA 7.0

6.3.1 Непараметрический дисперсионный анализ при оценке зависимых переменных (дисперсионный анализ Фридмана)

Рассмотрим этот вид анализа на данных по учёту численности жуков жужелиц на опушке смешанного леса в различные месяцы полевого периода исследований (таблица 6.3).

Таблица 6.3 – Данные по учёту количества жужелиц в почвенных ловушках на опушке леса за вегетационный период

в экземплярах

Вид	Месяцы						
	IV	V	VI	VII	VIII	IX	X
Broscus cephalotes	2	6	6	4	2	1	0
Carabus nemoralis	4	12	15	6	6	3	1
Carabus hortensis	2	3	4	5	2	1	1
Carabus arvensis	1	1	3	2	1	0	0
Pterostichus niger	6	15	12	15	6	3	2
Pterostichus melanarius	3	11	15	12	8	2	1
Calathus erratus	8	23	34	12	4	23	4
Calathus melanocephalus	2	3	4	2	1	1	0
Amara aenea	3	6	11	10	6	3	1
Harpalus latus	4	5	3	6	4	8	2

Шаг 1. Создание электронной таблицы с данными.

При создании этой таблицы названия видов жужелиц следует писать в заголовках строк значений варианты, а именам переменных

присвоить номера (или названия) месяцев. Пример заполненной электронной таблицы представлен на рисунке 6.31.

Data: жуки (7v by 10c)

	1 IV	2 V	3 VI	4 VII	5 VIII	6 IX	7 X
Broscus cephalotes	2	6	6	4	2	1	0
Carabus nemoralis	4	12	15	6	6	3	1
Carabus hortensis	2	3	4	5	2	1	1
Carabus arvensis	1	1	3	2	1	0	0
Pterostichus niger	6	15	12	15	6	3	2
Pterostichus melanarius	3	11	15	12	8	2	1
Calathus erratus	8	23	34	12	4	23	4
Calathus melanocephalus	2	3	4	2	1	1	0
Amara aenea	3	6	11	10	6	3	1
Harpalus latus	4	5	3	6	4	8	2

Рисунок 6.31 – Электронная таблица данных для расчёта дисперсионного анализа Фридмана

Шаг 2. Выбор модуля.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **Non-parametrics** (*Непараметрические методы*) (рисунок 6.32), далее – **Comparing multiple dependent samples** (*Сравнение нескольких зависимых выборок*) (рисунок 6.33) и нажать **ОК**.

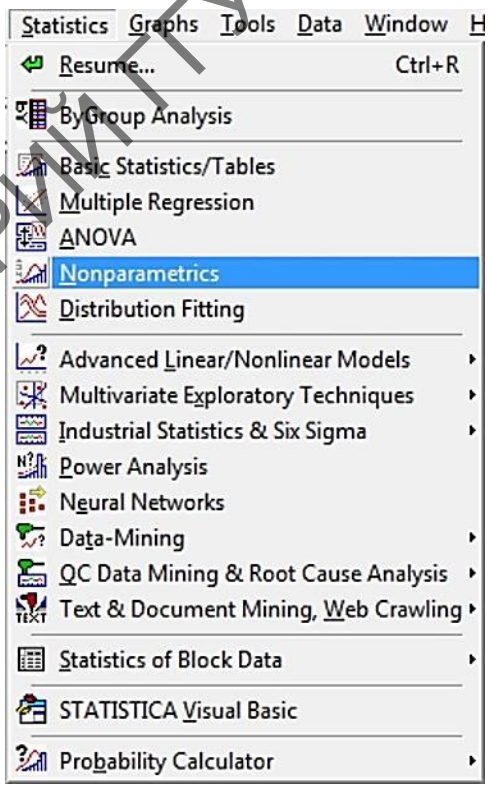


Рисунок 6.32 – Выбор опции **Nonparametrics**

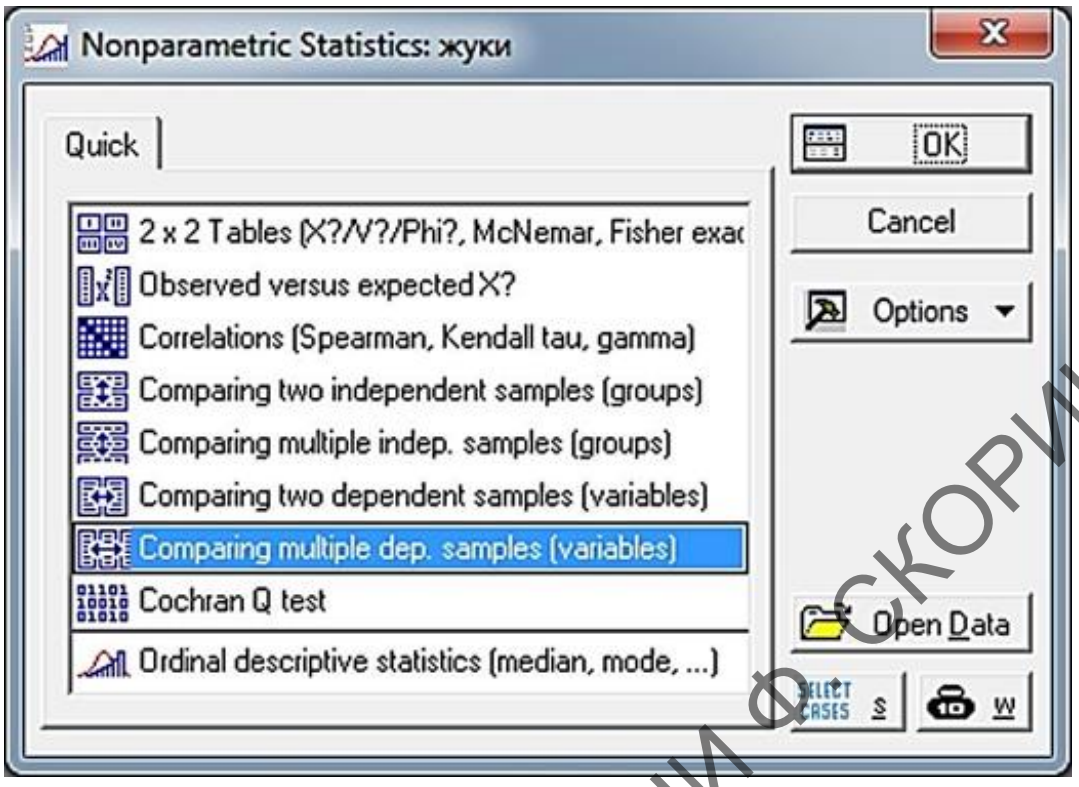


Рисунок 6.33 – Выбор опции **Comparing multiple dependent samples**

Шаг 3. Выбор переменных.

В появившемся на экране диалоговом окне (рисунок 6.34) необходимо нажать кнопку **Variables** (*Переменные*), выбрать переменные, которые должны участвовать в анализе (рисунок 6.35), и нажать **ОК**.

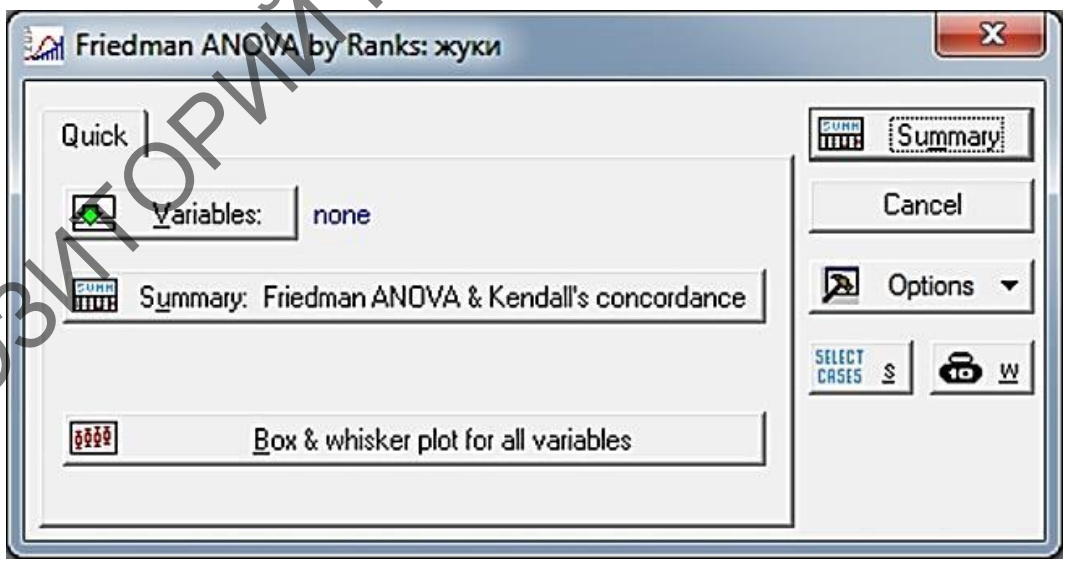


Рисунок 6.34 – Диалоговое окно модуля **Friedman ANOVA**

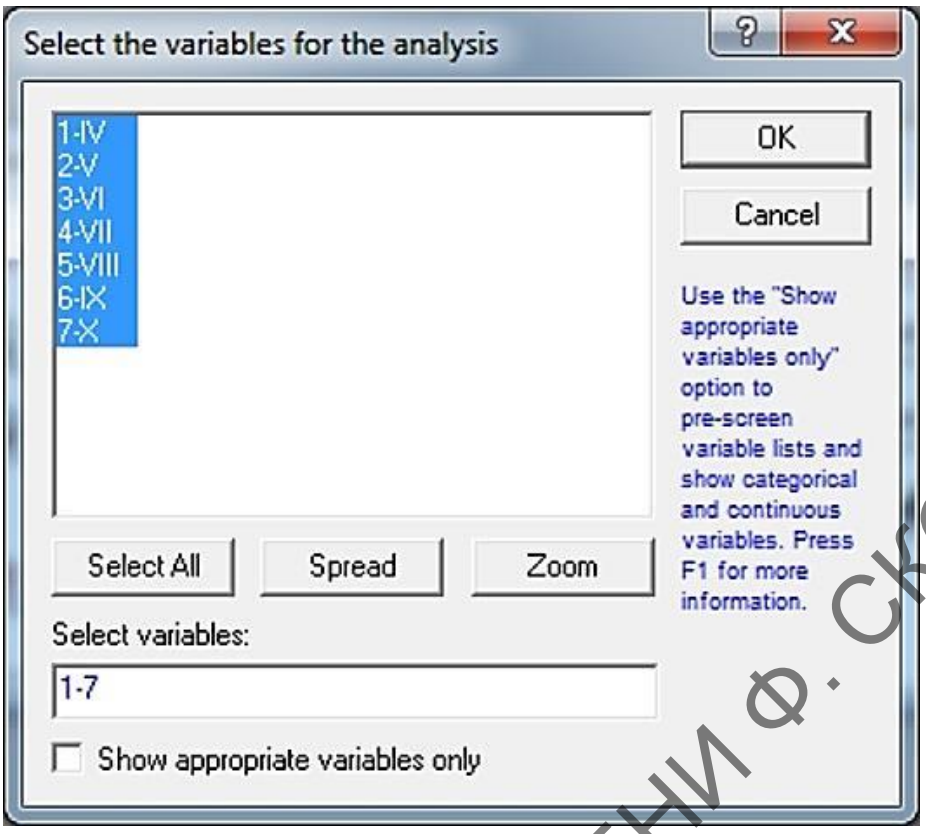


Рисунок 6.35 – Выбор переменных для анализа

Шаг 4. Запуск анализа.

После выбора переменных и нажатия кнопки **ОК** программа возвращается к первоначальному окну модуля **Friedman ANOVA** (*Дисперсионный анализ Фридмана*) (рисунок 6.36).

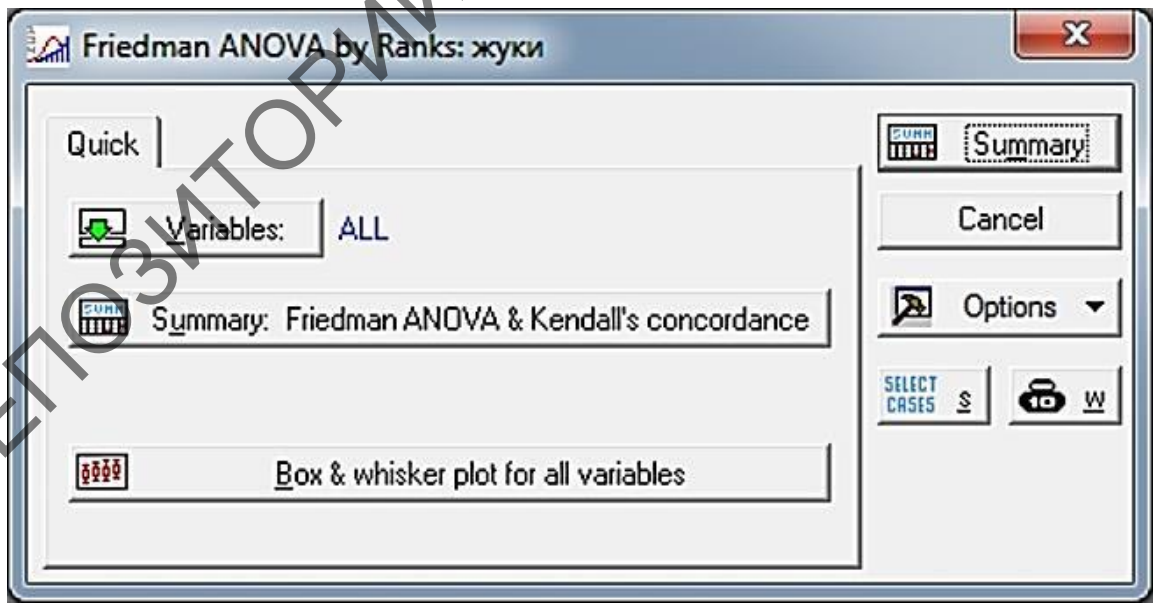


Рисунок 6.36 – Диалоговое окно модуля **Friedman ANOVA**, подготовленное к проведению анализа

Затем необходимо запустить анализ, нажав кнопку **Summary: Friedman ANOVA and Kendall's concordance** (*Результат: ANOVA по Фридману и критерий согласованности Кендалла*).

Шаг 5. Интерпретация результатов.

После нажатия указанной выше кнопки на экране появится таблица с результатами (рисунок 6.37).

Friedman ANOVA and Kendall Coeff. of Concordance (жуки)
ANOVA Chi Sqr. (N = 10, df = 6) = 42,23290 p = ,00000
Coeff. of Concordance = ,70388 Aver. rank r = ,67098

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
IV	3,400000	34,00000	3,50000	2,121320
V	5,400000	54,00000	8,50000	6,770032
VI	6,150000	61,50000	10,70000	9,499123
VII	5,550000	55,50000	7,40000	4,550946
VIII	3,500000	35,00000	4,00000	2,449490
IX	2,850000	28,50000	4,50000	6,867799
X	1,150000	11,50000	1,20000	1,229273

Рисунок 6.37 – Таблица результатов **Friedman ANOVA**

В этой таблице имеются следующие столбцы:
– **Average Rank** – *Средний ранг*;
– **Sum of Ranks** – *Сумма рангов*;
– **Mean** – *Средняя арифметическая*;
– **Std. Dev.** – *Стандартное отклонение*.

В заголовке таблицы указывается величина уровня значимости p для нулевой гипотезы о том, что на протяжении 7 месяцев численность жужелиц не изменялась. При $p < 0{,}05$ (как в нашем случае) следует сделать вывод о наличии статистически значимых различий между группами.

В заголовке также указан коэффициент согласованности Кендалла (*Kendall Coeff. of Concordance*) – усреднение коэффициентов корреляции Спирмена для каждой пары участвующих в анализе групп (чем больше различия между группами, тем ближе коэффициент согласованности Кендалла к 1).

Шаг 6. Графическое отображение результата анализа.

В первоначальном окне модуля **Friedman ANOVA** (*Дисперсионный анализ Фридмана*) (рисунок 6.36) необходимо нажать кнопку **Box & whisker plot for all variables** (*Диаграмма размахов для всех переменных*), выбрать тип диаграммы **Mean/SE/SD** (Средняя/Стандартная ошибка/Стандартное отклонение) и нажать **ОК**. Итог отображён на рисунке 6.38.

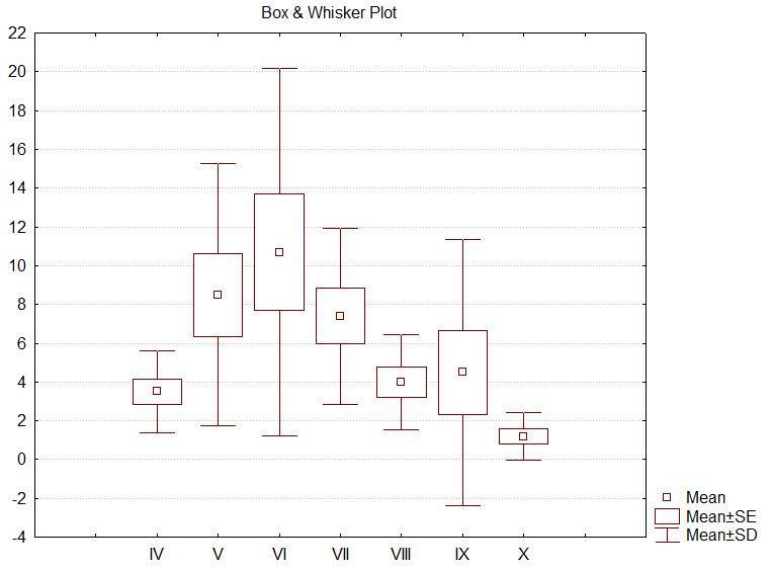


Рисунок 6.38 – Диаграмма размахов итогов дисперсионного анализа Фридмана

6.3.2 Непараметрический дисперсионный анализ Крускала-Уоллиса

Не секрет, что в большинстве полевых (и не только!) биологических исследований распределение выборки далеко не всегда подчиняется закону нормального распределения. В связи с этим применение параметрического дисперсионного анализа невозможно. В этом случае используется непараметрический дисперсионный анализ Крускала-Уоллиса.

Рассмотрим данный вид анализа на примере оценки влияния на численность беспозвоночных и их удалённости от площадки с работающей установкой прокачки нефти (таблица 6.4), изучаемых в летние месяцы.

Таблица 6.4 – Численность беспозвоночных в почвенных ловушках по линии удаления от скважины нефти

в экземплярах

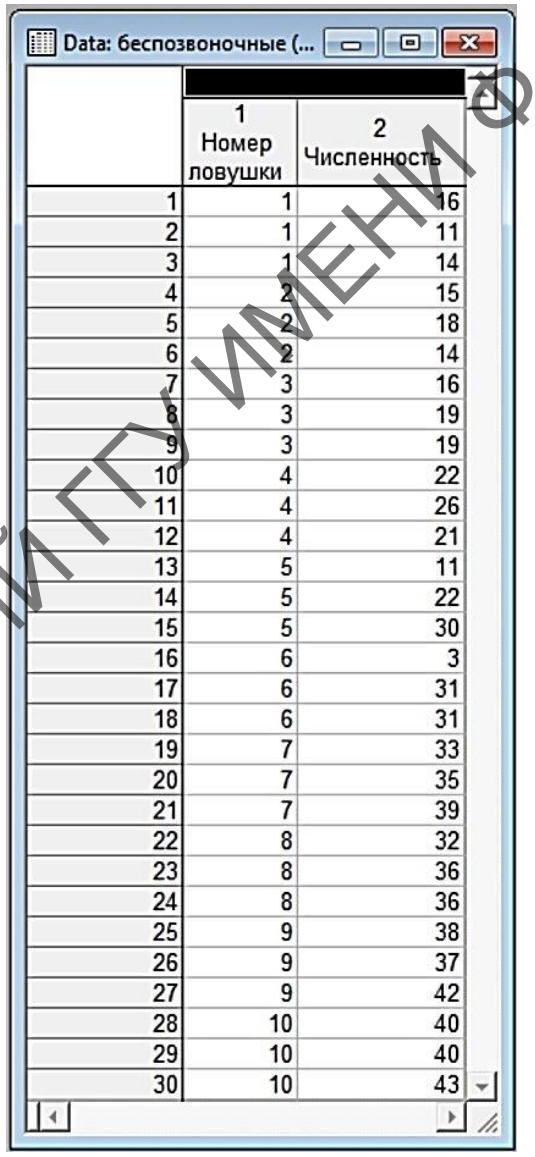
Номер ловушки	Июнь	Июль	Август
1	16	11	14
2	15	18	14
3	16	19	19
4	22	26	21
5	11	22	30
6	30	31	31
7	33	35	39
8	32	36	36
9	38	37	42
10	40	40	43

Шаг 1. Создание электронной таблицы с данными.

При создании этой таблицы названия видов жужелиц следует использовать 2 переменные, как в случае параметрического однофакторного дисперсионного анализа. Пример заполненной электронной таблицы представлен на рисунке 6.39.

Шаг 2. Выбор модуля.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **Non-parametrics** (*Непараметрические методы*) (рисунок 6.32), далее – **Comparing multiple independent samples (groups)** (*Сравнение нескольких независимых выборок*) (рисунок 6.40) и нажать **ОК**.



	1 Номер ловушки	2 Численность
1	1	16
2	1	11
3	1	14
4	2	15
5	2	18
6	2	14
7	3	16
8	3	19
9	3	19
10	4	22
11	4	26
12	4	21
13	5	11
14	5	22
15	5	30
16	6	3
17	6	31
18	6	31
19	7	33
20	7	35
21	7	39
22	8	32
23	8	36
24	8	36
25	9	38
26	9	37
27	9	42
28	10	40
29	10	40
30	10	43

Рисунок 6.39 – Электронная таблица данных для расчёта дисперсионного анализа Крускала-Уоллиса

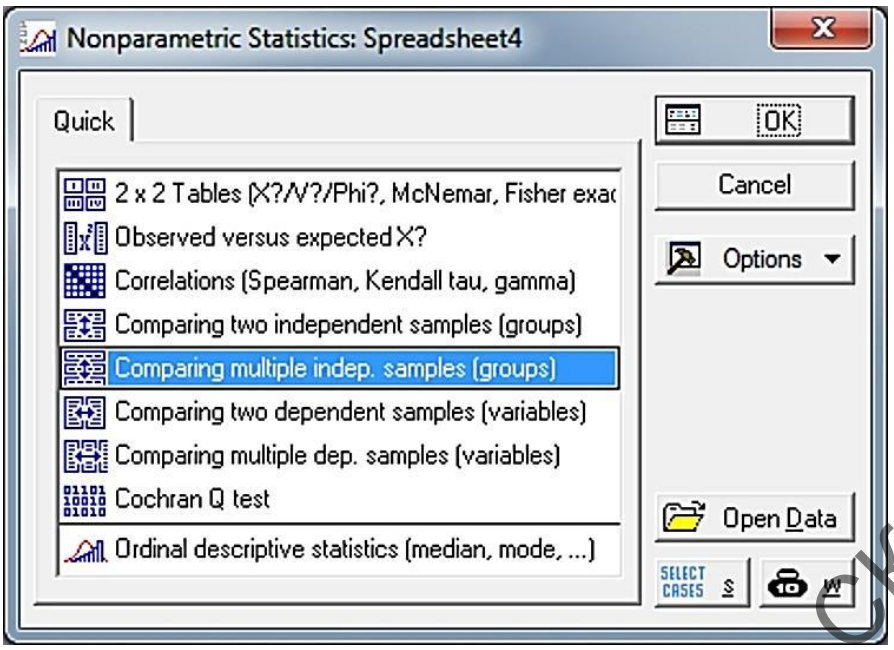


Рисунок 6.40 – Выбор опции **Comparing multiple independent samples**

Шаг 3. Выбор переменных.

