МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
РОССИЙСКОЙ ФЕДЕРАЦИИ

РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ –
МСХА имени К.А. ТИМИРЯЗЕВА

Р.Р. Усманов

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ АГРОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ В ПРОГРАММЕ «STATISTICA»

Учебно-методическое пособие

Москва

РГАУ-МСХА имени К.А. Тимирязева

2020

УДК 311.21: 633/635(075.8)
ББК 65.051:65.325.1я73
У 757

*Рецензенты: **Джалилов Ф.С.-У**., доктор биологических наук, профессор, заведующий кафедрой защиты растений ФГБОУ ВО РГАУ-МСХА имени К.А. Тимирязева, **Милюкова Н.А.,** кандидат биологических наук, старший научный сотрудник ВНИИ сельскохозяйственной биотехнологии*

У 757 **Усманов, Р. Р.** Статистическая обработка данных агрономических исследований в программе «STATISTICA» : учебно-методическое пособие / Р. Р. Усманов ; Российский государственный аграрный университет – МСХА имени К. А. Тимирязева. – Москва : РГАУ-МСХА имени К. А. Тимирязева, 2020. – 177 с. – Текст : электронный

DOI: 10/34677/2020.004

Учебно-методическое пособие охватывает вопросы обработки экспериментальных данных. В пособии в доступной форме изложены методы статистической обработки результатов агрономических исследований в пакете «Statistica» с их биологической интерпретацией.

Учебно-методическое пособие предназначено для магистров по направлениям подготовки: «Агрономия» и «Биотехнология», а также аспирантов по направлению подготовки «Сельское хозяйство». Пособие может быть использовано студентами, научными сотрудниками и преподавателями при подготовке дипломных проектов, диссертационных работ и научных статей.

Рекомендовано к изданию учебно-методической комиссией факультета агрономии и биотехнологии РГАУ-МСХА имени К.А. Тимирязева, протокол № 11 от 10 февраля 2020 г.

Usmanov, R. R. Statistical processing of agronomic research data in the "STATISTICA" program : educational and methodological guide / R. R. Usmanov, Russian state agrarian University-Moscow state Agricultural Academy named after K. A. Timiryazev. – Moscow : RSAU-MTAA named after K. A. Timiryazev, 2020. – 177 p. – Text: electronic.
DOI: 10/34677/2020.004

The training manual covers the processing of experimental data. The manual describes in an accessible form the methods of statistical processing of the results of agronomic research in the "Statistica" package with their biological interpretation.

The training manual is intended for masters in the areas of training: "Agronomy" and "Biotechnology", as well as graduate students in the field of training "Agriculture". The manual can be used by students, researchers and teachers in the preparation of diploma projects, dissertations and scientific articles.

Recommended for publication by the educational and methodological Commission of the faculty of agronomy and biotechnology of the Timiryazev state agrarian University - MSHA, Protocol no. of February 2020.

Оглавление

Глава 1. ОБЩИЕ СВЕДЕНИЯ О СТАТИСТИЧЕСКОМ ПАКЕТЕ STATISTICA

1.1 Структура пакета Statistica

Программа **Statistica** запускается по значку на рабочем столе  или из меню **ПУСК – ПРОГРАММЫ – STATISTICA.** После запуска на экране монитора появится рабочее окно программы (рис. 1.1), которое обладает стандартным для любого Windows-приложения видом. В самом верху слева находится заголовок окна, в нашем случае файл с расширением. **sta: «Statistica – Примеры 1-2.sta»**,. а также кнопки управления состоянием окна, которые расположены в крайней правой области строки заголовка.