В появившемся на экране диалоговом окне анализа Крускала-Уоллиса (рисунок 6.41), в закладке **Quick** (*Быстрый анализ*) необходимо нажать кнопку **Variables** (*Переменные*) и в диалоговом окне выбора переменных указать слева зависимую переменную, а справа – независимую (группирующую) переменную (рисунок 6.42) и далее щелкнуть левой клавишей мыши на **ОК**.

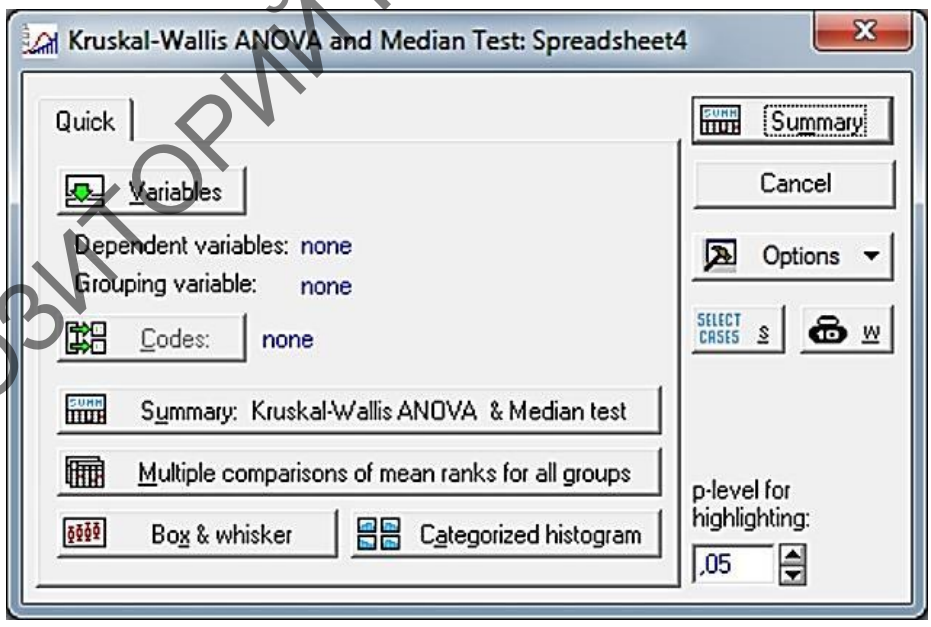


Рисунок 6.41 – Диалоговое окно модуля **Kruskal-Wallis ANOVA**

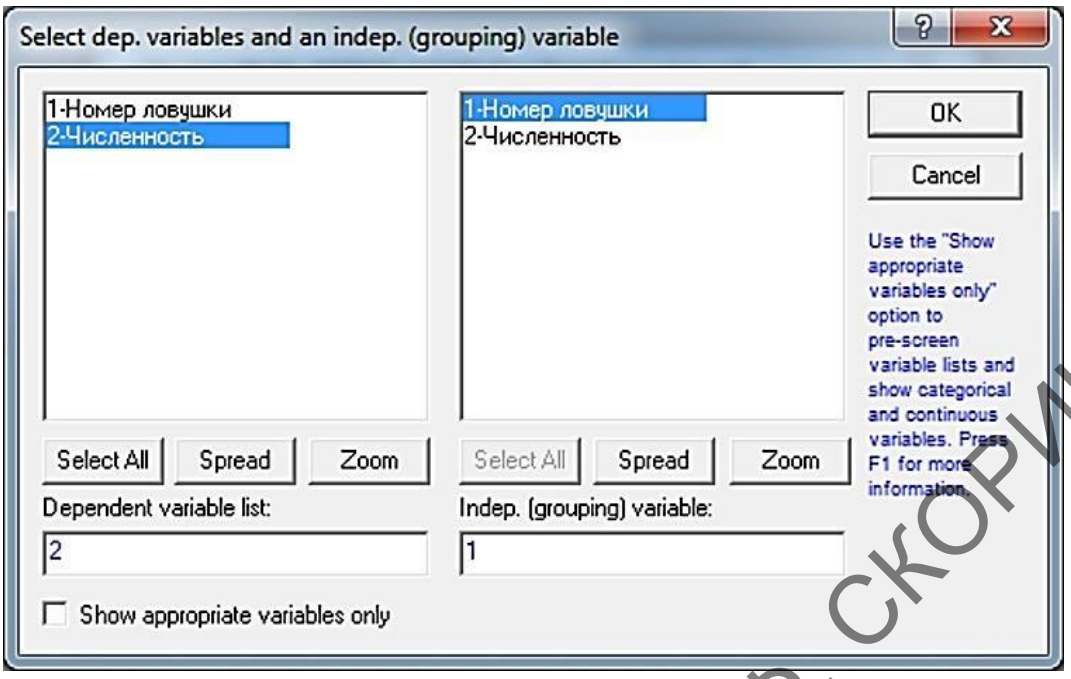


Рисунок 6.42 – Выбор переменных для анализа

Шаг 4. *Выбор фактора.*

После выбора переменных и нажатия кнопки **ОК** программа возвращается к первоначальному окну модуля **Kruskal-Wallis ANOVA** (*Дисперсионный анализ Крускала-Уоллиса*) (рисунок 6.41).

Для определения нужных нам факторов (для расчёта) в этом же окне необходимо нажать кнопку **Codes** (*Коды*) и в появившемся окне (рисунок 6.43) нужно указать название фактора, перечисленного в первой переменной, влияние которого мы пытаемся учитывать.

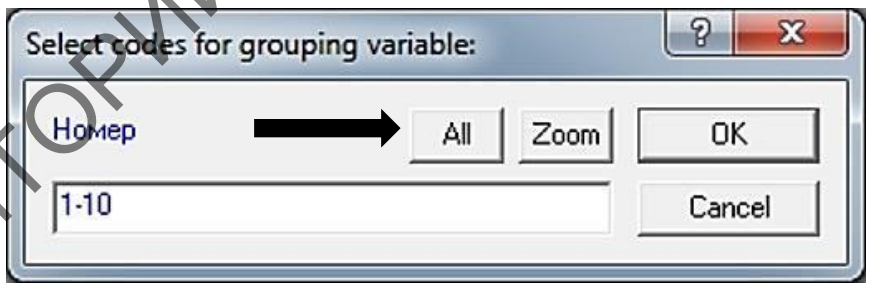


Рисунок 6.43 – Выбор фактора

В нашем случае учитывается влияние всех факторов, поэтому нужно нажать кнопку **All** (*Все*) и далее щелкнуть левой клавишей мыши на **ОК**. Программа вернётся в окно модуля **Kruskal-Wallis ANOVA** (*Дисперсионный анализ Крускала-Уоллиса*), которое принимает окончательный вид перед проведением анализа (рисунок 6.44). После необходимо кликнуть левой клавишей мыши на **ОК**.

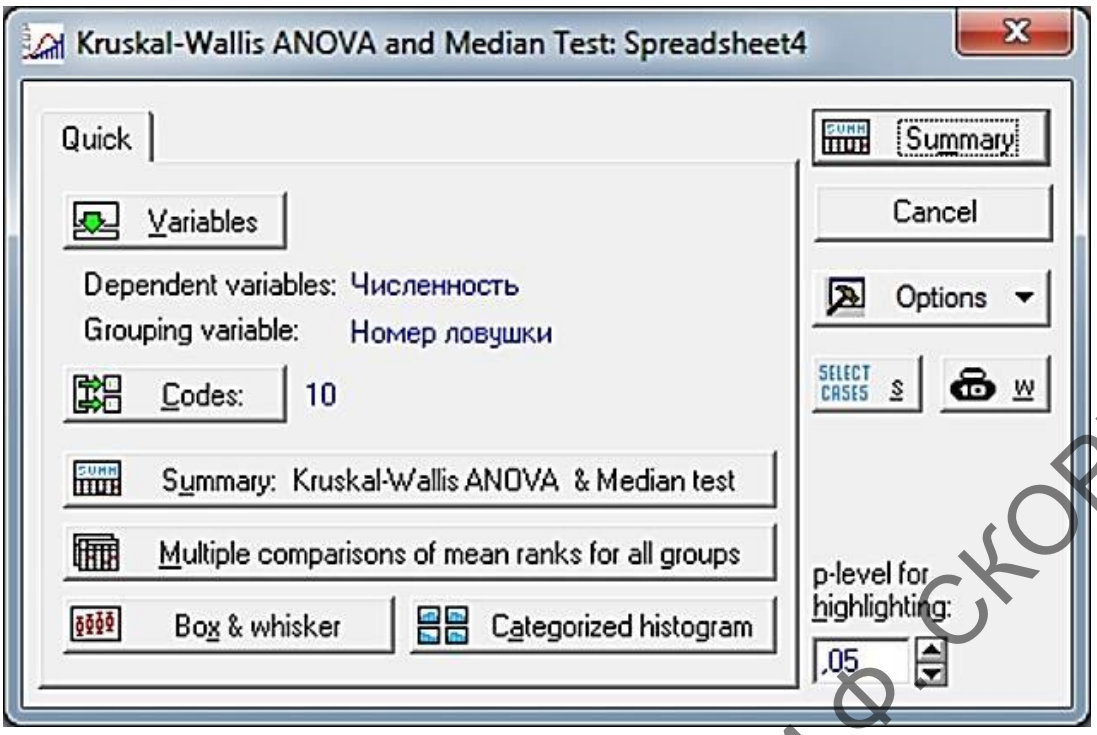


Рисунок 6.44 – Диалоговое окно модуля **Kruskal-Wallis ANOVA**, подготовленное для анализа

Шаг 5. Просмотр и интерпретация результатов.

После нажатия на кнопки **ОК** на предыдущем шаге выполнения анализа на экране появляется диалоговое окно результатов (рисунок 6.44).

Далее необходимо нажать на кнопку **Summary: Kruskal-Wallis ANOVA and Median test** (*Результат: ANOVA по Крускалу-Уоллису и медианный тест*). На экране появится таблица с результатами (рисунок 6.45).

Median Test, Overall Median = 19,0000; Численность (беспозвоночные)
Independent (grouping) variable: Номер ловушки
Chi-Square = 14,61279 df = 9 p = ,1021

Dependent: Численность	2	3	4	5	6	7	8	9	10	Total
<= Median: observed	3,00000	3,00000	0,00000	1,000000	1,000000	0,00000	0,00	0,00	0,00	11,00000
expected	1,65000	1,65000	1,65000	1,650000	1,650000	1,10000	0,00	0,00	0,00	
obs.-exp.	1,35000	1,35000	-1,65000	-0,650000	-0,650000	-1,10000	0,00	0,00	0,00	
> Median: observed	0,00000	0,00000	3,00000	2,000000	2,000000	2,00000	0,00	0,00	0,00	9,00000
expected	1,35000	1,35000	1,35000	1,350000	1,350000	0,90000	0,00	0,00	0,00	
obs.-exp.	-1,35000	-1,35000	1,65000	0,650000	0,650000	1,10000	0,00	0,00	0,00	
Total: observed	3,00000	3,00000	3,00000	3,000000	3,000000	2,00000	0,00	0,00	0,00	20,00000

Рисунок 6.45 – Участок итоговой таблицы анализа **Kruskal-Wallis ANOVA**

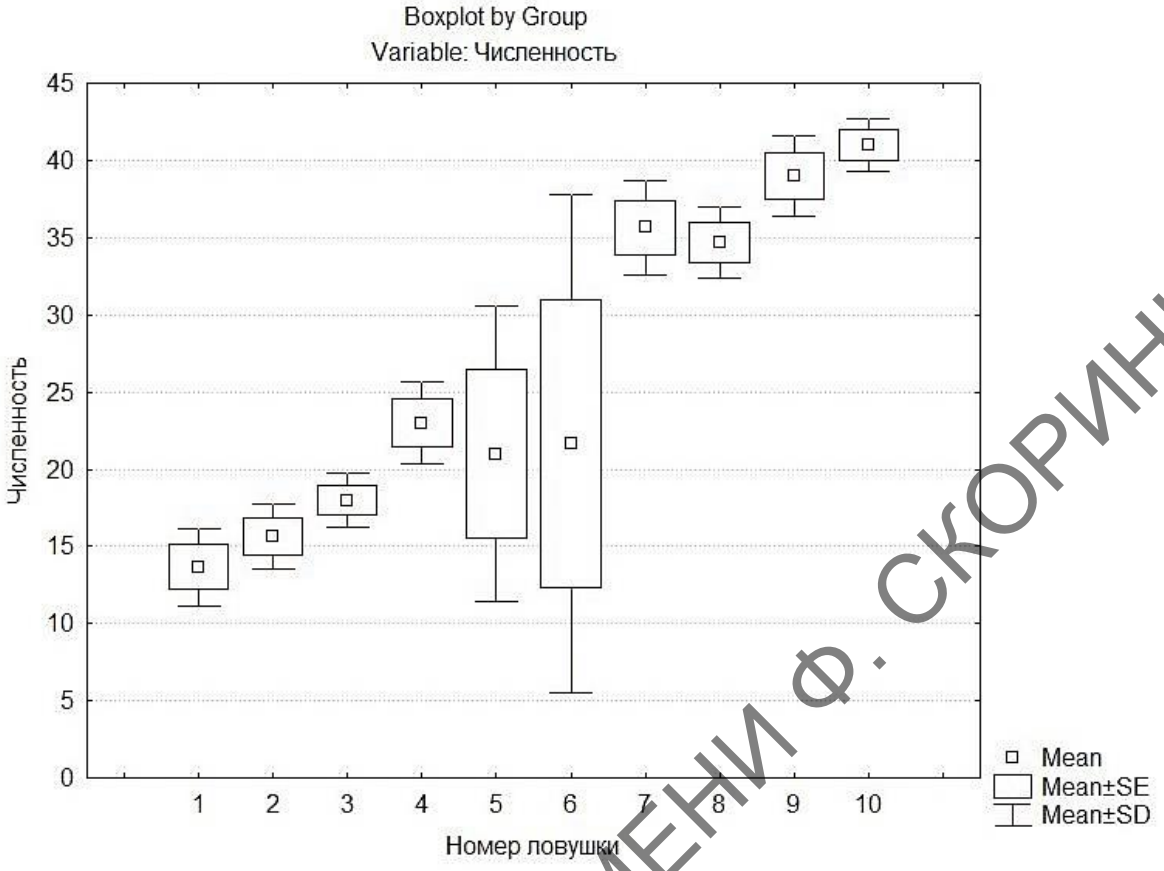


Рисунок 6.46 – Диаграмма размахов итогов дисперсионного анализа Крускала-Уоллиса

В заголовке таблицы указывается величина уровня значимости *p* для нулевой гипотезы о том, что по мере удалённости ловушек численность беспозвоночных не изменялась. При $p > 0{,}05$ (как в нашем случае) следует сделать вывод об отсутствии статистически значимых различий между группами.

Шаг 6. Графическое отображение результата анализа.

В первоначальном окне модуля **Kruskal-Wallis ANOVA** (*Дисперсионный анализ Крускала-Уоллиса*) (рисунок 6.44), необходимо нажать кнопку **Box & whisker** (*Диаграмма размахов*), выбрать тип диаграммы **Mean/SE/SD** (Средняя/Стандартная ошибка/Стандартное отклонение) и нажать **ОК**. Итог отображён на рисунке 6.46.

Задания

1) Были получены данные по внесению смеси удобрений (*a*) и их влиянию на прирост ели в питомнике, в см (*b*).

a	9,4	8,4	4,9	13,4	11,1	3,6	15,1	13,2	12,0	11,6	18,2
b	21,1	8,2	16,6	10,5	18,4	11,3	18,3	11,2	19,4	10,0	11,2
a	12,0	16,7	7,3	5,6	5,8	7,8	12,4	8,5	5,7	10,7	11,3
b	6,6	14,7	18,2	9,2	9,5	10,6	3,1	24,9	10,0	19,0	16,2
a	15,2	9,8	8,0	12,0	13,3	6,7	5,7	12,1	8,0	10,4	12,4
b	6,3	16,3	18,8	10,1	17,6	15,8	7,8	16,8	19,3	19,1	11,1
a	4,7	11,8	5,4	10,3	6,5	10,1	14,2	9,8	10,0	12,4	12,5
b	2,1	24,5	17,2	13,5	14,7	7,2	16,0	4,9	19,4	11,6	20,9

Выясните с помощью Excel и STATISTICA, влияет ли фактор внесения данной смеси удобрений на прирост ели. Постройте график «ящики и усы» для наглядной демонстрации модели.

2) Были получены данные по внесению смеси удобрений (*a*) и их влиянию на прирост ели в питомнике, в см (*b*).

a	12,5	12,1	12,1	15,4	2,0	9,5	12,1	11,0	5,8	16,0	12,1
b	21,0	5,0	24,1	14,8	29,1	20,1	11,2	19,5	20,5	22,9	17,7
a	27,8	19,8	18,1	28,2	43,7	19,1	28,0	20,1	15,0	20,7	12,6
b	1,6	0,9	1,1	0,7	1,6	0,7	1,0	1,1	1,1	0,8	0,9
a	16,6	15,5	11,5	9,3	14,4	12,2	13,2	6,2	8,1	31,3	10,8
b	13,8	9,3	20,4	16,7	18,3	29,1	13,1	19,5	20,0	1,1	0,7
a	11,5	16,6	22,9	12,0	21,0	15,1	26,2	24,0	26,2	8,0	35,8
b	1,1	1,0	1,1	1,3	1,4	0,7	0,7	1,1	0,9	1,6	0,9

Выясните с помощью Excel и STATISTICA, влияет ли фактор внесения данной смеси удобрений на прирост ели. Постройте график «ящики и усы» для наглядной демонстрации модели.

3) Были получены данные по внесению смеси кормов (*a*) и их влиянию на привес телят, в кг (*b*).

a	24,4	15,8	19,5	16,5	16,5	15,0	21,4	17,5	17,2	11,5	11,5	12,0	15,3	30,2	33,3	15,4
b	21,5	23,2	16,8	20,5	8,2	19,4	28,6	21,3	19,8	12,7	23,5	20,2	19,1	29,8	25,9	22,5
a	17,3	22,8	11,0	17,8	23,5	11,9	15,0	24,6	19,9	20,8	22,8	29,2	25,4	19,8	25,6	10,8
b	33,0	14,2	18,3	27,2	5,6	22,6	31,4	26,4	33,9	26,6	14,1	20,7	25,8	21,7	29,4	27,0

Выясните с помощью Excel и STATISTICA, влияет ли фактор внесения данной смеси кормов на привес телят. Постройте график «ящики и усы» для наглядной демонстрации модели.

4) Было проведено исследование степени заболевания гастритом работников заводов концерна.

Номер завода	Цех		
	Доменный	Прокатный	Сборочный
1	23,4	26,4	40,1
2	13,6	45,6	49,6
3	26,1	37,5	32,1
4	22,4	29,4	38,0

Выясните с помощью Excel и STATISTICA, влияет ли фактор цеха и фактор завода на заболеваемость гастритом рабочих.

5) Был проведён сбор показателей концентрации детергента на 5 разных створах реки вниз по течению на протяжении весенних и осенних месяцев.

Створ	Март	Апрель	Май	Сентябрь	Октябрь	Ноябрь
1	0,6	0,5	0,6	1,2	0,9	0,9
2	0,8	0,7	0,7	1,9	1,8	2,2
3	0,6	0,7	0,4	0,8	0,8	0,7
4	0,4	0,3	0,3	0,7	0,5	0,8
5	0,1	0,1	0,2	0,9	1,1	1,0

Выясните при помощи дисперсионного анализа Фридмана, влияет ли сезон на концентрацию детергента на створах реки.

6) Была изучена плотность популяции беззубки на разной глубине озера в летние месяцы.

Факторы	VI	VI	VI	VI	VI	VII	VII	VII	VII	VII	VIII	VIII	VIII	VIII	VIII
Глубина, м	1	1	1	1	1	3	3	3	3	3	5	5	5	5	5
Количество	13	0	5	9	12	0	0	6	8	15	20	22	4	15	10

Выясните при помощи дисперсионного анализа Крускала-Уоллиса, влияет ли глубина на плотность популяции беззубки в озере.

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

4 Мастицкий, С. Э. Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований / С. Э. Мастицкий. – Минск : РУП «Институт рыбного хозяйства», 2009. – 76 с.

ТЕМА 7. ДИСКРИМИНАНТНЫЙ АНАЛИЗ В STATISTICA 7.0

7.1 Краткая характеристика дискриминантного анализа.
7.2 Реализация дискриминантного анализа в STATISTICA 7.0.

7.1 Краткая характеристика дискриминантного анализа

Дискриминантный анализ является одним из методов многомерного статистического анализа и относится к так называемым анализам классификации, а именно к классификации при наличии обучающих выборок. *Классификация* – это разделение рассматриваемой совокупности объектов или явлений на однородные группы.

Цель дискриминантного анализа – на основе измерения различных характеристик (признаков, параметров) объекта классифицировать его, то есть отнести к одной из нескольких групп (классов) определённым оптимальным способом. Под этим оптимальным способом понимается либо минимум математического ожидания потерь, либо минимум вероятности ложной классификации.

Данный вид анализа является многомерным, так как измеряются несколько параметров объекта, по крайней мере, больше одного, например, температура, влажность в технологическом процессе, давление, состав крови, температура больного и т. д.

Для практических целей реализовано два общих метода дискриминантного анализа: стандартный и пошаговый (включения и исключения). Указанные методы дискриминантного анализа аналогичны методам множественной регрессии.

При *стандартом методе* в случае двух групп путём наименьших квадратов строится регрессионная прямая (зависимая переменная – номер группы, все остальные переменные – независимые). Если групп несколько, то можно представить себе, что вначале строится дискриминация между группами 1 и 2, затем между 2 и 3, и так далее.

В *пошаговом методе* модель строится последовательно по шагам. Для метода включения на каждом шаге оценивает вклад в функцию дискриминации не включенных в модель переменных. Переменная, дающая наибольший вклад, включается в модель, далее система переходит к следующему шагу. Если применяется так называемый пошаговый метод исключения, то вначале в модель включаются все переменные, затем производится их последовательное исключение.

Типичные области применения дискриминантного анализа – биология, медицина, управление производством, экономика, геология, контроль качества.

7.2 Реализация дискриминантного анализа в STATISTICA 7.0

Знакомство с возможностями проведения дискриминантного анализа в программном пакете STATISTICA лучше всего начать с разбора уже апробированного примера. Таким примером может являться анализ цветков ириса Фишера.

Цель классификации состоит в том, чтобы по результатам измерения длины и ширины чашелистиков и лепестков цветков ириса отнести ирис к одному из трех сортов: SETOSA, VERSICOLOR и VIRGINIC. Все необходимые данные для этого примера уже находятся в файле Irisdat.sta, который содержит результаты измерений 150 цветков ириса, по 50 для каждого сорта.

Шаг 1. Открытие электронной таблицы с данными.

Для открытия готовой электронной таблицы примера необходимо сначала в главном меню программы последовательно нажать на пункт меню **File** (*Файл*), а затем – **Open** (*Открыть*) и в открывшемся диалоговом окне выбрать необходимый нам файл Irisdat.sta (рисунок 7.1), который затем откроется как новая стандартная электронная таблица пакета STATISTICA (рисунок 7.2).

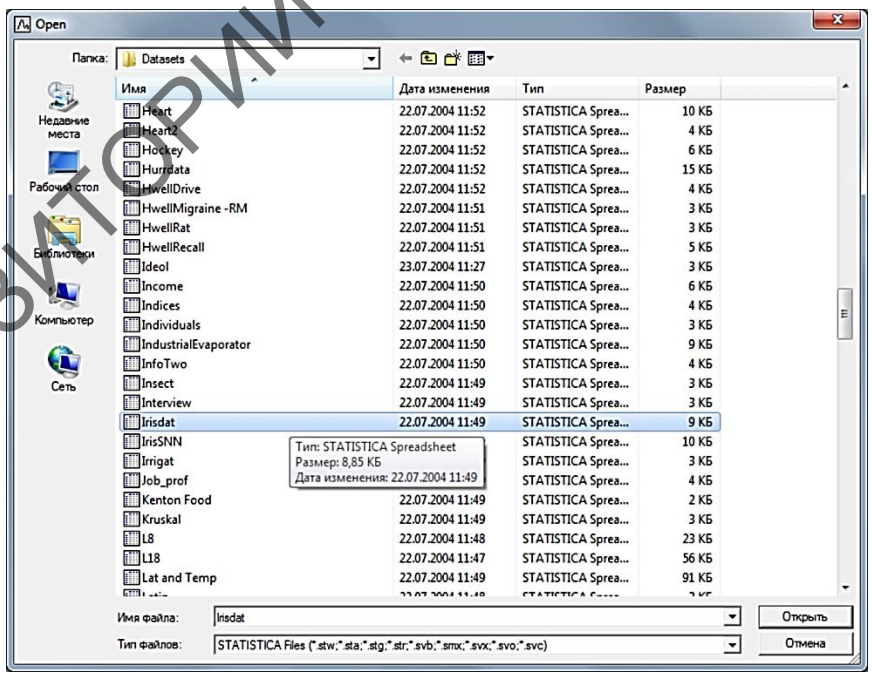
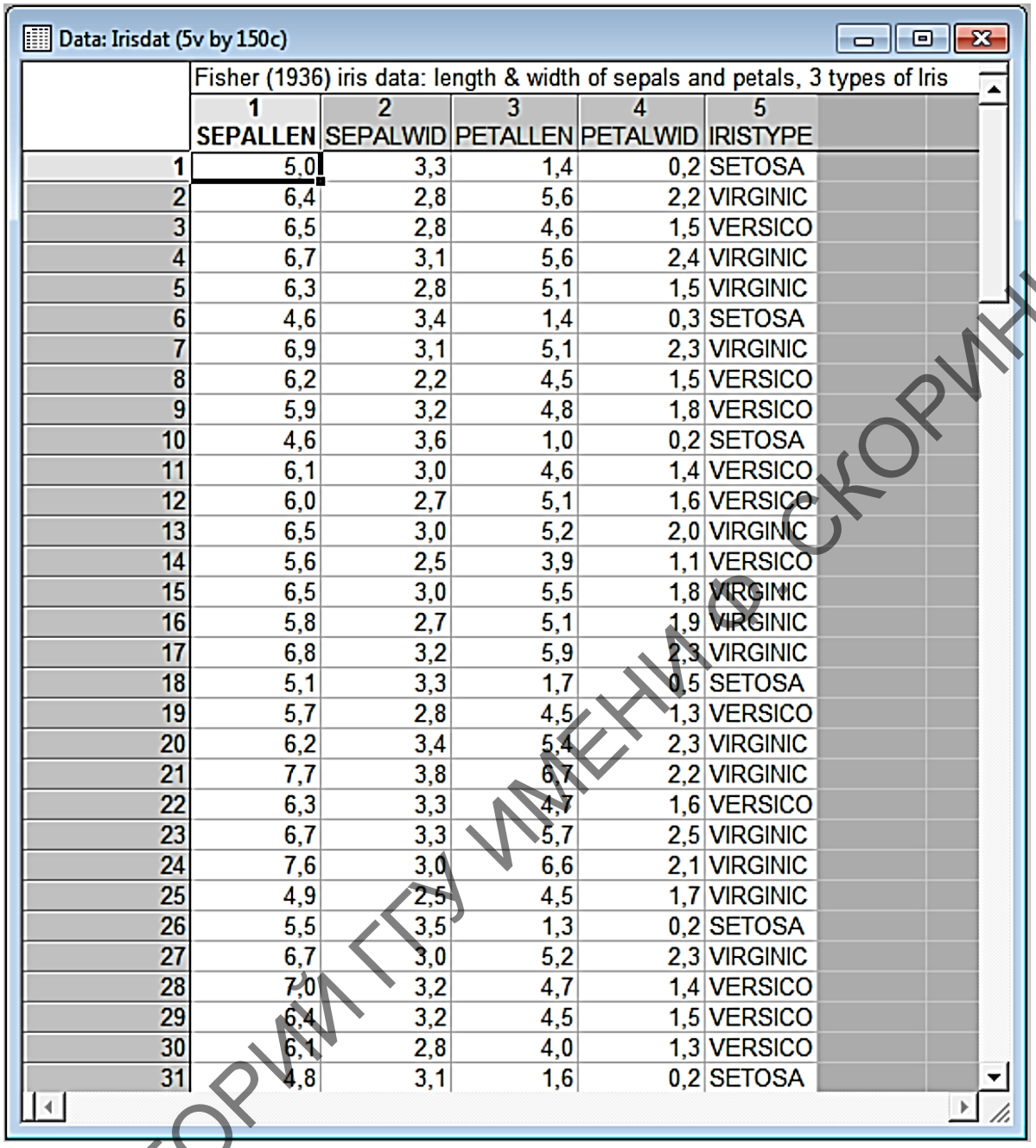


Рисунок 7.1 – Диалоговое окно **Open** в пакете **STATISTICA**



Data: Irisdat (5v by 150c)

Fisher (1936) iris data: length & width of sepals and petals, 3 types of Iris

	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
1	5,0	3,3	1,4	0,2	SETOSA
2	6,4	2,8	5,6	2,2	VIRGINIC
3	6,5	2,8	4,6	1,5	VERSICO
4	6,7	3,1	5,6	2,4	VIRGINIC
5	6,3	2,8	5,1	1,5	VIRGINIC
6	4,6	3,4	1,4	0,3	SETOSA
7	6,9	3,1	5,1	2,3	VIRGINIC
8	6,2	2,2	4,5	1,5	VERSICO
9	5,9	3,2	4,8	1,8	VERSICO
10	4,6	3,6	1,0	0,2	SETOSA
11	6,1	3,0	4,6	1,4	VERSICO
12	6,0	2,7	5,1	1,6	VERSICO
13	6,5	3,0	5,2	2,0	VIRGINIC
14	5,6	2,5	3,9	1,1	VERSICO
15	6,5	3,0	5,5	1,8	VIRGINIC
16	5,8	2,7	5,1	1,9	VIRGINIC
17	6,8	3,2	5,9	2,3	VIRGINIC
18	5,1	3,3	1,7	0,5	SETOSA
19	5,7	2,8	4,5	1,3	VERSICO
20	6,2	3,4	5,4	2,3	VIRGINIC
21	7,7	3,8	6,7	2,2	VIRGINIC
22	6,3	3,3	4,7	1,6	VERSICO
23	6,7	3,3	5,7	2,5	VIRGINIC
24	7,6	3,0	6,6	2,1	VIRGINIC
25	4,9	2,5	4,5	1,7	VIRGINIC
26	5,5	3,5	1,3	0,2	SETOSA
27	6,7	3,0	5,2	2,3	VIRGINIC
28	7,0	3,2	4,7	1,4	VERSICO
29	6,4	3,2	4,5	1,5	VERSICO
30	6,1	2,8	4,0	1,3	VERSICO
31	4,8	3,1	1,6	0,2	SETOSA

Рисунок 7.2 – Содержимое файла **Irisdat.sta**

Шаг 2. Выбор анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **Multivariate Exploratory Techniqes** (*Многомерные поисковые методы*), а затем – **Discriminant Analysis** (*Дискриминантный анализ*) (рисунок 7.3) и нажать **ОК**.

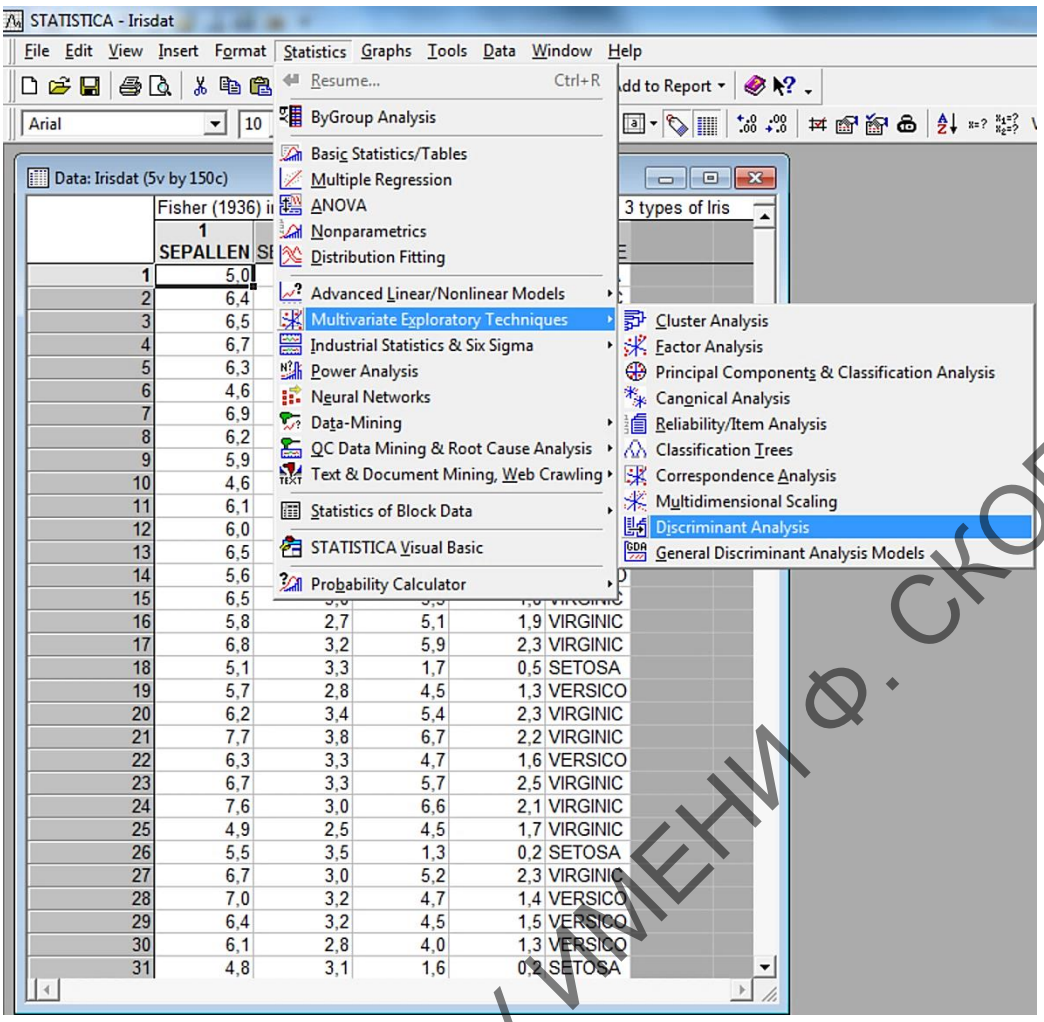


Рисунок 7.3 – Расположение модуля **Discriminant Analysis** в меню пакета STATISTICA

В результате появится диалоговое окно модуля (рисунок 7.4).

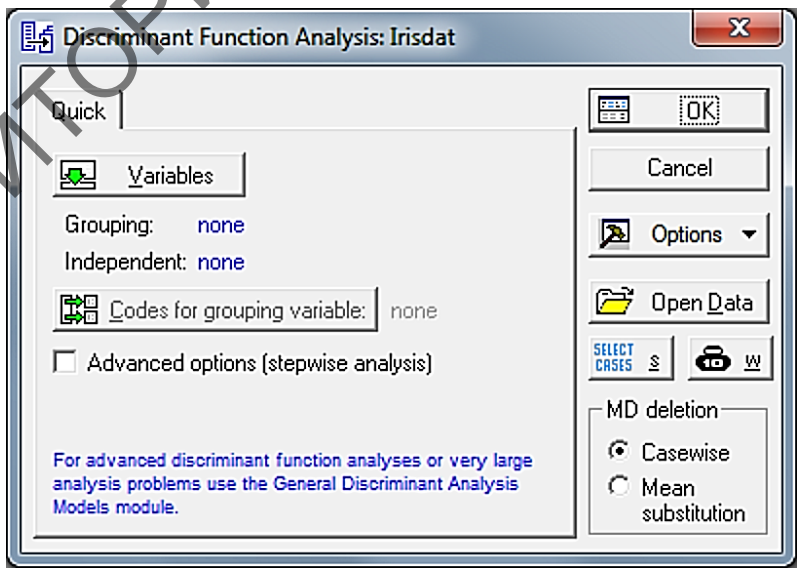


Рисунок 7.4 – Диалоговое окно модуля **Discriminant Analysis**

Шаг 3. Выбор переменных.

Для выбора переменных необходимо в диалоговом окне (рисунок 7.4) нажать кнопку **Variables** (*Переменные*) и выбрать соответствующие переменные для анализа. В данном случае группирующей переменной **Grouping variable** (*Группирующая переменная*) будет выступать сорт ириса – IRISTYPE. В качестве независимых переменных **Independent variables** (*Независимые переменные*) необходимо будет обозначить переменные **SEPALLEN** (*Длина чашелистика*), **SEPALWID** (*Ширина чашелистика*), **PETALLEN** (*Длина лепестка*), **PETALWID** (*Ширина лепестка*) (рисунок 7.5) и далее щелкнуть левой клавишей мыши на **ОК**.

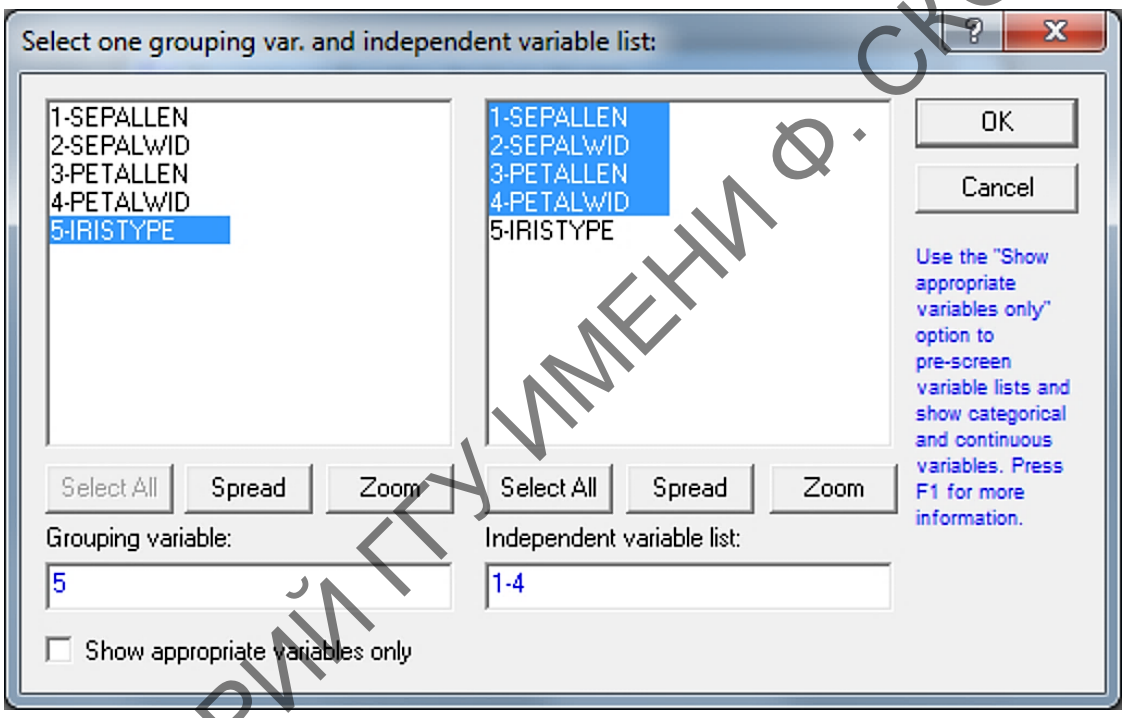


Рисунок 7.5 – Выбор переменных для дискриминантного анализа

Шаг 4. Установка кодов группирующей переменной.

После выбора переменных и нажатия кнопки **ОК** программа возвращается к первоначальному окну модуля **Discriminant Analysis** (*Дискриминантный анализ*) (рисунок 7.4).

Для установки нужных кодов независимой переменной в этом же окне необходимо нажать кнопку **Codes for grouping variable** (*Коды* для группирующей переменной) и в появившемся окне (рисунок 7.6) нужно указать название того кода переменной, который необходимо учесть. В нашем случае используются все. Поэтому нужно нажать кнопку **All** (*Все*) (рисунок 7.6) и нажать **ОК.**

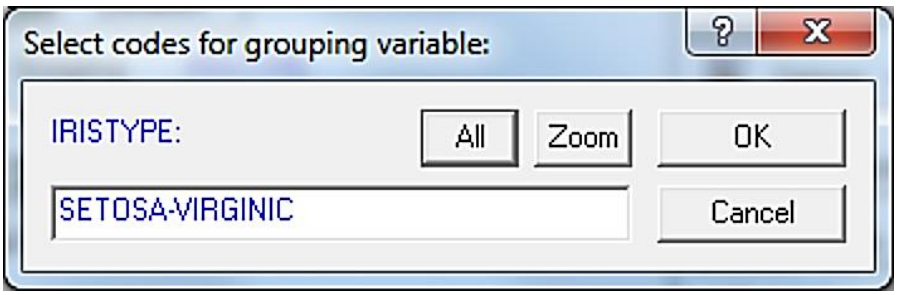


Рисунок 7.6 – Выбор кода для группирующей переменной

Программа вернётся в окно модуля **Discriminant Analysis** (*Дискриминантный анализ*), которое принимает окончательный вид перед проведением анализа (рисунок 7.7). После необходимо кликнуть левой клавишей мыши на **ОК**.

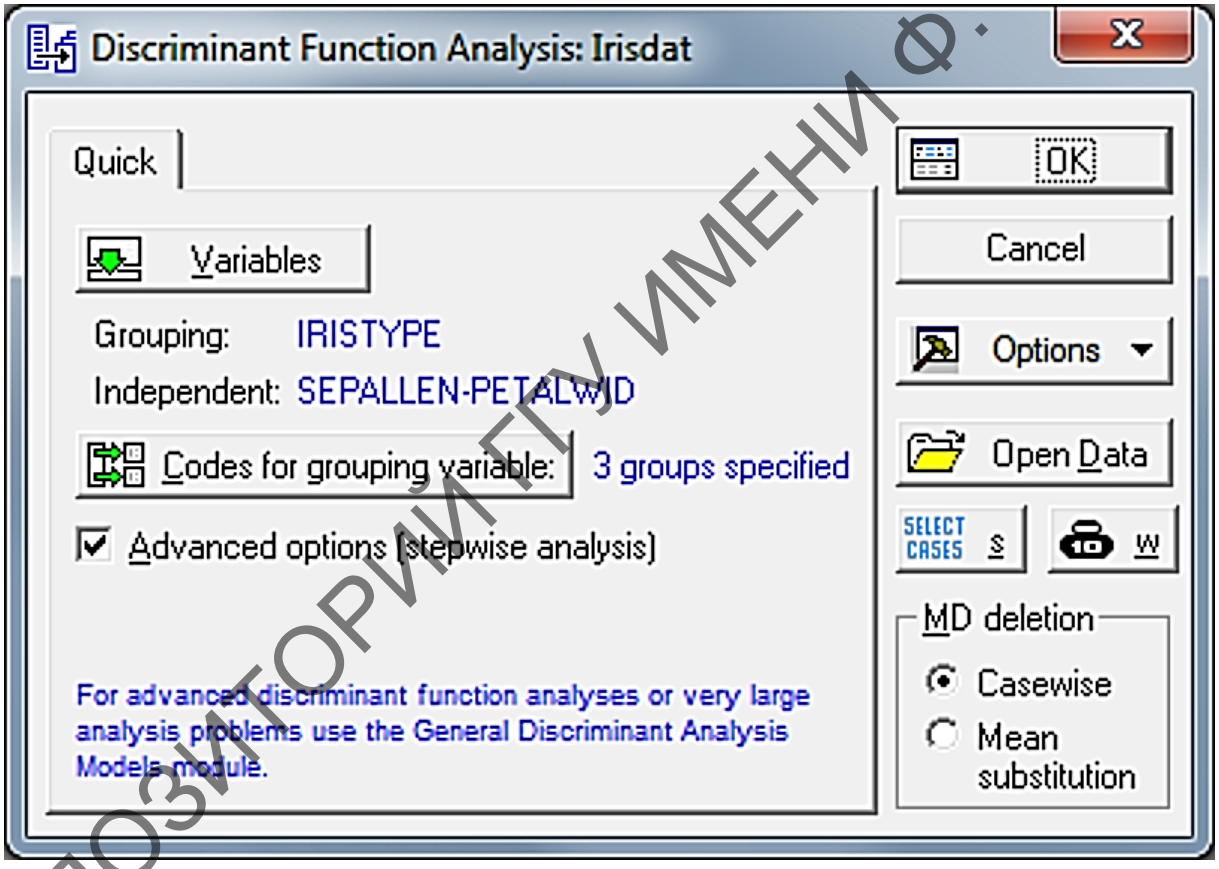


Рисунок 7.7 – Диалоговое окно модуля **Discriminant Analysis**, подготовленное для анализа

Шаг 5. Выбор способа проведения анализа.