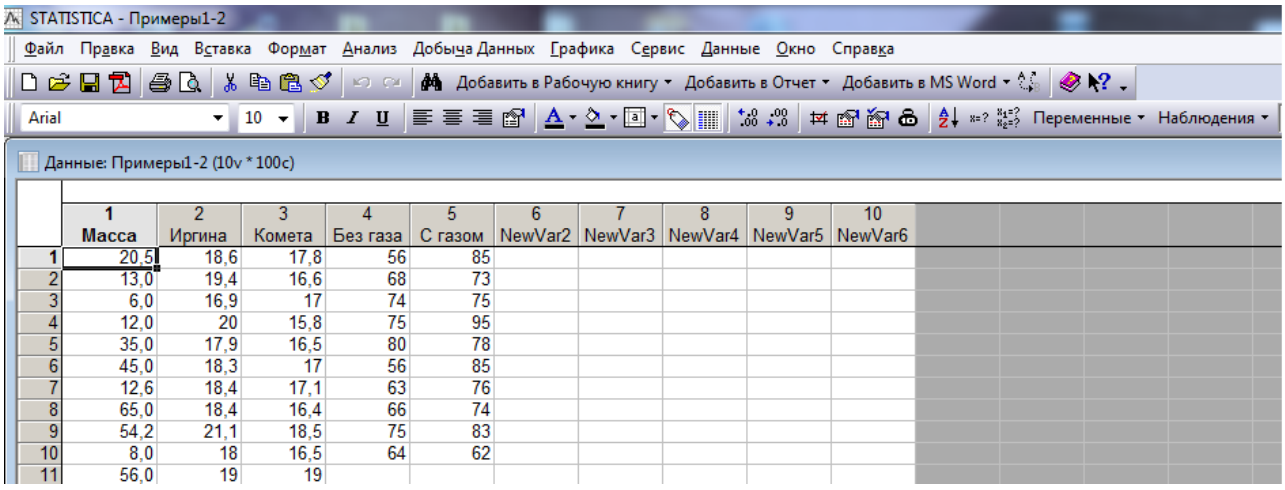
Ниже располагается строка меню, с привычными для Windows – приложений пунктами: **Файл (File), Правка (Edit), Вид (View), Вставка (Insert), Формат (Format), Окно (Window), Помощь (Help).**

Под строкой меню находится стандартная панель инструментов, четвертая строка – панель форматирования, ниже которой располагается рабочая область, занимающая большую часть окна программы, в которой представлены исходные данные, результаты расчетов или диаграммы.

В нижней части экрана находится строка состояния окна, в которой представлена обычно информация о содержимом объектов окна и другая информация в зависимости от работающей программы.

В настоящих методических указаниях приводится описание методов обработки данных агрономических исследований в русскоязычном интерфейсе программы «Statistica – 10», который имеет более широкие возможности по сравнению с русскоязычной программой «Statistica – 6». Основными преимуществами программы «Statistica – 10» являются удобный интерфейс, новые модули для проведения статистических расчетов; возможность напрямую импортировать файлы Office 2010, 2016, а также разнообразные интерактивные графические процедуры.

При выполнении расчетов в англоязычных версиях программы «Statistica» пользователь может ориентироваться на термины, указанные на английском языке в скобках.



	1 Масса	2 Иргина	3 Комета	4 Без газа	5 С газом	6 NewVar2	7 NewVar3	8 NewVar4	9 NewVar5	10 NewVar6
1	20,5	18,6	17,8	56	85					
2	13,0	19,4	16,6	68	73					
3	6,0	16,9	17	74	75					
4	12,0	20	15,8	75	95					
5	35,0	17,9	16,5	80	78					
6	45,0	18,3	17	56	85					
7	12,6	18,4	17,1	63	76					
8	65,0	18,4	16,4	66	74					
9	54,2	21,1	18,5	75	83					
10	8,0	18	16,5	64	62					
11	56,0	19	19							

Рис. 1.1. **Рабочее окно программы Statistica 10**

Важными специфическими разделами меню, позволяющими выполнять все расчеты и строить графики, являются: **Анализ (Statistics), Графика (Graphs), Data (Данные).** При щелчке мышью на указанные разделы меню открывается выпадающий список с пунктами данного раздела меню. Так, при нажатии на **Анализ (Statistics),** выпадает меню основных статистических приемов, а далее еще подменю (рис.1.2).