После нажатия на кнопки **ОК** на предыдущем шаге выполнения анализа на экране появляется диалоговое окно выбора способа проведения дискриминантного анализа (модели) (рисунок 7.8).

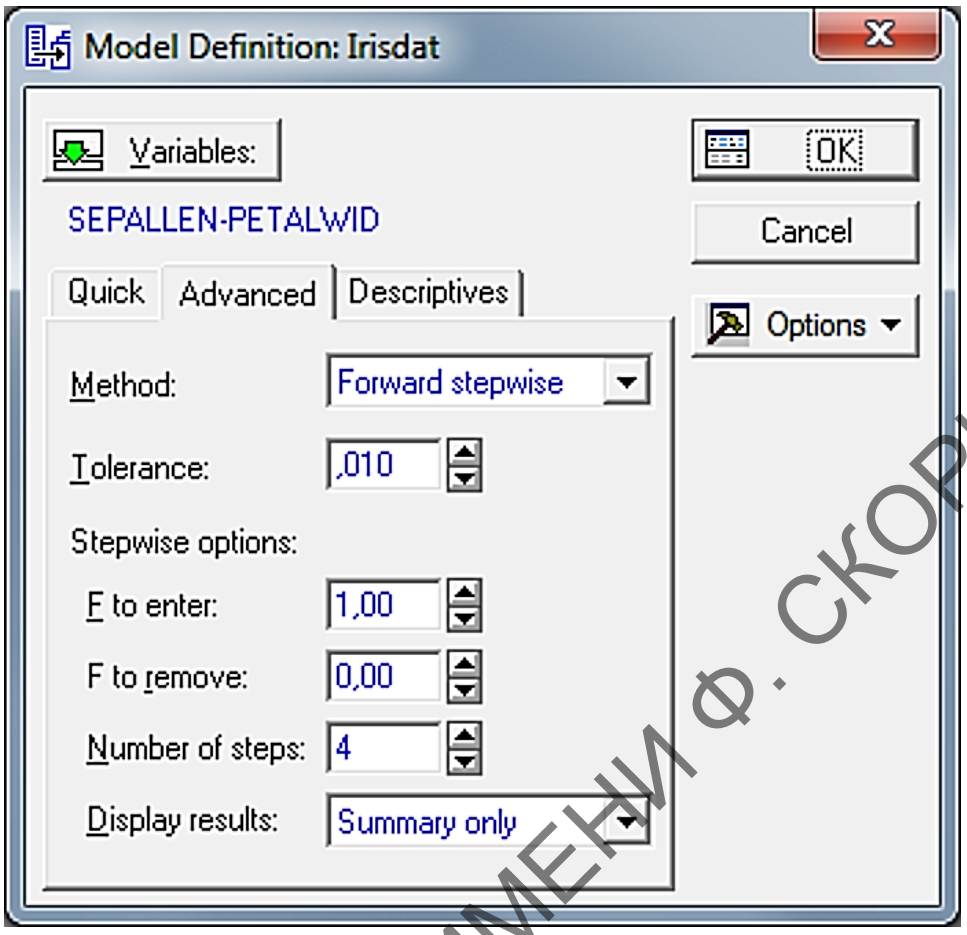


Рисунок 7.8 – Диалоговое окно модуля **Model Definition**

Для выбора метода необходимо перейти на закладку **Advanced** (*Расширенные настройки*) и обратиться к полю выбора с выпадающим списком напротив заголовка **Method** (*Метод*). Раскрывающийся список содержит следующие методы:

– **Standard** (*Стандартный*);
– **Forward stepwise** (*Пошаговый включения*);
– **Backward stepwise** (*Пошаговый исключения*).

В нашем случае нужно выбрать из списка **Forward stepwise** (*Пошаговый включения*), нажав на треугольник справа от поля выбора метода и указав его непосредственно в списке. Остальные настройки оставить без изменения (рисунок 7.8) и нажать левой клавишей мыши на кнопку **ОК**.

Шаг 6. Просмотр результатов.

После нажатия на кнопку **ОК** на предыдущем шаге выполнения анализа на экране появляется диалоговое окно результатов дискриминантного анализа (рисунок 7.9).

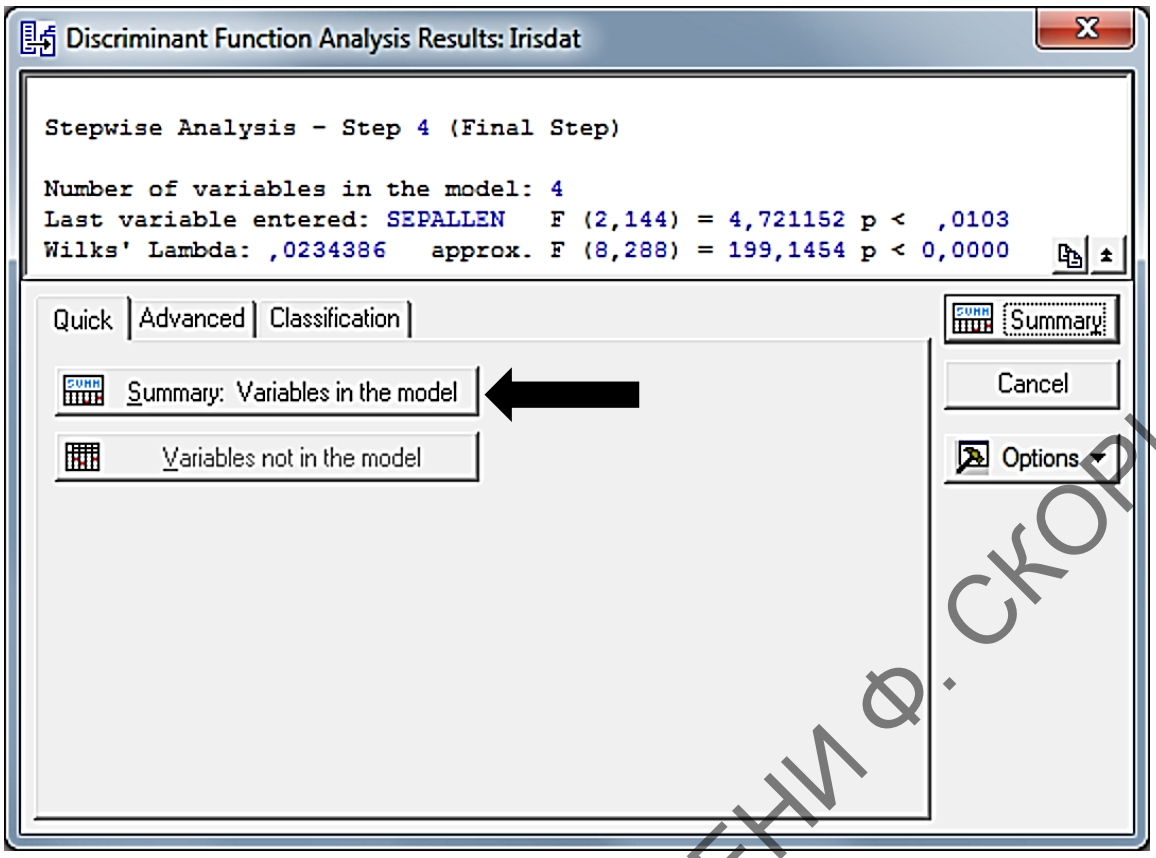


Рисунок 7.9 – Диалоговое окно результатов дискриминантного анализа

Информационная часть окна сообщает, что использован:

– **Stepwise analysis** (*Пошаговый анализ включения*);

– **Step 4 Final step** (*Шаг 4 – заключительный*);

– **Number of variables in the model** (*Число переменных в модели*): 4;

– **Last variable entered** (*Последняя включенная переменная*): SEPALLEN, соответствующее значение статистики F-критерия F (2,144) = = 4,72, уровень значимости $p < 0{,}0103$;

– **Wilks lambda** (*Значение лямбды Уилкса*): 0,0234386;

– **approx. F** (*Приближенное значение критерия F*), связанного с лямбдой Уилкса (8,288) = 199,1454; уровень значимости $p < 0{,}0000$;

– ***p*** *– уровень значимости F-критерия для значения 199,1454*.

Значения статистики лямбда Уилкса лежат в интервале [0, 1]. Значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации. Значения статистики Уилкса, лежащие около 1, свидетельствуют о плохой дискриминации.

Шаг 7. Просмотр переменных, включённых в модель.

Для просмотра переменных, которые в результате классификации были включены в модель, необходимо на закладке **Quick** (*Быстрые*

результаты) или **Advanced** (*Расширенные результаты*) нажать кнопку **Summary: Variables in the model** (*Результат: Переменные, включенные в модель*) (рисунок 7.9). На экране появится итоговая таблица анализа (рисунок 7.10).

N=150	Discriminant Function Analysis Summary (Irisdat) Step 4, N of vars in model: 4; Grouping: IRISTYPE (3 grps) Wilks' Lambda: ,02344 approx. F (8,288)=199,15 p<0,0000					
	Wilks' Lambda	Partial Lambda	F-remove (2,144)	p-level	Toler.	1-Toler. (R-Sqr.)
PETALLEN	0,035025	0,669206	35,59018	0,000000	0,365126	0,634874
SEPALWID	0,030580	0,766480	21,93593	0,000000	0,608859	0,391141
PETALWID	0,031546	0,743001	24,90433	0,000000	0,649314	0,350686
SEPALLEN	0,024976	0,938464	4,72115	0,010329	0,347993	0,652007

Рисунок 7.10 – Таблица переменных, включённых в модель

Таблица переменных, включённых в модель, имеет следующие столбцы:

– **Wilks Lambda** (*Значение лямбды Уилкса*);

– **Partial Lambda** (*Частичная лямбда*);

– F-remove (*Извлечённое значение критерия Фишера*);

– p-level (*Уровень значимости результата*);

– Toler. (*Значение толерантности*);

– 1-Toler. (R-Sqr.) (*Значение, обратное толерантности,* R^2).

В данном случае все классифицируемые переменные были включены в модель (уровень значимости гораздо выше минимального 0,05), благодаря чему подкрашены красным цветом.

Шаг 8. Графическая визуализация групп.

Для графической визуализации классифицированных групп необходимо в итоговом окне результатов (рисунок 7.9) перейти на закладку **Advanced** (*Расширенные результаты*) и нажать кнопку **Perform Canonical Analysis** (*Выполнение канонического анализа*), после чего будет отображено диалоговое окно канонического анализа (рисунок 7.11). В этом окне нужно нажать кнопку **Scatterplot of canonical scores** (*Диаграмма рассеяния канонических значений*) (рисунок 7.11). На экране появится итоговый график (рисунок 7.12).

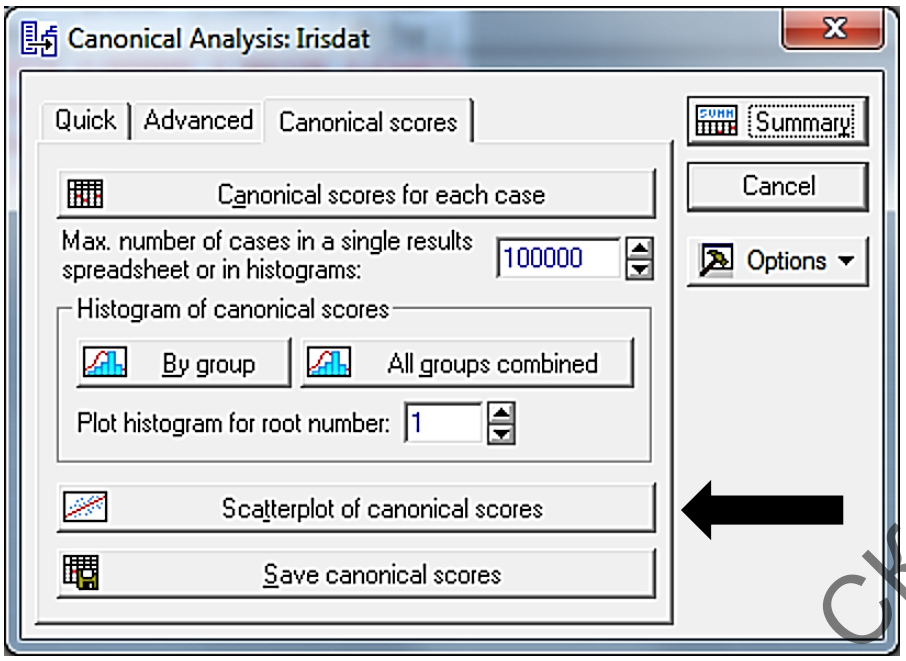


Рисунок 7.11 – Диалоговое окно модуля **Canonical Analysis**

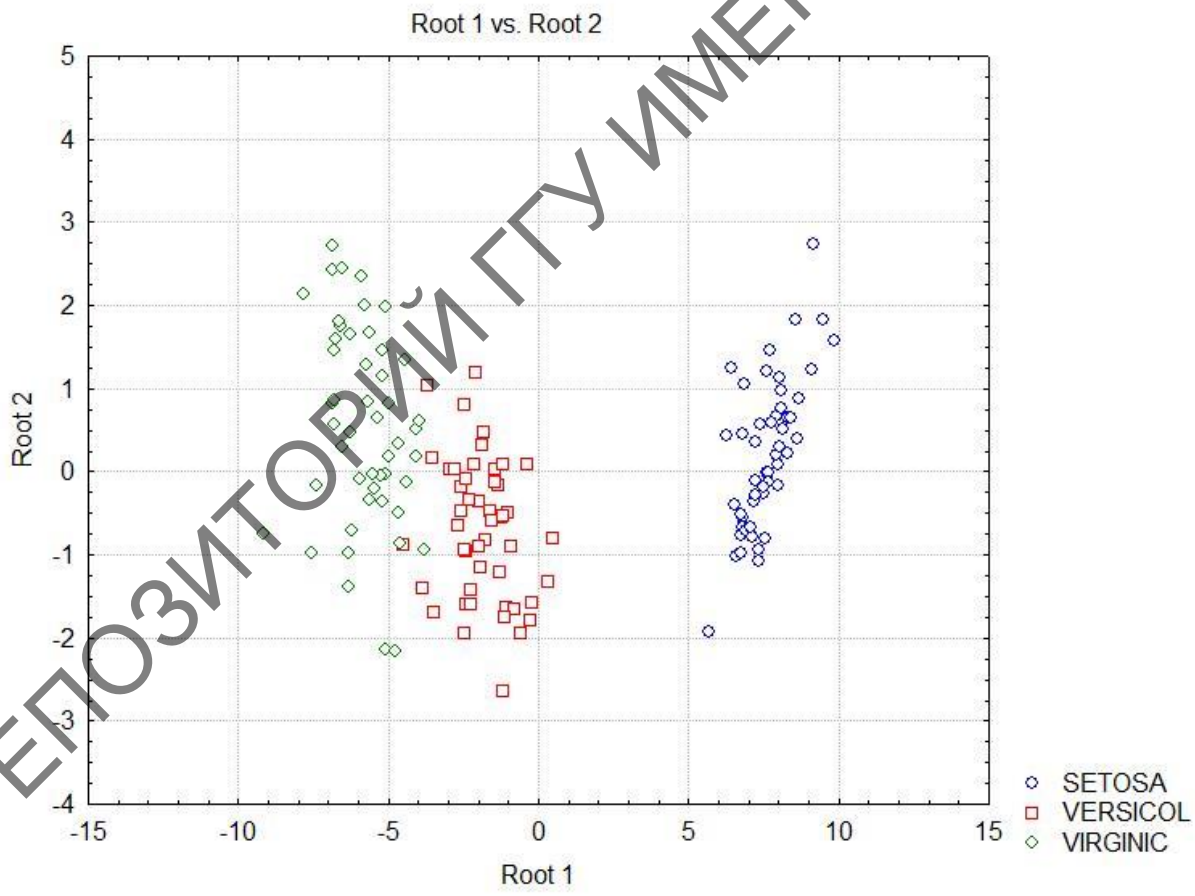


Рисунок 7.12 – График визуализации полученной классификации

Шаг 9. Расчёт функций классификации.

Для расчёта функций с целью соотнесения каждого из признаков к тому или иному сорту ириса необходимо в итоговом окне результатов (рисунок 7.9) перейти на закладку **Classification** (*Классификация*) и нажать кнопку **Classification functions** (*Функции классификации*) (рисунок 7.13).

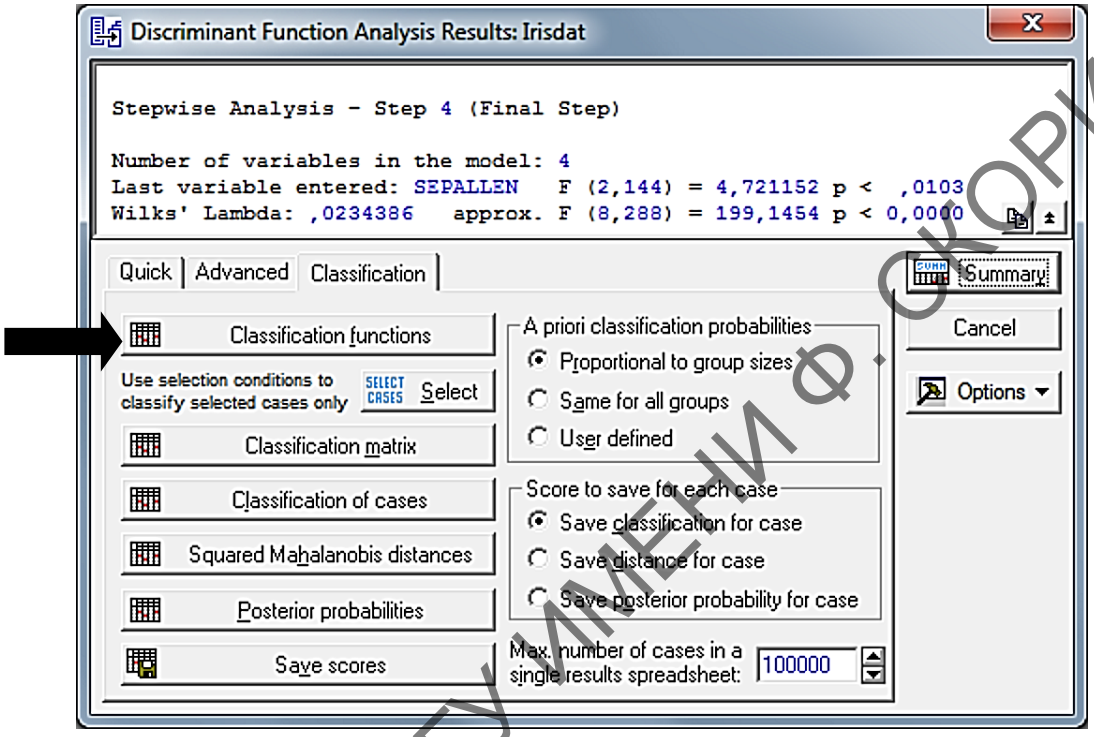


Рисунок 7.13 – Выбор опции **Classification functions**

В итоге появится таблица с коэффициентом к каждому аргументу функций (признаку) для получения значения функции (конкретного сорта) (рисунок 7.14).

Variable	Classification Functions; grouping: IRISTYPE (Irisdat)		
	SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
PETALLEN	-16,4306	5,2115	12,767
SEPALWID	23,5879	7,0725	3,685
PETALWID	-17,3984	6,4342	21,079
SEPALLEN	23,5442	15,6982	12,446
Constant	-86,3085	-72,8526	-104,368

Рисунок 7.14 – Итоговая таблица **Classification functions**

При помощи полученных функций мы можем составить классификационные уравнения для каждого из сортов:

SETOSA = – 16,43*PL + 23,69*SW – 17,4*PW + 23,54*SL – 86,31;

VERSICOL = 5,21* PL + 7,07* SW – 6,43* PW + 15,70* SL – 72,85;

VIRGINIC = 12,76* PL + 3,69* SW – 21,08* PW + 12,5* SL – 104,37,

где:

– PL – PETALLEN;

– SW – SEPALWID;

– PW– PETALWID;

– SL – SEPALLEN.

В дальнейшем при наличии нового цветка с новыми значениями: PETALLEN, SEPALWID, PETALWID, SEPALLEN. Для определения, к какому сорту цветов его отнести следует подставить эти значения в приведенные выше формулы и вычислить классификационные значения SETOSA, VERSICOL, VIRGINIC. Новый цветок будет относиться к тому сорту, для которого классификационное значение будет максимально.

Шаг 10. Расчёт расстояния Махаланобиса.

Для просмотра квадратов расстояния Махаланобиса от точек (случаев) до центров групп необходимо в диалоговом окне результатов анализа (рисунок 7.9) на закладке **Classification** (*Классификация*) нажать кнопку **Squared Mahalanobis distance** (*Квадрат расстояния Махаланобиса*) и вы увидите таблицу расстояний (рисунок 7.15).

Squared Mahalanobis Distances from Group Centroids (Irisdat)
Incorrect classifications are marked with *

Case	Observed Classif.	SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
1	SETOSA	0,2419	90,6602	181,5587
2	VIRGINIC	208,5713	27,3188	1,8944
3	VERSICOL	105,2663	2,2329	13,0720
4	VIRGINIC	207,9180	31,7492	4,4506
* 5	VIRGINIC	133,0668	5,2529	7,2359
6	SETOSA	1,3337	84,0118	170,0569
7	VIRGINIC	173,1838	26,5620	11,0484
8	VERSICOL	131,6617	8,4307	14,7647

Squared Mahalanobis Distances from Group Centroids (Irisdat)

Рисунок 7.15 – Итоговая таблица **Squared Mahalanobis distance**

Считается, что случай относится к группе, до которой расстояние Махаланобиса минимально.

Шаг 11. Расчёт вероятности принадлежности случая к группе.

Для проведения анализа подобных апостериорных вероятностей необходимо в диалоговом окне результатов анализа (рисунок 7.9) на закладке **Classification** (*Классификация*) нажать кнопку **Posterior probabilities** (*Апостериорные вероятности*) и вы увидите таблицу с апостериорными вероятностями принадлежности объекта к определенному классу (рисунок 7.16).