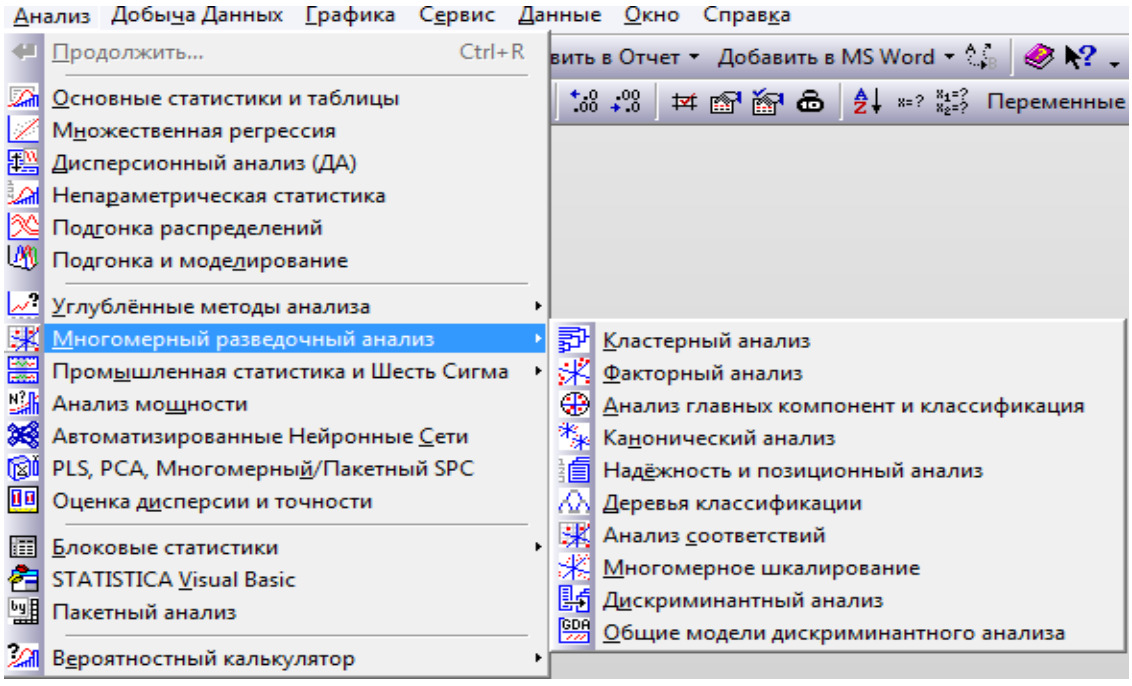


Рис.1.2. **Список статистических методов в меню первого и второго уровня раздела *Анализ***

При выборе раздела **Графика (Graphs)** (рис. 1.3) выпадает следующее меню, выбрав которое можно построить заданную диаграмму.