Posterior Probabilities (Irisdat)
Incorrect classifications are marked with *

Case	Observed Classif.	SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
1	SETOSA	1,000000	0,000000	0,000000
2	VIRGINIC	0,000000	0,000003	0,999997
3	VERSICOL	0,000000	0,995590	0,004410
4	VIRGINIC	0,000000	0,000001	0,999999
* 5	VIRGINIC	0,000000	0,729388	0,270612
6	SETOSA	1,000000	0,000000	0,000000
7	VIRGINIC	0,000000	0,000428	0,999572
8	VERSICOL	0,000000	0,959573	0,040427

Рисунок 7.16 – Итоговая таблица **Posterior probabilities**

Интерпретировать полученные результаты нужно следующим образом. В первом столбце указан сорт ириса для каждого случая. Во втором, третьем, четвертом столбцах даны апостериорные вероятности отнесения каждого конкретного цветка к определенному сорту: цветок относится к группе с максимальной апостериорной вероятностью.

Знаком * отмечаются неправильно классифицированные при использовании данного правила случаи (5, 9, 12).

Шаг 12. Классификация новых случаев.

Для изучения возможности добавления новых случаев в уже классифицированную таблицу необходимо закрыть окна дискриминантного анализа и добавить в таблицу исходных данных новый случай, указанный на рисунке 7.17. О добавлении новых случаев и переменных в таблицу см. тему 1.

Data: Irisdat* (5v by 151c)

	Fisher (1936) iris data: length & width of sepals and petals, 3				
	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
151	5,3	3,1	2,7	0,5	

Рисунок 7.17 – Новое наблюдение в данных **Iris.sta**

Для того чтобы понять, к какому классу относится этот объект, необходимо повторить шаги с 1 по 6 и затем в диалоговом окне результатов анализа (рисунок 7.9) на закладке **Classification** (*Классификация*) нажать кнопку **Posterior probabilities** (*Апостериорные вероятности*). После этого на экране появится та же таблица с постериорными вероятностями, к которой будет добавлена строка (рисунок 7.18).

Case	Posterior Probabilities (Irisdat) Incorrect classifications are marked with *			
	Observed Classif.	SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
141	VERSICOL	0,000000	0,999917	0,000083
142	VERSICOL	0,000000	0,998254	0,001746
143	SETOSA	1,000000	0,000000	0,000000
144	SETOSA	1,000000	0,000000	0,000000
145	SETOSA	1,000000	0,000000	0,000000
146	VIRGINIC	0,000000	0,000001	0,999999
147	VERSICOL	0,000000	0,689213	0,310787
148	VIRGINIC	0,000000	0,000000	1,000000
149	SETOSA	1,000000	0,000000	0,000000
150	VERSICOL	0,000000	1,000000	0,000000
151	---	0,999874	0,000126	0,000000

Рисунок 7.18 – Итоговая таблица **Posterior probabilities** с новыми данными

Согласно результату, новое наблюдение с вероятностью 0,999 можно отнести к сорту SETOSA.

Для подтверждения нашего предположения необходимо проверить квадраты расстояния Махаланобиса от точек (случаев) до центров групп, проведя порядок действий, описанный в шаге 10. В результате на экране появится таблица расстояний, содержащая новый случай (рисунок 7.19).

Case	Squared Mahalanobis Distances from Group Ce Incorrect classifications are marked with *			
	Observed Classif.	SETOSA p=,33333	VERSICOL p=,33333	VIRGINIC p=,33333
144	SETOSA	1,0069	92,3452	179,2350
145	SETOSA	0,3420	96,1764	188,2198
146	VIRGINIC	216,2140	37,6554	10,0253
147	VERSICOL	124,8026	4,6670	6,2599
148	VIRGINIC	245,7223	47,8027	10,2841
149	SETOSA	1,2454	105,2551	199,3038
150	VERSICOL	68,4668	5,4822	37,4651
151	---	16,2240	34,1747	97,2396

Рисунок 7.19 – Итоговая таблица **Squared Mahalanobis distance**

Расстояние от нового наблюдения до центра групп минимально именно для сорта SETOSA. Следовательно, с высокой степенью вероятности новый цветок – это ирис сорта SETOSA.

Задания

1) Создайте пустую электронную таблицу Spreadsheet.sta. Внесите данные для выполнения расчетов, находящиеся в таблице 7.1. Выполните процедуры дискриминантного анализа. Дайте объяснение полученным результатам.

Таблица 7.1

1-е задание					2-е задание			3-е задание		
Класс	1-й признак	2-й признак	3-й признак	4-й признак	Класс	1-й признак	2-й признак	Класс	1-й признак	2-й признак
a	1,01	0,48	0,87	1,12	*d*	0,247	0,295	*a*	9,23	5,26
a	1,14	1,11	1,38	1,17	*d*	0,491	0,495	*c*	19,82	14,00
a	1,22	1,44	1,02	0,79	*c*	0,768	0,240	*a*	11,59	8,06
a	1,09	0,94	0,91	1,1	*c*	0,838	0,354	*a*	11,06	6,99
a	0,65	0,89	1,09	1,12	*c*	0,921	0,320	*a*	7,58	2,83
a	1,21	1,22	1,11	1,4	*c*	0,837	0,409	*a*	9,35	5,73
a	0,95	1,04	0,76	1,15	*b*	0,642	0,309	*a*	10,27	6,24
b	2,01	1,24	0,78	0,8	*b*	0,754	0,288	*d*	25,95	20,80
b	1,64	0,93	0,96	1,48	*c*	0,844	0,340	*a*	11,04	6,83
b	2,13	1,38	1,15	0,93	*d*	0,433	0,242	*b*	16,03	11,53
a	1,41	0,85	1,15	0,73	*c*	0,923	0,289	*e*	31,41	26,44
a	1,22	1,03	1,1	0,87	*c*	0,824	0,243	*b*	16,92	11,96
b	2,81	1,13	1	0,65	*c*	0,963	0,335	*b*	16,02	11,47
b	2,05	1,25	0,91	0,88	*c*	0,800	0,223	*b*	17,74	12,96
b	2,31	0,87	1,14	1,37	*a*	0,828	0,696	*b*	17,00	11,96
b	1,8	0,84	0,94	0,89	*c*	1,000	0,366	*b*	18,36	13,36
a	1,15	1,27	1,25	1,16	*b*	0,715	0,354	*b*	15,73	11,25
a	0,74	2	0,86	0,85	*a*	0,885	0,573	*b*	17,98	13,18
a	0,89	2,05	0,9	0,87	*b*	0,658	0,278	*b*	17,33	12,11
a	0,68	1,39	0,71	0,99	*a*	0,739	0,632	*a*	12,62	8,28
a	0,94	2,07	1,09	0,78	*b*	0,684	0,357	*c*	21,71	16,34
a	1,06	1,92	1,01	0,82	*a*	0,684	0,592	*c*	22,13	16,88
a	0,76	2,27	1,09	1,17	*d*	0,512	0,428	*e*	30,27	26,10
a	1,12	1,86	0,9	1,26	*a*	0,763	0,777	*c*	20,92	15,50
a	1,11	1,9	1,33	0,93	*d*	0,452	0,339	*e*	30,17	25,93
a	1,03	1,25	0,91	1,02	*d*	0,468	0,428	*c*	22,83	17,16
a	1,57	2,1	0,94	1,09	*c*	0,849	0,298	*c*	21,02	15,63
a	1,05	0,56	1,82	0,95	*c*	0,761	0,252	*c*	21,84	16,83

Продолжение таблицы 7.1

1-е задание					2-е задание			3-е задание		
Класс	1-й признак	2-й признак	3-й признак	4-й признак	Класс	1-й признак	2-й признак	Класс	1-й признак	2-й признак
a	0,92	0,93	2,17	1,17	*d*	0,298	0,201	*e*	31,82	26,62
a	1,11	0,9	2,36	1,03	*c*	0,820	0,438	*d*	25,06	19,97
a	1,44	1,29	2,11	0,95	*c*	0,886	0,252	*d*	26,06	21,15
a	0,81	1,03	2,36	1,12	*a*	0,761	0,534	*c*	20,60	15,17
a	0,69	1,29	1,79	0,92	*c*	0,827	0,361	*d*	24,51	19,51
a	1,04	0,83	2,76	1,26	*a*	0,894	1,000	*e*	28,94	23,35
a	0,98	1,17	1,62	0,93	*d*	0,381	0,459	*d*	26,12	21,25
a	1,23	0,88	2,69	0,63	*b*	0,750	0,301	*d*	23,90	18,94
a	1,25	1,11	2,52	1,38	*a*	0,613	0,512	*d*	23,74	18,34
c	1,26	0,95	0,66	2,44	*d*	0,443	0,369	*d*	25,30	20,05
c	1,03	1,27	1,39	2,15	*c*	0,853	0,231	*c*	20,51	15,11

2) Из таблицы 7.2 внесите новые данные. Запустите процедуру дискриминантного анализа. Внесите поправки в исходную таблицу, пополнив обучающую выборку новой информацией.

Таблица 7.2

1-е задание					2-е задание			3-е задание		
Класс	1-й признак	2-й признак	3-й признак	4-й признак	Класс	1-й признак	2-й признак	Класс	1-й признак	2-й признак
	1,14	1,26	0,99	2,06		0,738	0,658		36,63	31,29
	0,79	0,84	1,17	2,72		0,612	0,243		24,84	19,63
	1,01	1,16	1,06	1,4		0,774	0,233		17,78	13,00
	0,97	1,11	0,73	0,98		0,933	0,271		5,17	1,92

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

ТЕМА 8. КЛАСТЕРНЫЙ АНАЛИЗ В СИСТЕМЕ STATISTICA 7.0

8.1 Краткая характеристика кластерного анализа.
8.2 Реализация кластерного анализа в STATISTICA 7.0.

8.1 Краткая характеристика кластерного анализа

Кластерный анализ объединяет различные процедуры, используемые для проведения классификации и является методом классификации без обучения (автоматическая классификация). В результате применения этих процедур исходная совокупность объектов разделяется на кластеры или группы (классы) схожих между собой объектов.

Кластер – это группа объектов, обладающих свойством плотности (плотность объектов внутри кластера выше, чем вне его), дисперсией, отделимостью от других кластеров, формой и размером.

Наиболее часто методы кластерного анализа используются в социологии, маркетинговых исследованиях, экономике, биологии, медицине, археологии.

В программном пакете STATISTICA доступны следующие меры сходства объектов:

– евклидова метрика;
– квадрат евклидовой метрики;
– манхэттенское расстояние, или «расстояние городских кварталов»;
– метрика Чебышева;
– метрика Минковского;
– обратный коэффициент корреляции Пирсона;
– обратный коэффициент совстречаемости.

В STATISTICA также реализованы следующие методы кластеризации:

а) агломеративные методы:
– *joining* (*tree clustering*);
– *two-way joining*;
б) метод k-средних (*k-means clustering*).

Обычно перед началом классификации данные стандартизуются (вычитается среднее и производится деление на корень квадратный из дисперсии). Полученные в результате стандартизации переменные имеют нулевое среднее и единичную дисперсию.

Кроме того, в STATISTICA можно выбрать следующие правила иерархического объединения кластеров:

− *Single linkage* – метод одиночной связи;

− *Complete linkage* – метод полной связи;

− *Unweighted pair group average* – невзвешенный метод «средней связи»;

− *Weighted pair group average* – взвешенный метод «средней связи»;

− *Weighted centroid pair group* (*median*) – взвешенный центроидный метод;

− *Ward method* – метод Уорда.

8.2 Реализация кластерного анализа в STATISTICA 7.0.

Знакомство с возможностями проведения кластерного анализа в программном пакете STATISTICA лучше всего начать с разбора уже апробированного примера, содержащегося в файле *Cars.sta*. Таким примером может являться совокупность автомобилей различных марок и технических характеристик.

Всего в файле содержатся данные о 22 машинах разных марок. Марки машин – это случаи.

Переменные в этом файле:

– PRICE – цена;

– ACCELERATION, BRAKING, HANDLING – технические характеристики;

– MILAGE – расход горючего (количество миль, пройденных на одном галлоне бензина).

Все характеристики машин уже стандартизованы (из значений переменной PRICE вычтена средняя цена, и разность поделена на корень квадратный из дисперсии).

Задача состоит в том, чтобы разбить автомобили на несколько групп, в которых они мало отличаются друг от друга (существенно меньше, чем в целом в совокупности). Разбив машины на группы, можно лучше в целом представить их совокупность, с тем, чтобы затем более обоснованно принимать решение, например при покупке или обмене одной машины на другую.

Шаг 1. Открытие электронной таблицы с данными.

Для открытия готовой электронной таблицы примера необходимо сначала в главном меню программы последовательно нажать на пункт меню **File** (*Файл*), а далее – **Open** (*Открыть*) и в открывшемся диалоговом окне выбрать нужный нам файл Cars.sta (рисунок 8.1), ко-

торый затем откроется как новая стандартная электронная таблица пакета STATISTICA (рисунок 8.2).

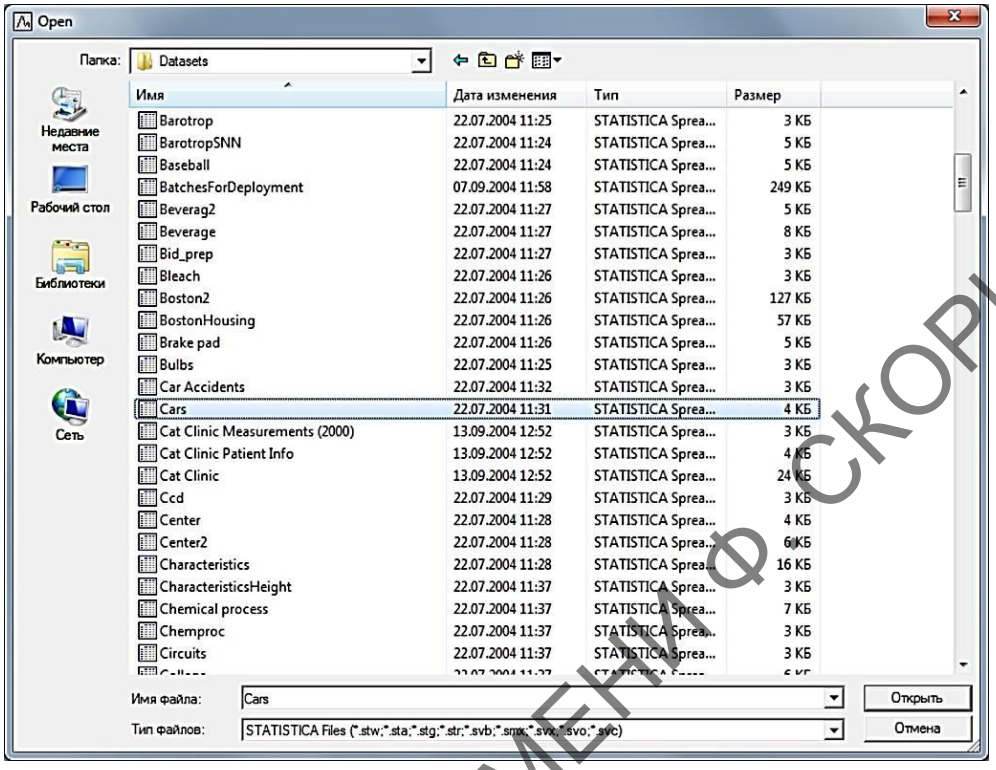


Рисунок 8.1 – Диалоговое окно **Open** в пакете STATISTICA

Data: Cars (5v by 22c)

	Performance, fuel economy, and approximate price for various automobiles				
	1 PRICE	2 ACCELERATION	3 BRAKING	4 HANDLING	5 MILEAGE
Acura	-0,521	0,477	-0,007	0,382	2,079
Audi	0,866	0,208	0,319	-0,091	-0,677
BMW	0,496	-0,802	0,192	-0,091	-0,154
Buick	-0,614	1,689	0,933	-0,210	-0,154
Corvette	1,235	-1,811	-0,494	0,973	-0,677
Chrysler	-0,614	0,073	0,427	-0,210	-0,154
Dodge	-0,706	-0,196	0,481	0,145	-0,154
Eagle	-0,614	1,218	-4,199	-0,210	-0,677
Ford	-0,706	-1,542	0,987	0,145	-1,724
Honda	-0,429	0,410	-0,007	0,027	0,369
Isuzu	-0,798	0,410	-0,061	-4,230	1,067
Mazda	0,126	0,679	-0,133	0,500	-1,724
Mercedes	1,051	0,006	0,120	-0,091	-0,154
Mitsub.	-0,614	-1,003	0,084	0,382	0,718
Nissan	-0,429	0,073	-0,007	0,263	0,997
Olds	-0,614	-0,734	0,409	0,382	2,114
Pontiac	-0,614	0,679	0,536	0,145	0,195
Porsche	3,454	-2,215	-0,296	0,618	-1,026
Saab	0,588	0,679	0,246	0,263	0,021
Toyota	-0,059	1,218	0,228	0,736	-0,851
VW	-0,706	-0,128	0,102	0,382	0,195
Volvo	0,219	0,612	0,138	-0,210	0,369

Рисунок 8.2 – Содержимое файла Cars.sta

Шаг 2. Выбор анализа.

В главном меню программы STATISTICA необходимо выбрать пункт **Statistics** (*Статистические процедуры*), в нём – модуль **Multivariate Exploratory Techniqes** (*Многомерные поисковые методы*), а затем – **Cluster Analysis** (*Кластерный анализ*) (рисунок 8.3) и нажать **ОК**.

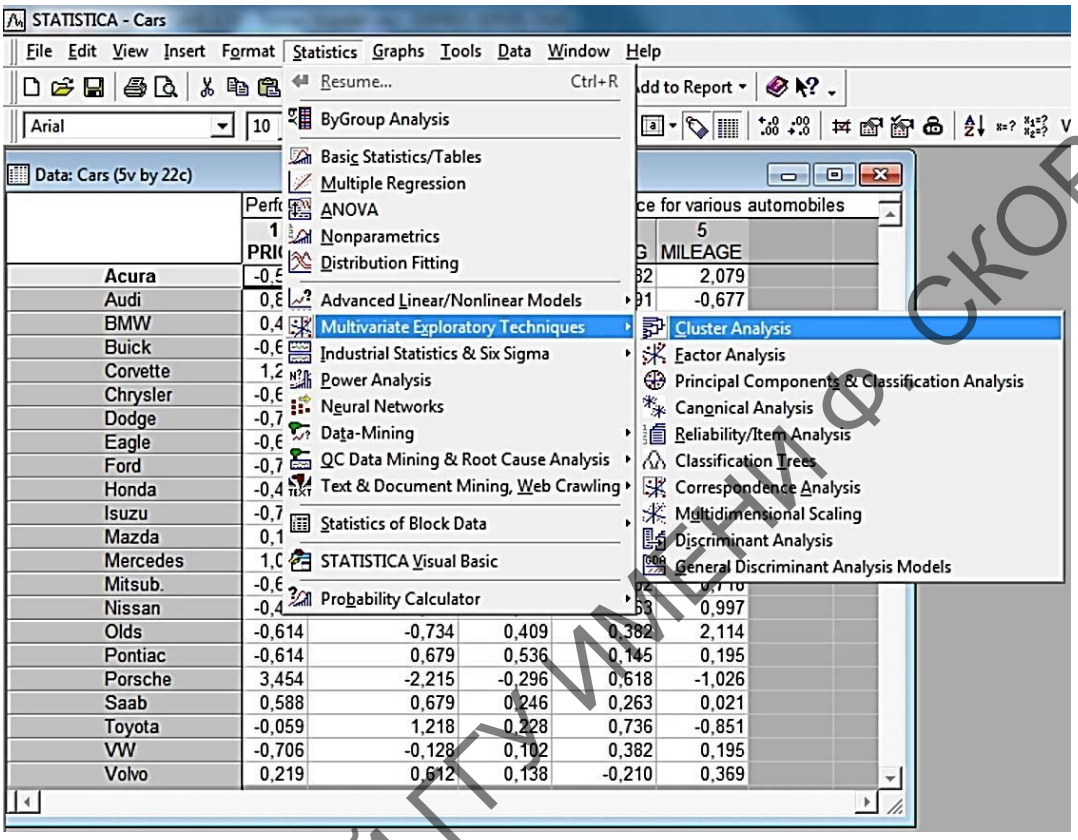


Рисунок 8.3 – Расположение модуля **Cluster Analysis** в меню пакета STATISTICA

Шаг 3. Выбор вида кластерного анализа.

После операций, произведённых на предыдущем шаге, на экране появится окно выбора кластерного анализа (рисунок 8.4).

Рисунок 8.4 – Стартовая панель модуля **Cluster Analysis**

Для проведения дальнейшего анализа необходимо определиться с тем, какой вид кластерного анализа нужно провести в зависимости от поставленных целей, выделить его в поле мышью и нажать **ОК**.

8.2.1 Проведение кластерного анализа методом k-средних

Шаг 1. Выбор вида кластерного анализа.

После вызова стартовой панели модуля кластерного анализа в списке видов анализа укажите мышью пункт **k-means** (*k-средних*) (рисунок 8.4) и нажмите кнопку **OK.** Диалоговое окно метода k-means появится на экране (рисунок 8.5).

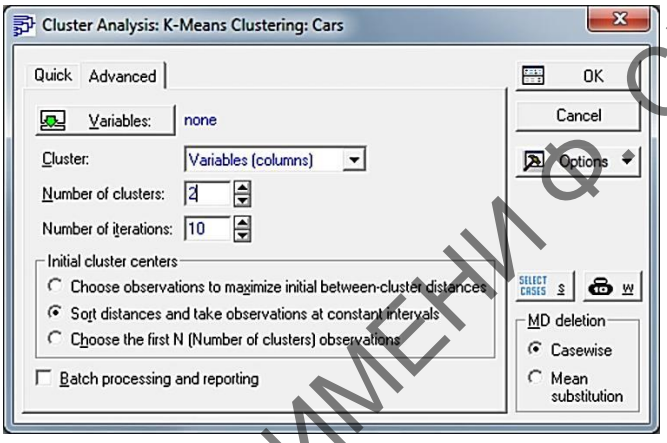


Рисунок 8.5 – Диалоговое окно метода анализа **k-means**

Шаг 2. Выбор переменных для анализа.

Для выбора нужных переменных необходимо нажать кнопку **Variables** (*Переменные*) и, таким образом, откроется диалоговое окно **Select variable for the analysis** (*Выбор переменных для анализа*) (рисунок 8.6).