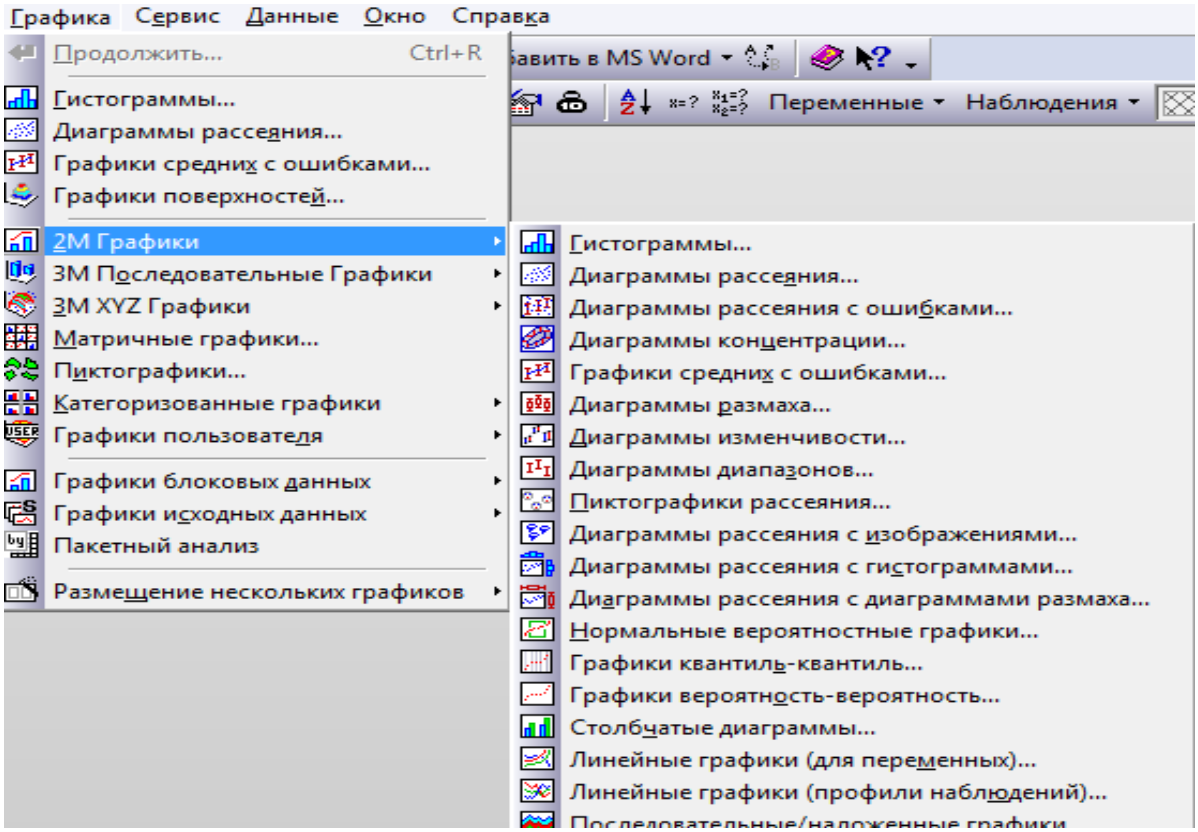


Рис.1.3. **Галерея графиков раздела *Графика***

Для открытия уже имеющего файла следует нажать кнопку **Открытие** или выбрать команду **Файл/Открыть документ**.

1.2 Ввод и редактирование данных

Для создания нового документа (файла) в пункте основного меню (**Файл**) выбрать (**Новый**); или нажать кнопку (белая страничка – «создать файл по умолчанию») на панели инструментов, в результате появится диалоговое окно создания нового документа **(Create new document**), в котором вкладками указаны разные формы документов. Так как мы создаем новую таблицу с данными, останемся на вкладке **Таблица (Spreadsheet),** которая по умолчанию предстает перед пользователем первой.

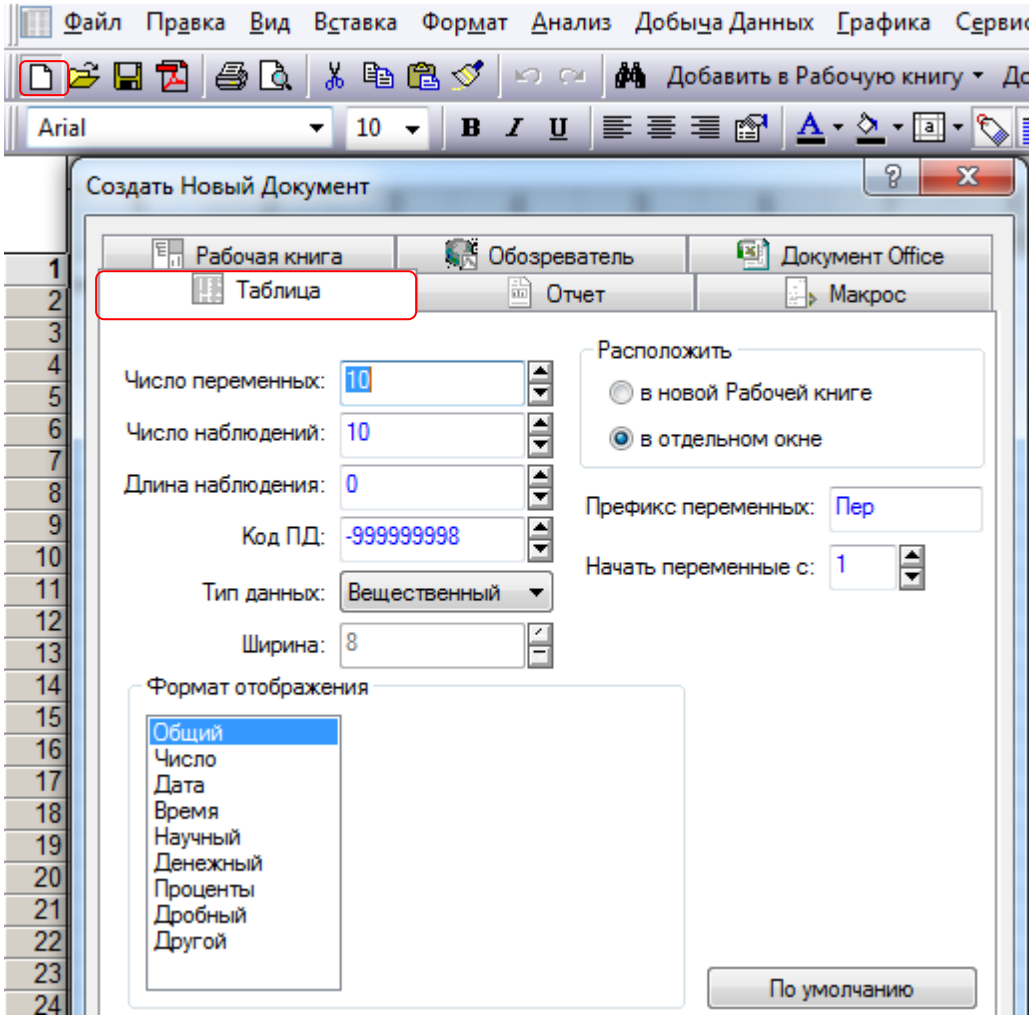


Рис. 1.4. **Диалоговое окно нового документа**

По умолчанию в поле **Число переменных (Number of variables) и Число наблюдений (Number of cases)** указано 10. **Переменные** – это изучаемые признаки, варианты опыта, число **наблюдений** – это объем выборки, повторность опыта. Если мы хотим создать таблицы с 5 переменными и объемом выборки 50 наблюдений, необходимо в поле **Число переменных (Number of variables)** выставить 5, а в поле **Число наблюдений (Number of cases)** – 50. Остальные опции закладки оставим без изменений. После нажатия кнопки **ОК** в рабочей области программы появится таблица с 5 столбцами и 50 строками (рис. 1.5).