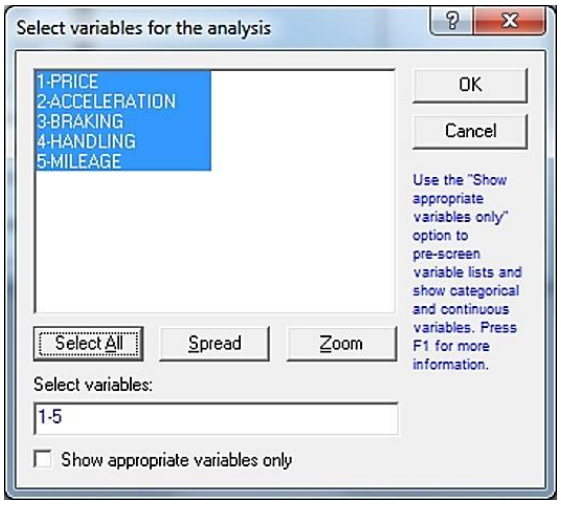


Рисунок 8.6 – Выбор переменных для кластерного анализа

Так как машины разбиты на группы и учитываются все параметры, то необходимо выбрать все переменные – нажать кнопку **Select All** (*Выбрать все*) (рисунок 8.6), а затем – кнопку **OK**. Программа вернётся в предыдущее окно (рисунок 8.5).

Шаг 3. Выбор параметров для анализа.

В поле **Cluster** (*Кластер*) необходимо нажать на треугольную стрелку рядом с полем и выбрать пункт **Cases** (*Случаи*). Альтернативный выбор был бы **Variables** (*Переменные*); им следует пользоваться, если нужно кластеризировать переменные.

В приведённом примере кластеризируются машины, которые являются случаями в исходном файле данных, поэтому и выбирается пункт **Cases** (*Случаи*).

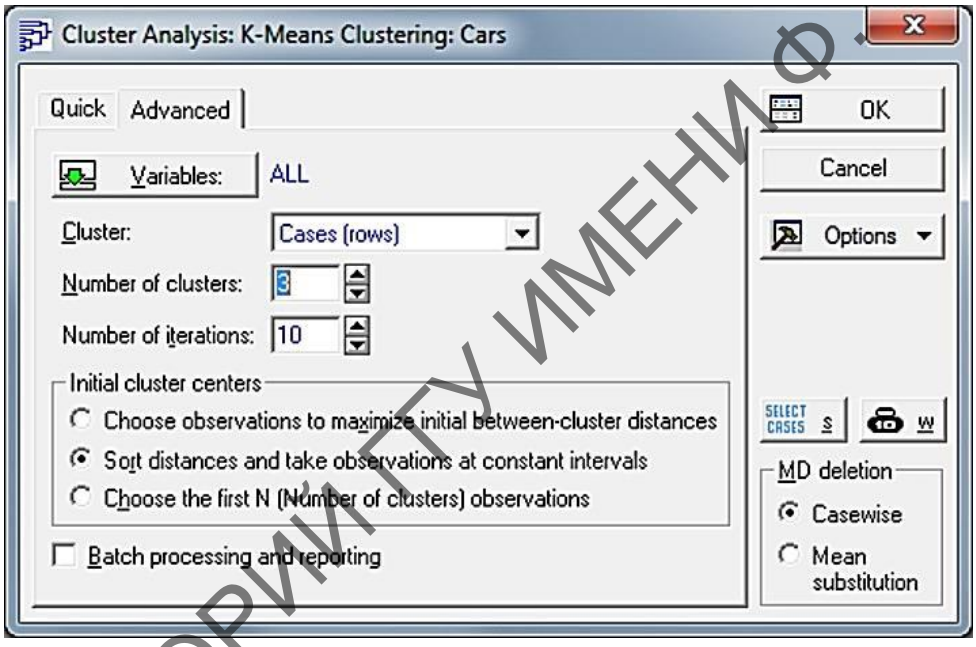


Рисунок 8.7 – Диалоговое окно метода анализа k-means, подготовленное для анализа

В поле **Number of clusters** (*Число кластеров*) нужно определить число групп, на которые необходимо разбить автомобили. Необходимо указать в этом поле число 3 (таким образом, мы предполагаем агрегацию машин в 3 кластера).

В строке **Number of iterations** (*Число итераций*) задается максимальное число итераций, используемых при построении классов. По умолчанию оно равно 10. Можно оставить это значение.

В строке **Missing data** (*Пропущенные данные*) задается способ обработки пропущенных значений в данных (например, для какой-то машины отсутствует значение некоторого параметра). В этом приме-

ре пропусков в данных нет и обработка пропущенных значений не происходит.

Группа опций **Initial cluster centers** (*Начальные центры кластеров*) позволяет задать начальные центры кластеров:

– **Choose observations to maximize initial between-cluster distance** (*Выберите наблюдения, чтобы максимизировать начальное расстояние между кластерами*);

– **Sort distances and take observations at constant intervals** (*Сортировка расстояний и наблюдение с постоянными интервалами*);

– **Choose the first N (Number of clusters) observation** (*Выберите первое наблюдение N* (*количество кластеров*)).

Необходимо указать все параметры так, как показано на рисунке 8.7, нажать кнопку **ОК** и запустить вычислительную процедуру.

Шаг 4. Просмотр результатов анализа.

В результате вычисления на экране появится окно результатов (рисунок 8.8).

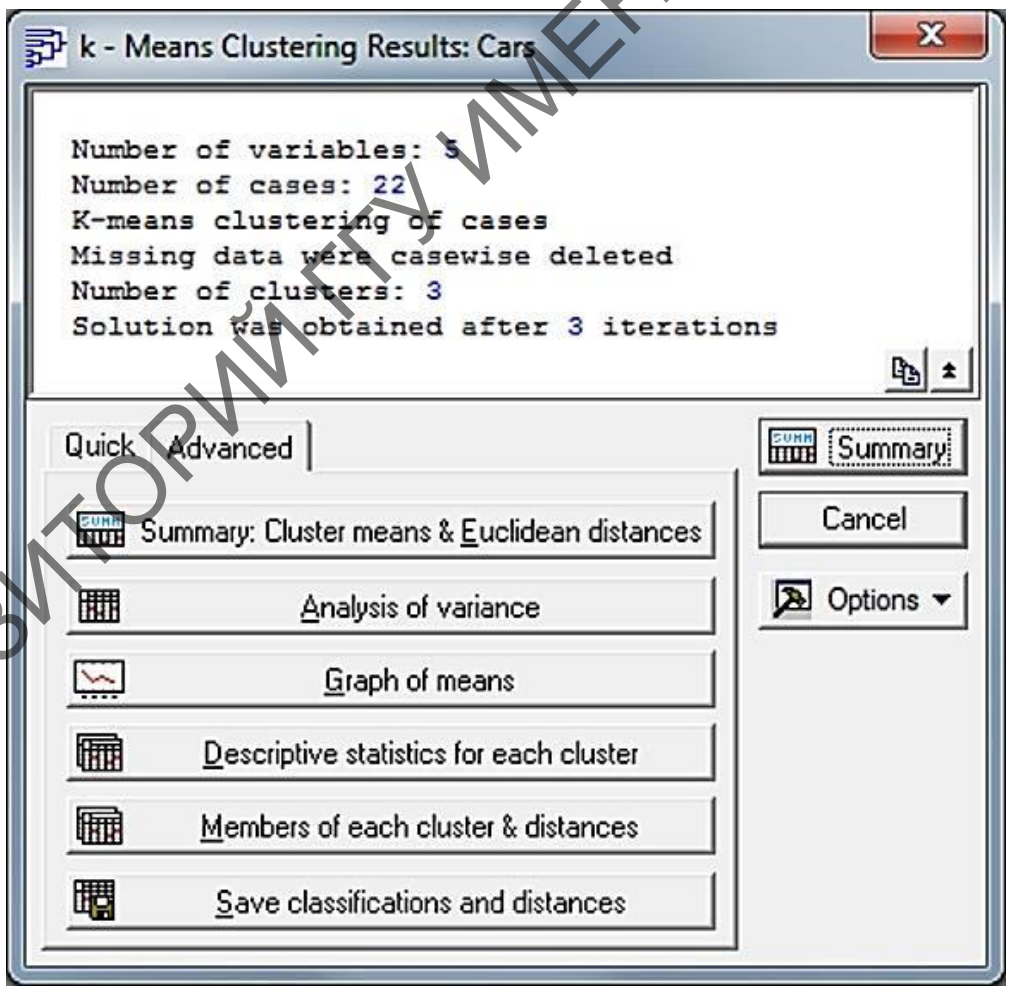


Рисунок 8.8 – Окно результатов кластеризации машин по методу k-means

В верхней части окна записана информация:

– **Number of variables** (*Число переменных*);

– **Number of cases** (*Число случаев*);

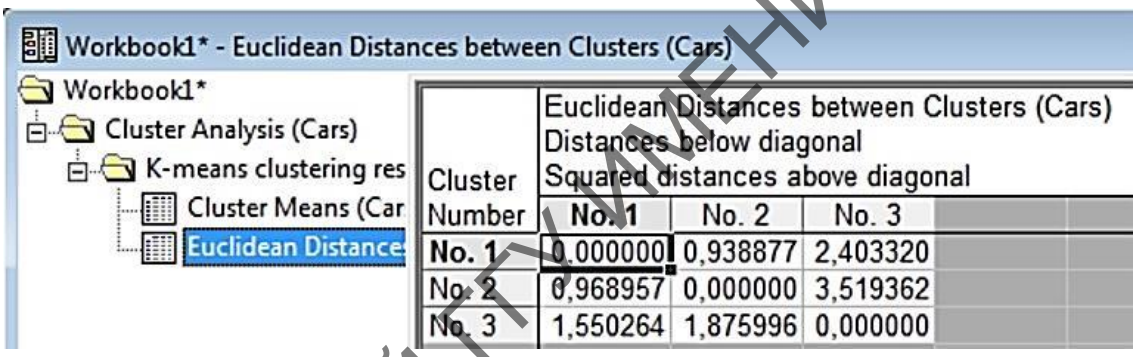
– **K-means clustering of cases** (*Кластеризация случаев методом k-средних*);

– **Number of clusters** (*Число кластеров*);

– **Solution was obtained after 2 iterations** (*Решение найдено после 2 итераций*).

Кнопки в нижней части окна на закладке **Advanced** (*Расширенные результаты*) позволяют провести подробный анализ результатов кластеризации.

Кнопка **Cluster Means&Euclidean Distances** (*Средние кластеров и евклидово расстояние*) позволяет вывести таблицы, в первой из которых указаны средние для каждого кластера (усреднение производится внутри кластера), во второй – евклидовы расстояния и квадраты евклидовых расстояний между кластерами (рисунок 8.9).



Euclidean Distances between Clusters (Cars)
Distances below diagonal
Squared distances above diagonal

Cluster Number	No. 1	No. 2	No. 3
No. 1	0,000000	0,938877	2,403320
No. 2	0,968957	0,000000	3,519362
No. 3	1,550264	1,875996	0,000000

Рисунок 8.9 – Расстояния между кластерами

В таблице даны евклидовы расстояния между средними кластеров (по каждому из параметров внутри кластера вычисляется среднее, получаются 3 точки в пятимерном пространстве, и между ними находится расстояние).

Исходя из данных, отражённых в таблице, видно, что расстояния между кластерами даны под диагональю – между 1 и 2 кластером 0,968957, а, например, между вторым и третьим – 1,875996. Над диагональю в таблице даны квадраты расстояний между кластерами.

Кнопка **Analysis of variation** (*Дисперсионный анализ*) позволяет просмотреть таблицу дисперсионного анализа.

Кнопка **Graph of means** (*Графики средних*) позволяет посмотреть средние значения для каждого кластера на линейном графике (рисунок 8.10).

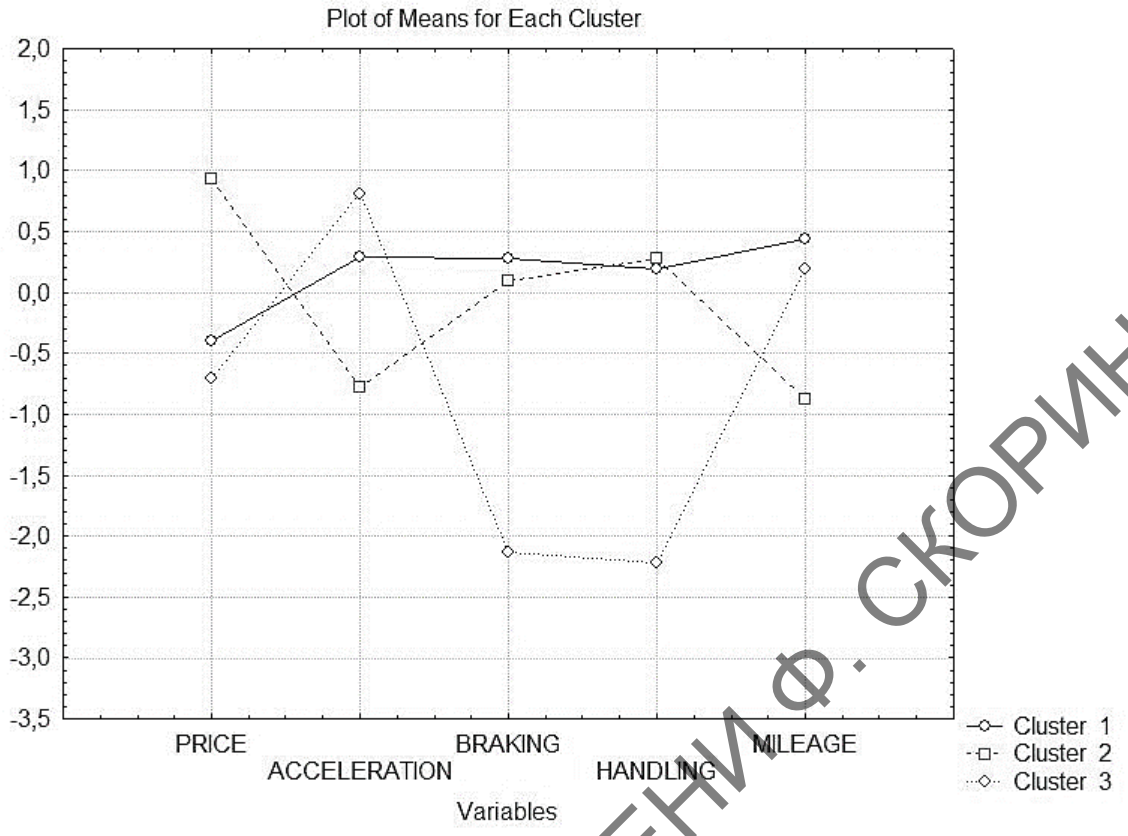
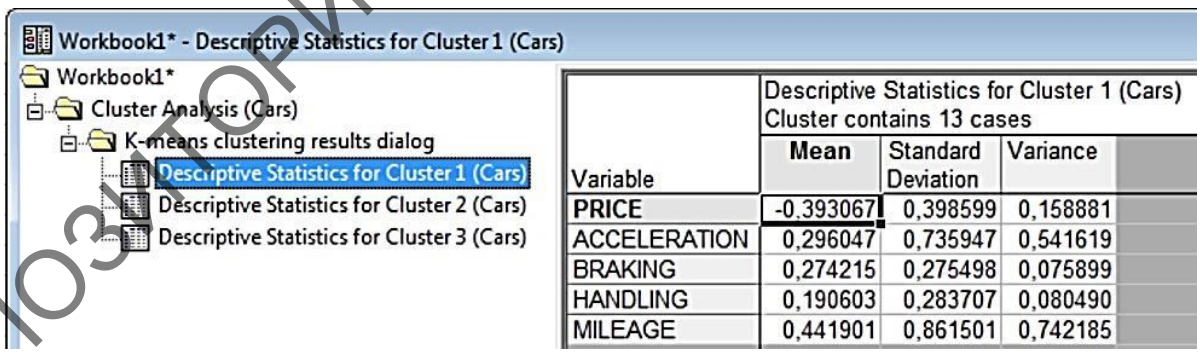


Рисунок 8.10 – График средних для каждого кластера

Кнопка **Descriptive Statistics for each clusters** (*Описательная статистика каждого кластера*) открывает электронную таблицу с описательными статистиками для каждого кластера (среднее, дисперсия и т. д.) (рисунок 8.11).

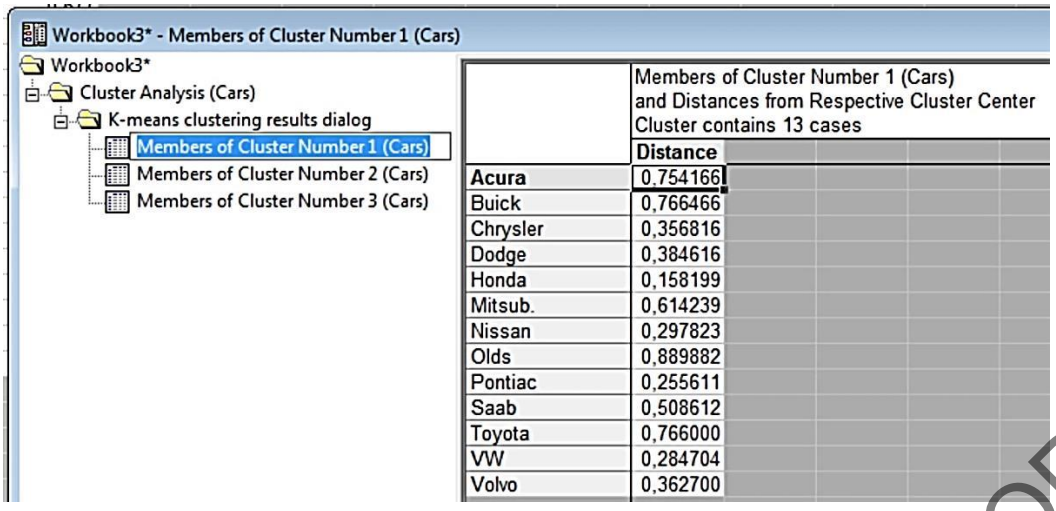


Descriptive Statistics for Cluster 1 (Cars)
Cluster contains 13 cases

Variable	Mean	Standard Deviation	Variance
PRICE	-0,393067	0,398599	0,158881
ACCELERATION	0,296047	0,735947	0,541619
BRAKING	0,274215	0,275498	0,075899
HANDLING	0,190603	0,283707	0,080490
MILEAGE	0,441901	0,861501	0,742185

Рисунок 8.11 – Описательная статистика для каждого кластера

Кнопка **Member of each cluster&distances** (*Члены каждого кластера и расстояния*) выдаёт таблицы с членами каждого из кластеров с расстояниями этих составляющих его членов до центра кластера (рисунок 8.12).



Members of Cluster Number 1 (Cars) and Distances from Respective Cluster Center
Cluster contains 13 cases

	Distance
Acura	0,754166
Buick	0,766466
Chrysler	0,356816
Dodge	0,384616
Honda	0,158199
Mitsub.	0,614239
Nissan	0,297823
Olds	0,889882
Pontiac	0,255611
Saab	0,508612
Toyota	0,766000
VW	0,284704
Volvo	0,362700

Рисунок 8.12 – Члены кластеров и расстояния от них до центра кластера

В столбцах таблиц указано расстояние от каждой машины до центра кластера.

Кнопка **Save classifications and distances** (*Сохранить классификации и расстояния*) позволяет сохранить результаты классификации в файле STATISTICA для дальнейшего исследования.

8.2.2 Проведение дендрограммного кластерного анализа

Шаг 1. Выбор вида кластерного анализа.

После вызова стартовой панели модуля кластерного анализа в списке видов анализа укажите мышью пункт **Joining (tree clustering)** (*Объединение* (*древовидная кластеризация*)) (рисунок 8.13) и нажмите кнопку **OK.** Диалоговое окно метода **Joining (tree clustering)** появится на экране (рисунок 8.14).

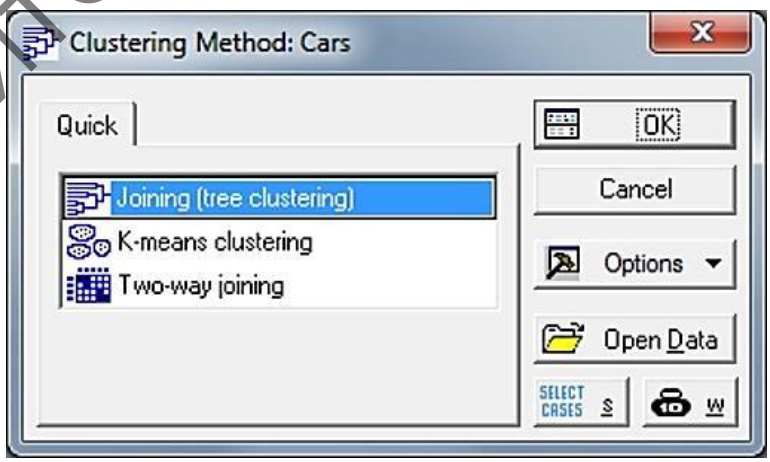


Рисунок 8.13 – Стартовая панель модуля **Cluster Analysis**

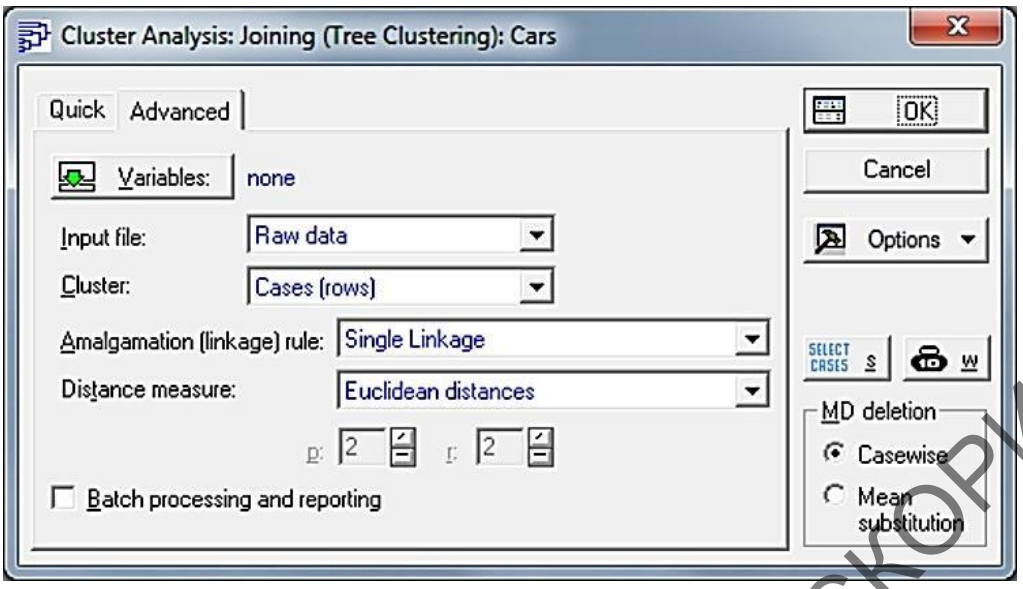


Рисунок 8.14 – Диалоговое окно метода анализа k-means

Шаг 2. Выбор переменных для анализа.

Для выбора нужных переменных необходимо нажать кнопку **Variables** (*Переменные*), и, таким образом, откроется диалоговое окно **Select variable for the analysis** (*Выбор переменных для анализа*) (рисунок 8.6).

Так как машины разбиты на группы и учитываются все параметры, то необходимо выбрать все переменные – нажать кнопку **Select All** (*Выбрать все*) (рисунок 8.15), а затем – кнопку **OK**. Программа вернётся в предыдущее окно (рисунок 8.13).

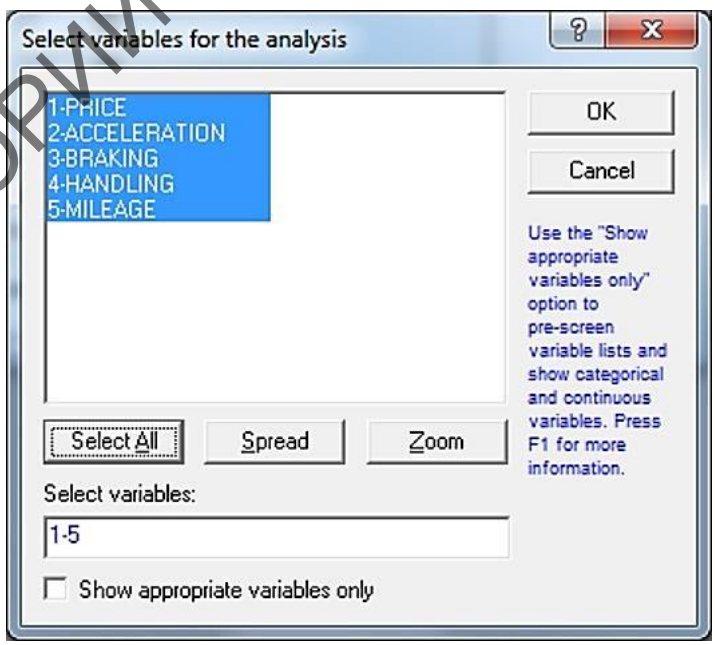


Рисунок 8.15 – Выбор переменных для кластерного анализа

Шаг 3. Выбор параметров для анализа.

В поле **Input file** (*Входные данные*) необходимо нажать на треугольную стрелку рядом с полем и выбрать пункт выбрать **Raw data** (*Ряды данных*); альтернатива – **Distance matrix** (*Матрица расстояний*) (рисунок 8.16).

В приведённом примере кластеризируются машины, которые являются случаями в исходном файле данных, поэтому в поле **Cluster** (*Кластер*) необходимо выбрать пункт **Cases (rows)** (*Случаи* (*ряды*)).

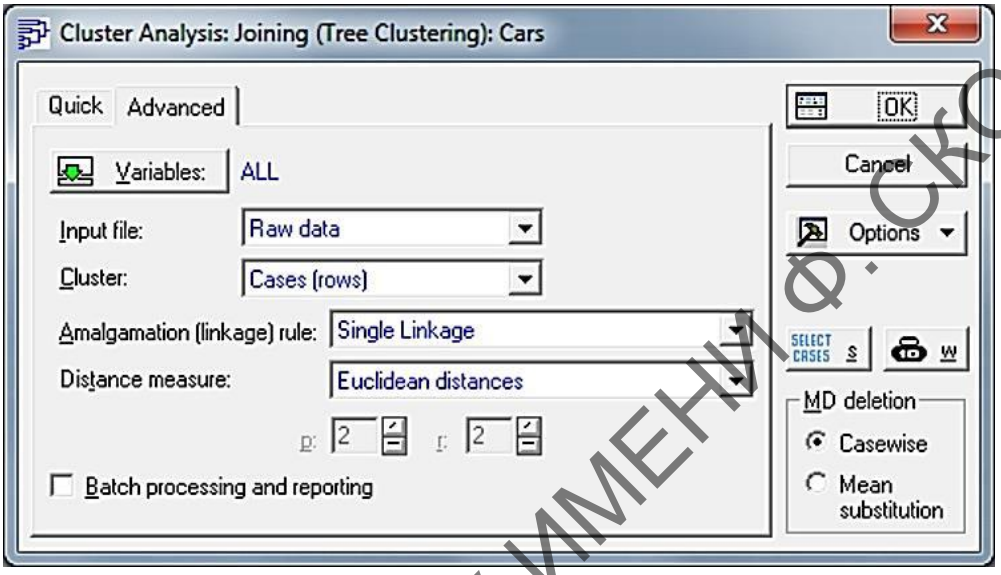


Рисунок 8.16 – Диалоговое окно метода анализа Joining (tree clustering), подготовленное для анализа

Ниже в окне имеются ещё 2 поля, которые нам понадобятся:

– **Amalgamation (linkage) rule** (*Правило объединения* (*связь*));

– **Distance measure** (*Измерение расстояния*).

В первом случае необходимо для наших целей выбрать опцию **Single linkage** (*Метод одиночной связи*), а во втором – **Euclidean distance** (*Евклидово расстояние*), после чего нажать **ОК** и запустить вычислительную процедуру.

Шаг 4. Просмотр результатов анализа.

В результате вычисления на экране появится окно результатов (рисунок 8.17). Верхняя часть окна содержит описательные характеристики анализа: количество вариант (5), количество случаев (22), что было объединено в кластеры (случаи), правило объединения (одиночная связь) и мера расстояния (евклидова метрика). Ниже, во второй части окна, представлены расчёты по кластерам.

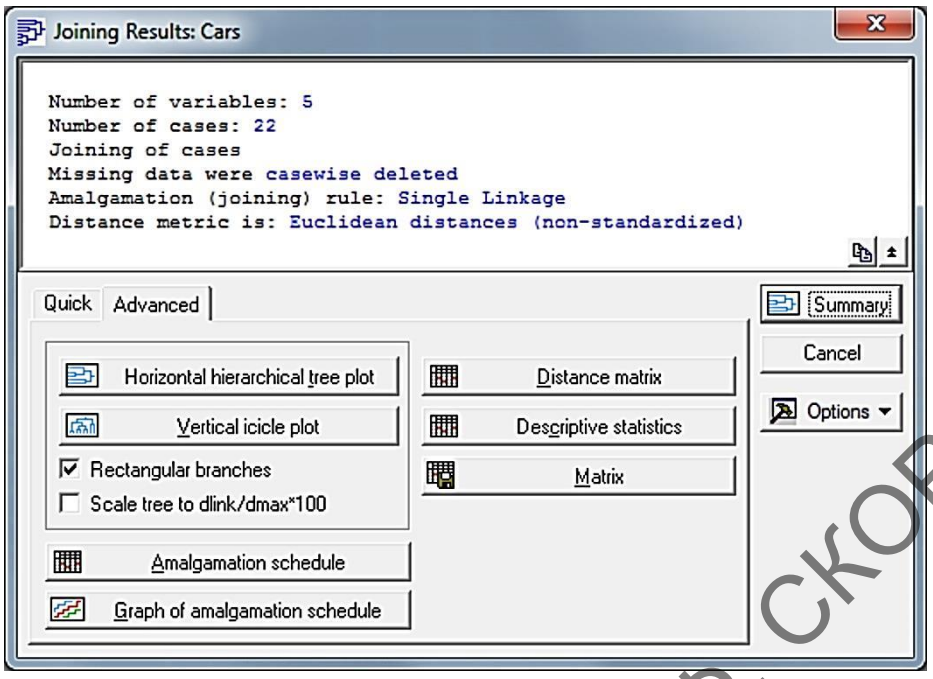


Рисунок 8.17 – Окно результатов кластеризации машин
по методу Joining (tree clustering)

Шаг 5. Графическое отображение результатов**.**

Для просмотра дендрограммы необходимо определиться с видом отображения данных: горизонтальным (рисунок 8.18) или вертикальным (рисунок 8.19).

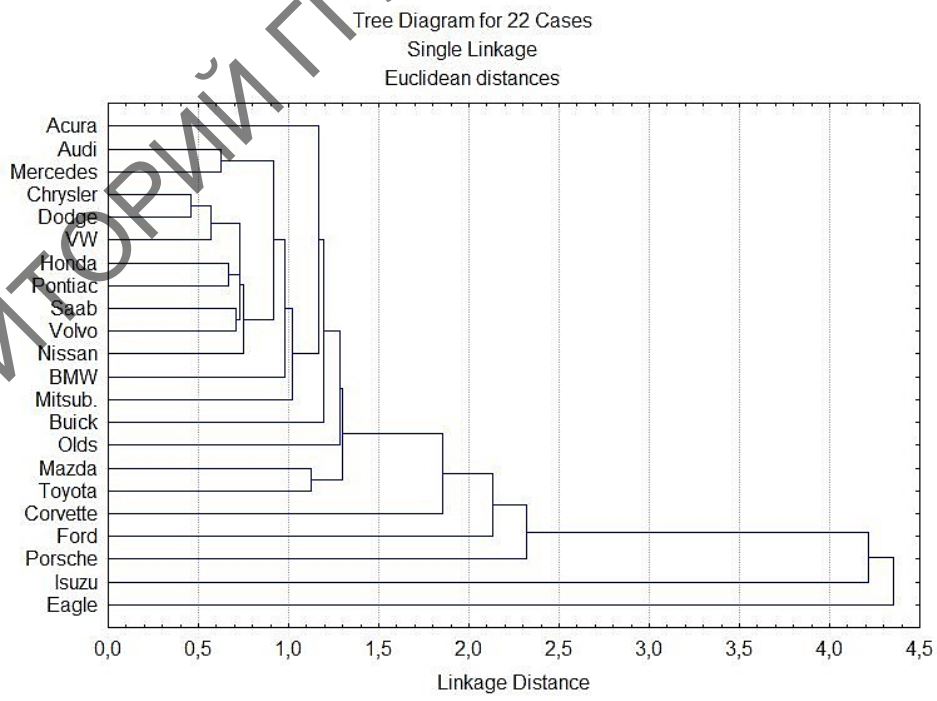


Рисунок 8.18 – Горизонтальное отображение дендрограммы сходства машин
по методу Joining (tree clustering)

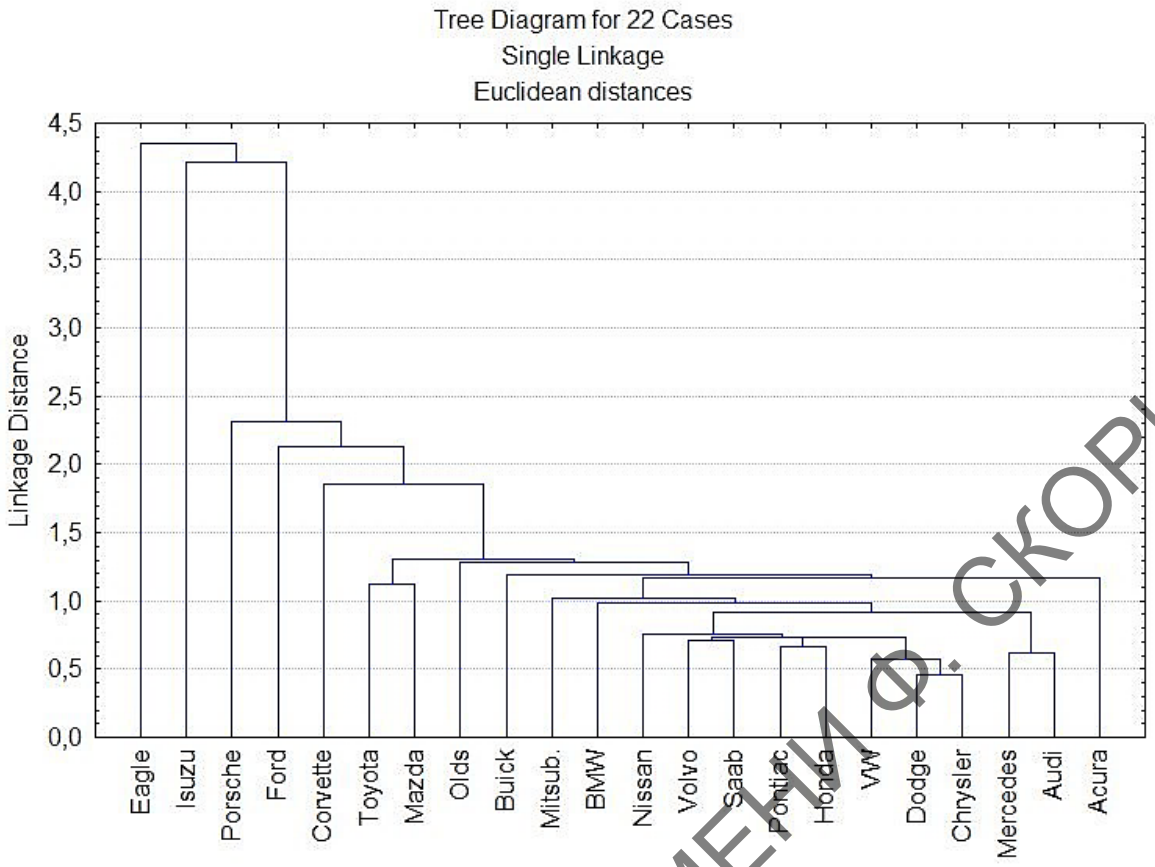


Рисунок 8.19 – Вертикальное отображение дендрограммы сходства машин по методу Joining (tree clustering)

Для горизонтального отображения необходимо в окне результатов кластеризации нажать кнопку **Horizontal hierarchical tree plot** (*График горизонтального иерархического древа*), а для вертикального – **Vertical icicle plot** (*Вертикальный график*) (рисунок 8.17).

На этом графике наиболее близкие машины имеют между собой наименьшее расстояние. Так, самые близкие по характеристикам будут Chrysler и Dodge, Audi и Mercedes, затем к первой группе близок Volkswagen и т. д.

Задание

Используя данные социально-экономического развития и загрязнённости радионуклидами (средние значения суммарной годовой эффективной индивидуальной дозы и удельной активности молока) 46 населённых пунктов, классифицируйте их при помощи кластерного анализа на 3 группы и дайте рекомендации о дальнейшей судьбе этих населённых пунктов.

№	Соц.-экон.	Радиолог.	№	Соц.-экон.	Радиолог.	№	Соц.-экон.	Радиолог.
1	0,24719328	0,29540060	17	0,71545293	0,35440414	33	0,8274442	0,36113045
2	0,49097333	0,49549700	18	0,88534920	0,57296465	34	0,89421901	1,00000000
3	0,76815312	0,24014938	19	0,65832115	0,27772873	35	0,38097797	0,45873837
4	0,83789641	0,35430513	20	0,73896951	0,63227917	36	0,7502056	0,30138691
5	0,92087342	0,31977714	21	0,68365850	0,35711118	37	0,61300178	0,51226606
6	0,83693198	0,40867035	22	0,68365850	0,59168568	38	0,44290375	0,36935573
7	0,64208613	0,30919644	23	0,51156682	0,42806438	39	0,85262386	0,23098892
8	0,75447239	0,28769677	24	0,76328686	0,77716795	40	0,78977027	0,39782878
9	0,84431658	0,33989094	25	0,45186181	0,33931740	41	0,74822401	0,52675872
10	0,43312923	0,24221188	26	0,46782343	0,42848911	42	0,89249235	0,36927655
11	0,92254974	0,28911136	27	0,84907526	0,29800685	43	0,73758750	0,65779692
12	0,82397599	0,24254329	28	0,76096615	0,25151951	44	0,61224608	0,24290558
13	0,96296219	0,33465852	29	0,29841344	0,20134516	45	0,77375605	0,23284717
14	0,80014315	0,22299802	30	0,82047265	0,43816498	46	0,93288537	0,27083840
15	0,82842084	0,69562282	31	0,88639612	0,25151590			
16	1,00000000	0,36552183	32	0,76052718	0,53358934			

Литература по теме

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

ТЕМА 9. РАСЧЁТ ПОКАЗАТЕЛЕЙ РАЗНООБРАЗИЯ ПРИ ПОМОЩИ ПАКЕТА ПРИКЛАДНЫХ ПРОГРАММ BIODIVERSITY PRO

9.1 Понятие о биоразнообразии.
9.2 Знакомство с пакетом BioDiversity Pro.

9.1 Понятие о биоразнообразии

Биологическое разнообразие – это вариабельность живых организмов из всех источников, включая наземные, морские и другие водные экосистемы и экологические комплексы, частью которых они являются; включает в себя разнообразие в рамках вида, между видами и экосистемами.

Биоразнообразие как экологическое понятие отражает функциональную роль его форм в жизни экосистем. Это не просто совокупность видов и не синоним видового разнообразия, а определенное их функциональное соотношение, сочетание. Считается, что разнообразие сообщества, включающего виды, относящиеся ко многим родам, выше, чем у такого, где большинство видов принадлежит к одному роду.

Единой классификации биоразнообразия не существует в связи с его сложностью и разномасштабностью. Наиболее часто встречается инвентаризационное разнообразие по Р. Уиттекеру (1960):

– α-*разнообразие* – видовое разнообразие в пределах одного сообщества, внутри одного однородного местообитания.

Как вариант альфа-разнообразия различают *точечное альфа-разнообразие*, отражающее разнообразие в микроместообитании, в выборке, полученной из однородного местообитания, в пределах небольшого гомогенного местообитания сообщества.

При оценке альфа-разнообразия принимаются во внимание два фактора: видовое богатство и выравненность обилий видов;

– β-*разнообразие* – позволяет сравнивать видовой состав разных сообществ. Обычно используется при установлении характера изменения видового состава сообществ, сменяющих друг друга по градиенту факторов среды или при переходе от одного местообитания (сообщества) к другому;

– γ-*разнообразие* – видовое разнообразие в пределах ландшафта, острова. (аналог альфа-разнообразия в большом пространстве и измеряется таким же путем);

– δ-*разнообразие* (добавлено Крюгером и Тейлором в 1979 г.) – географическая дифференциация, изменение растительности вдоль климатических градиентов или между географическими регионами.

Связано с крупными частями биома или биогеографическими регионами, отражает градиент разнообразия (подобно бета-разнообразию служит также для оценки варьирования между сообществами);

– ε-*разнообразие* – отражает глобальный градиент разнообразия в системе зонально-поясных биомов. Это наиболее высокий уровень, соответствующий природным зонам.

Таким образом, альфа-, гамма- и эпсилон-разнообразие – это оценка разнообразия сообщества разного масштаба; бета-, дельта-разнообразие – сравнение, оценка варьирования между сообществами разного масштаба.

9.2 Знакомство с пакетом BioDiversity Pro

Пакет BioDiversity Pro предназначен для расчета показателей биоразнообразия (индексов, рангов видов, дисперсии, моделей распределения, кластерного анализа и др.). Пакет является достаточно мощным средством, которое в значительной степени облегчает работу специалистам в области биоразнообразия.

Изучим основные возможности программы на конкретном примере встречаемости жужелиц в 4 биотопах окрестностей г. Гомеля (таблица 9.1).

Таблица 9.1 – Данные по обилию видов жужелиц в 4 биотопах отвалов фосфогипса Гомельского химического завода

	Вид	Б1	Б2	Б3	Б4		Вид	Б1	Б2	Б3	Б4
1	*Agonum fuliginosum*	0	0	0	1	22	*Carabus hortensis*	0	1	0	0
2	*Amara aenea*	0	4	1	2	23	*Cicindela hybrida*	4	0	0	0
3	*Amara communis*	0	2	0	1	24	*Cychrus caraboides*	0	0	0	15
4	*Amara consularis*	0	2	0	0	25	*Dyschirius arenosus*	0	0	1	0
5	*Amara majuscula*	0	1	0	0	26	*Harpalus calceatus*	0	0	0	1
6	*Anisodactylus binotatus*	0	0	1	1	27	*Harpalus flavescens*	2	0	0	0
7	*Anisodactylus signatus*	1	0	0	0	28	*Harpalus latus*	0	0	0	1
8	*Badister lacertosus*	0	0	0	2	29	*Harpalus rufipes*	1	5	15	0
9	*Badister unipustulatus*	0	0	1	0	30	*Harpalus rubripes*	0	0	1	0

Продолжение таблицы 9.1

	Вид	Б1	Б2	Б3	Б4		Вид	Б1	Б2	Б3	Б4
10	*Bembidion azurescens*	0	1	0	0	31	*Harpalus tardus*	0	1	1	1
11	*Bembidion lampros*	0	1	0	1	32	*Leistus ferrugineus*	0	0	0	5
12	*Bembidion properans*	2	0	0	0	33	*Leistus rufescens*	0	1	1	1
13	*Bembidion varium*	0	0	1	0	34	*Licinus depressus*	0	1	0	0
14	*Bembidion velox*	0	2	25	0	35	*Microlestes maurus*	0	1	47	5
15	*Broscus cephalotes*	11	0	0	0	36	*Oxypselaphus obscurus*	0	0	4	1
16	*Calathus erratus*	0	87	90	1	37	*Pterostichus niger*	0	1	4	44
17	*Calathus fuscipes*	0	2	4	0	38	*Pterostichus strenuus*	0	1	0	1
18	*Calathus melanocephalus*	0	9	0	0	39	*Pterostichus vernalis*	0	0	0	1
19	*Calathus micropterus*	0	0	0	1	40	*Stenolophus mixtus*	0	0	1	0
20	*Carabus glabratus*	0	0	0	91	41	*Synuchus vivalis*	0	0	1	9
21	*Carabus granulatus*	0	0	0	7						

Запустите программу и вы увидите рабочее окно (рисунок 9.1).

Рисунок 9.1 – Рабочее окно программного пакета BioDiversity Pro

Шаг 1. Создание нового файла.

Для создания нового файла в меню **File** (*Файл*) необходимо выбрать опцию **New Data** (*Новые данные*) (или нажать на клавиатуре стандартную комбинацию клавиш ***Ctrl*+*N***). Программа выдаст диалоговое окно, в которое нужно внести количество видов (рядов) и количество примеров (столбцов) (рисунок 9.2):

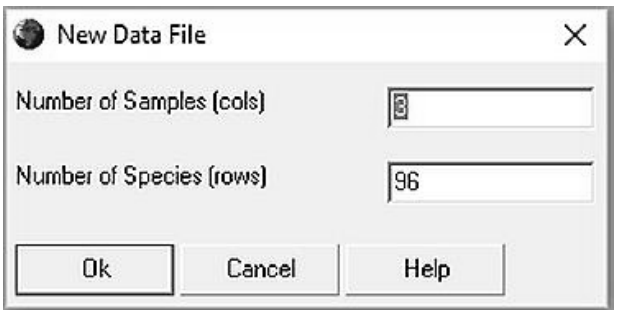


Рисунок 9.2 – Ввод числа видов и примеров

В поле **Number of Samples (cols)** (*Количество примеров, столбцы*) необходимо ввести количество биотопов (в нашем случае – 4), а в поле **Number of Species (rows)** (*Количество видов, ряды*) – количество видов (в нашем случае – 41). В итоге на экране появится базовая таблица (рисунок 9.3).

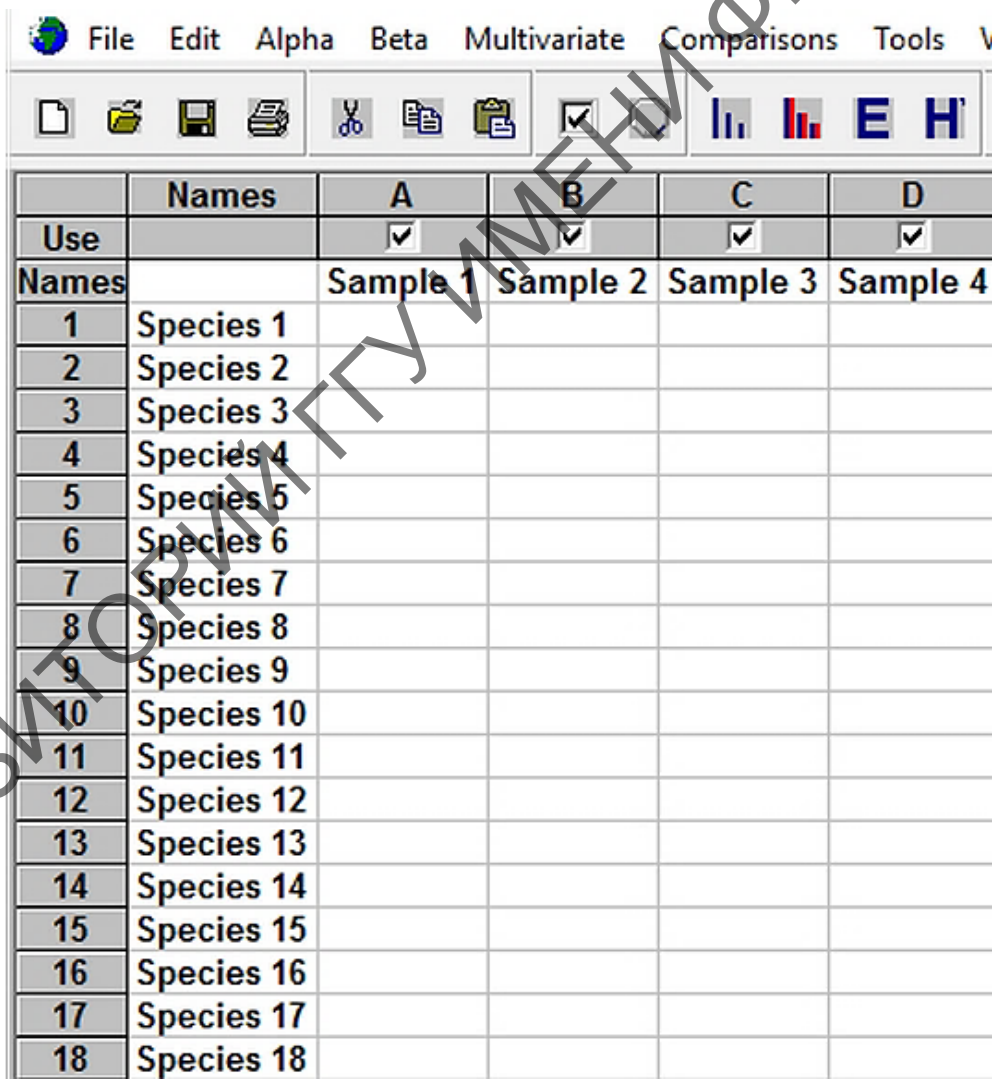
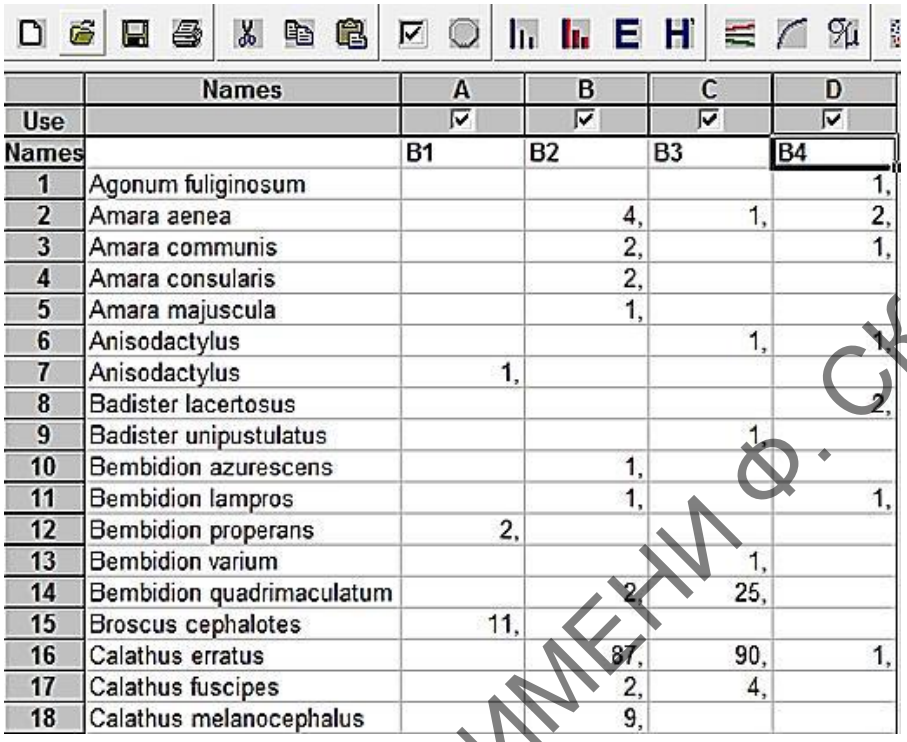


Рисунок 9.3 – Рабочая таблица
(из-за экономии места показаны только первые 18 рядов)

Шаг 2. Ввод данных.

Необходимо перенести данные из таблицы 9.1 в рабочую таблицу. В итоге рабочая таблица примет вид, как на рисунке 9.4 (цифра «0» в таблице не отражается, но ставить её нужно обязательно!).



	Names	A	B	C	D
Use		☑	☑	☑	☑
Names		B1	B2	B3	B4
1	Agonum fuliginosum				1,
2	Amara aenea		4,	1,	2,
3	Amara communis		2,		1,
4	Amara consularis		2,		
5	Amara majuscula		1,		
6	Anisodactylus			1,	1,
7	Anisodactylus	1,			
8	Badister lacertosus				2,
9	Badister unipustulatus			1,	
10	Bembidion azurescens		1,		
11	Bembidion lampros		1,		1,
12	Bembidion properans	2,			
13	Bembidion varium			1,	
14	Bembidion quadrimaculatum		2,	25,	
15	Broscus cephalotes	11,			
16	Calathus erratus		87,	90,	1,
17	Calathus fuscipes		2,	4,	
18	Calathus melanocephalus		9,		

Рисунок 9.4 – Заполненная рабочая таблица
(из-за экономии места показаны только первые 18 рядов)

Шаг 3. Расчёт индексов разнообразия.

Для примера рассчитаем показатель информационного *индекса Шеннона.* Для этого в пункте меню **Alpha** (*Альфа разнообразие*) нужно выбрать подменю **Diversity Indices** (*Индексы разнообразия*) и **Shannon** (*индекс Шеннона*) (рисунок 9.5).

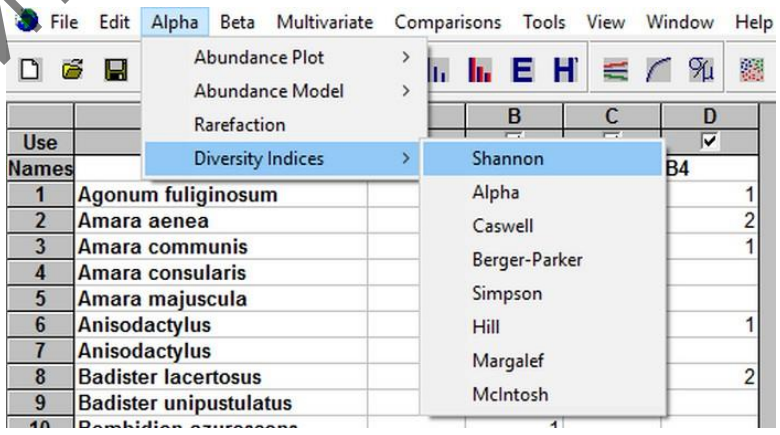


Рисунок 9.5 – Выбор в меню опции расчета индекса Шеннона

Программа покажет графическое изображение значений индекса. Чтобы увидеть его численное значение, зайдите в пункт меню **Window** (*Окно*) и выберите опцию **Shannon Index Results** (*Результаты расчета индекса Шеннона*), как показано на рисунке 9.6. После этого программа отобразит численные значения индекса в первой строке (рисунок 9.7).

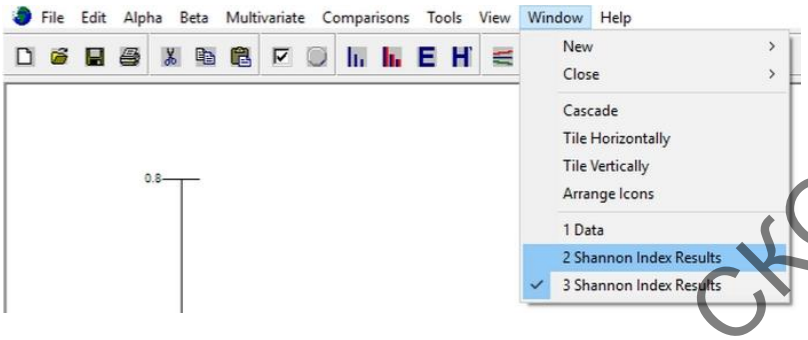


Рисунок 9.6 – Выбор показа числовых значений индекса Шеннона

	A	B	C	D	E
1	Index	B1	B2	B3	B4
2	Shannon H' Log Base 10,	0,76	1,229	1,19	1,307
3	Shannon Hmax Log Base 10,	0,778	1,255	1,23	1,342
4	Shannon J'	0,977	0,979	0,967	0,974

Рисунок 9.7 – Числовые значения индекса Шеннона

Аналогично рассчитываются и остальные индексы.

Шаг 4. Модель «ранг-обилие».

Для построения модели «ранг-обилие» необходимо в рабочей таблице выбрать пункт меню **Alpha** (*Альфа разнообразие*), подменю **Abundance Plot** (*Графики обилия*), и кликнуть левой клавишей мыши на опцию **Rank** (*Ранговое распределение*). Вы увидите графики «ранг-обилие» для каждого из биотопов (рисунок 9.8).

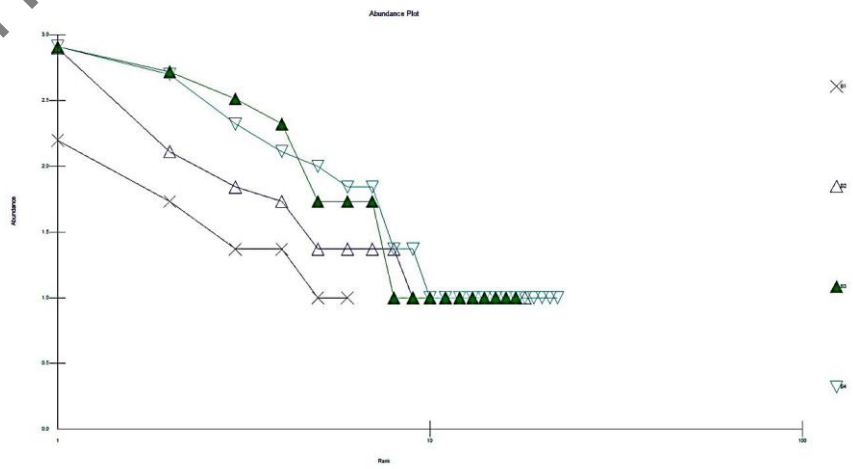


Рисунок 9.8 – Графики распределения «ранг-обилие»

Шаг 5. Редактирование графиков.

Для редактирования графика щёлкните правой кнопкой мыши в области графика и увидите контекстное меню (рисунок 9.9).

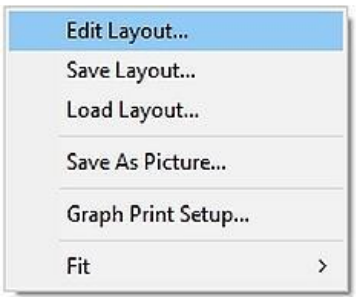


Рисунок 9.9 – Контекстное меню графика «ранг-обилие»

В контекстном меню (рисунок 9.9) нужно выбрать опцию **Edit Layout…** (*Редакция раскладки*). Появится диалоговое окно, отображённое на рисунке 9.10.

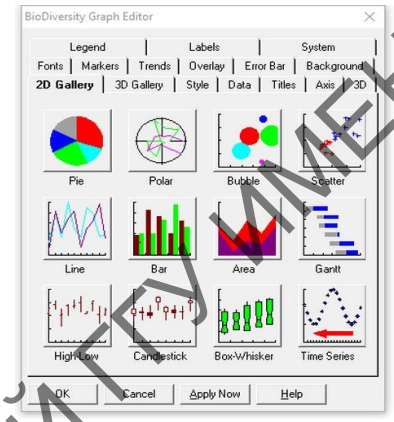


Рисунок 9.10 – Диалоговое окно редакции графика «ранг-обилие»

Рассмотрите каждую закладку и опции в них присутствующие. Для использования той или иной опции нужно её выбрать и нажать кнопку **Apply Now** (*Применить сейчас*).

Шаг 6. Настройки кластерного анализа.

Для проведения кластерного анализа сначала необходимо зайти в меню настроек. Для этого – выбрать пункт меню **Tools** (*Инструменты*) и подменю **Options** (*Опции*). В закладке **Cluster Analysis** (*Кластерный анализ*) выбрать метод кластеризации по коэффициенту Жаккара (*Jaccard*) и одиночное расстояние (*Single Linkage*).

Шаг 7. Проведение кластерного анализа.

Для проведения непосредственно самого анализа в окне рабочей таблицы необходимо выбрать пункт меню **Multivariate** (*Многообра-*

зие) и подменю **Cluster Analysis** (*Кластерный анализ*) или нажать специальную кнопку на панели инструментов, как показано на рисунке 9.11.

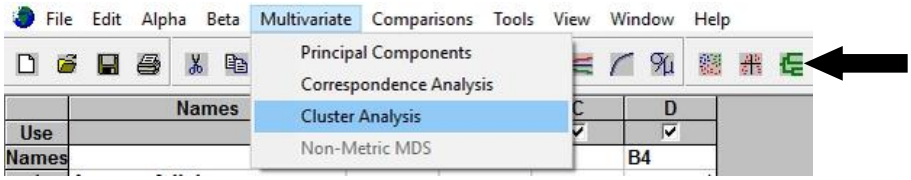


Рисунок 9.11 – Запуск кластерного анализа

Шаг 8. Результаты анализа.

Результаты дендрограммного кластерного анализа показаны на рисунке 9.12.

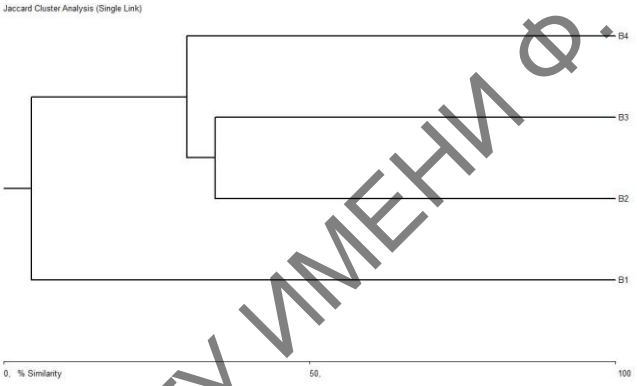


Рисунок 9.12 – Дендрограмма кластерного анализа

Шаг 9. Настройка графика анализа.

Настроить рисунок можно, кликнув правой кнопкой мыши на области дендрограммы и вызвать контекстное меню (рисунок 9.13).

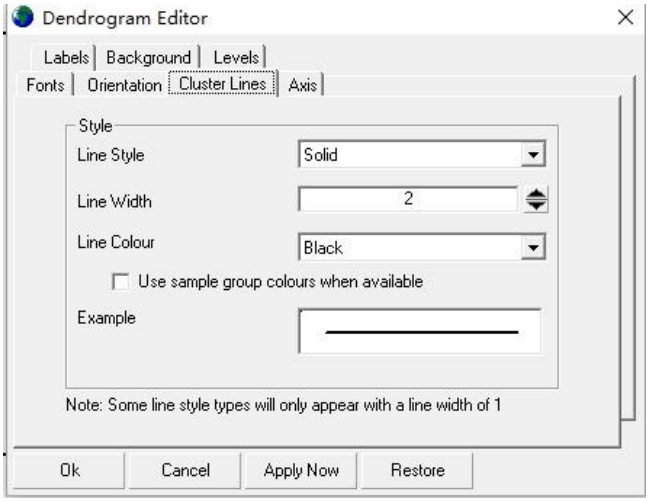


Рисунок 9.13 – Контекстное меню дендрограммы кластерного анализа

Задание

Используя данные о встречаемости жесткокрылых в шести прибрежных сообществах (таблица 9.2), рассчитайте индексы Шеннона, Симпсона, Бергера-Паркера, Маргалефа и МакИнтоша, постройте графики «ранг-обилие» и проведите кластерный анализ сходства этих сообществ по коэффициенту Жаккара.

Таблица 9.2 – Встречаемость жесткокрылых в шести прибрежных сообществах

Вид	1	2	3	4	5	6	Вид	1	2	3	4	5	6
Byrrhus pilula	3	1	1	0	0	0	*Pterostichus melanarius*	0	0	1	0	0	0
Amara aenea	0	0	0	0	0	1	*Pterostichus niger*	4	4	1	0	0	3
Anisodactylus signatus	0	0	1	0	0	0	*Oulema erichsonii*	0	0	0	0	0	1
Calathus erratus	30	10	19	20	6	20	*Phyllobius argentatus*	0	0	0	0	0	1
Calathus fuscipes	0	6	4	33	5	0	*Dermestes laniarius*	1	0	0	0	0	0
Carabus granulatus	2	0	0	3	1	5	*Agriotes lineatus*	1	0	0	0	0	0
Chlaenius tristis	0	0	0	1	0	0	*Agriotes obscurus*	1	1	2	0	2	0
Chlaenius vestitus	6	0	0	2	0	1	*Agrypnus murinus*	2	0	0	0	0	3
Curtonotus aulicus	0	0	0	0	1	0	*Prosternon tesellatum*	0	0	0	0	0	1
Elaphrus riparius	0	0	0	1	0	0	*Selatosomus aeneus*	2	0	2	0	0	0
Harpalus affinis	0	0	0	0	2	0	*Hydrous aterrimus*	0	0	0	1	0	0
Harpalus rufipes	0	0	1	8	0	0	*Hydrochara caraboides*	0	0	0	0	0	1
Harpalus tardus	0	0	3	6	2	0	*Glischrochilus 4punctatus*	2	0	0	0	0	0
Loricera pilicornis	0	0	0	1	0	0	*Phalacrus caricis*	0	0	0	4	0	0
Nebria brevicollis	0	0	0	0	1	0	*Silpha carinata*	2	0	0	0	0	0
Oodes helopioides	0	0	0	2	0	0	*Silpha obscura*	0	0	0	3	1	0
Platynus assimilis	1	0	1	42	4	2	*Nicrophorus vespillo*	0	0	0	0	0	2
Poecilus versicolor	29	19	9	0	4	6	*Crypticus quisquilis*	1	0	0	0	0	0

Литература

1 Боровиков, В. П. Программа STATISTICA для студентов и инженеров / В. П. Боровиков. – М. : КомпьютерПресс, 2001. – 301 с.

2 Боровиков, В. П. Популярное введение в программу Statistica / В. П. Боровиков. – М. : КомпьютерПресс, 1998. – 69 с.

3 Жученко, Ю. М. Статистическая обработка информации с применением персональных компьютеров : практическое руководство для студентов 5 курса / Ю. М Жученко. – Гомель : ГГУ им. Ф. Скорины, 2007. – 101 с.

4 Ивантер, Э. В. Основы биометрии: Введение в статистический анализ биологических явлений и процессов / Э. В. Ивантер, А. В. Коросов. – Петрозаводск : Издательство ПГУ, 1992. – 168 с.

5 Мастицкий, С. Э. Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований / С. Э. Мастицкий. – Минск : РУП «Институт рыбного хозяйства», 2009. – 76 с.

6 Рокицкий, П. Ф. Биологическая статистика / П. Ф. Рокицкий. – Минск : «Вышэйшая школа», 1973. – 320 с.

ПРИЛОЖЕНИЕ А

(обязательное)

Способы решения статистических задач [4]

Задача	Статистический показатель	Статистический метод и критерий
1 Оценить различие:		
1) двух признаков по величине	M, x – средняя величина	Сравнение средних по t критерию Стьюдента; долей – по F критерию Фишера с φ (фи)-преобразованием
2) двух признаков по изменчивости	σ – среднее квадратическое отклонение; σ^2 – дисперсия	Сравнение сигм по t критерию Стьюдента. Сравнение дисперсий по F критерию Фишера
3) нескольких признаков по величине	η^2 – сила влияния	Однофакторный дисперсионный анализ, F критерий Фишера
4) двух эмпирических распределений	частоты по классам	Метод χ^2 (хи-квадрат)
5) эмпирического распределения от теоритического	частоты по классам	Метод χ^2 (хи-квадрат)
2 Оценить влияние:		
1) одного признака на другой	R – коэффициент регрессии	t критерий Стьюдента
2) взаимное	r – коэффициент корреляции	t критерий Стьюдента
3) одного фактора на признак	η^2 – сила влияния	Однофакторный дисперсионный анализ, F критерий Фишера
4) двух факторов на признак	η^2 – сила влияния	Двухфакторный дисперсионный анализ, F критерий Фишера
3 Оценить принадлежность:		
1) варианты к одномерной совокупности	T – нормированное отклонение	Оценка «выскакивающих» значений T
2) объекта к двумерной (многомерной совокупности)	Z – значение дискриминантной функции	Дискриминантный анализ, F критерий Хотеллинга
	D – евклидово расстояние	Кластерный анализ

Учебное издание

Галиновский Николай Геннадьевич,
Зятьков Сергей Александрович

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Пособие

2-е издание, стереотипное

Редактор *В. И. Шкредова*
Корректор *В. В. Калугина*

Подписано в печать 31.08.2021. Формат 60x84 1/16.
Бумага офсетная. Ризография. Усл. печ. л. 11,63.
Уч.-изд. л. 12,71. Тираж 40 экз. Заказ 434.

Издатель и полиграфическое исполнение:
учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 3/1452 от 17.04.2017.
Специальное разрешение (лицензия) № 02330 / 450 от 18.12.2013.
Ул. Советская, 104, 246028, Гомель.