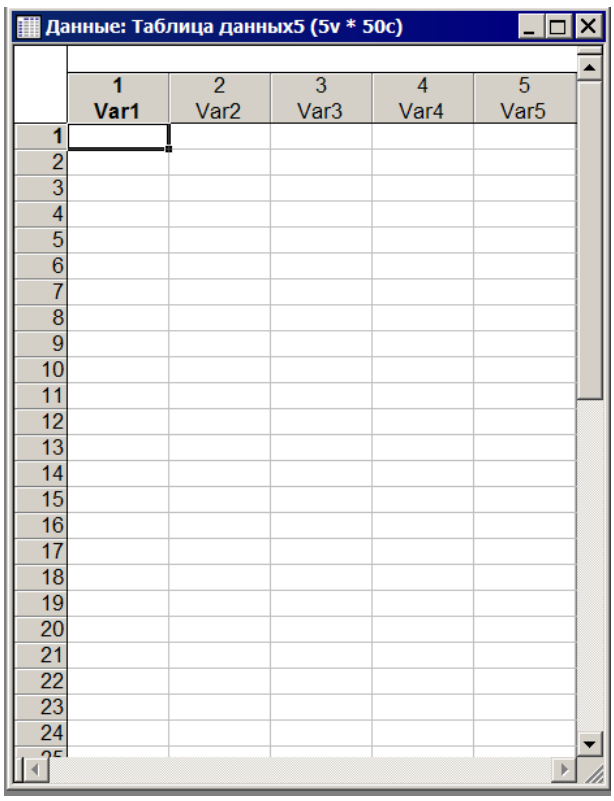


Рис. 1.5. **Рабочая область программы Statistica**

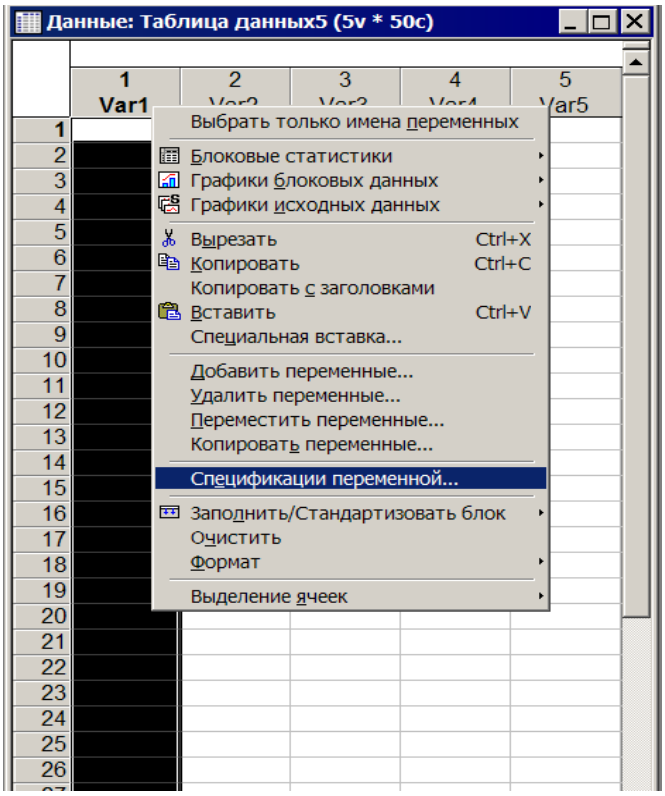


Рис. 1.6. **Окно для работы с переменными**

Создадим файл исходных данных на примере массы 50 клубней картофеля.

Пример. Произведено взвешивание 50 клубней картофеля, г.

70 64 130 76 83 12 75 60 145 85 11 112 56 42 109 94 67 98 75 150 86 49 35 114 83 89 125 76 45 9 34 85 75 63 102 99 73 65 156 35 65 35 84 134 124 76 115 59 70 82

Анализируемые данные представляются в Statistica в виде электронной таблицы, подобно тому, как это делается, например, в программе MS Excel. Однако электронная таблица с данными в программе Statistica имеет принципиальные отличия. Если в обычных электронных таблицах столбцы и строки равноправны, то в таблице программы Statistica в качестве столбцов выступают только **Переменные (Variables),** а в качестве строк – **Наблюдения (Cases).** Переменными в агрономических исследованиях являются исследуемые признаки (масса, высота, пораженность, качество и т.д.), изучаемые варианты опыта. Под наблюдениями же понимаются конкретные

значения, которые принимают переменные при отдельных измерениях. Statistica может обрабатывать не только числовые, но и текстовые данные.

В случае необходимости можно добавлять, удалять и т.п. переменные и наблюдения, щелкнув в строке форматирования по кнопке **Переменные** или **Наблюдения** и в появившемся подменю указать какое количество переменных, и после какой переменной (**Var**) добавить или с какой и по какую переменную удалить (**Var**). Аналогичным образом добавляются или удаляются наблюдения (строки).

Перед тем как начать вводить данные в таблицу, необходимо выполнить определенную предварительную подготовку созданной нами таблицы. Столбцы имеют порядковый номер и по умолчанию обозначаются **Var1, Var2**, и т.д. Для удобства работы столбцам рекомендуется присваивать краткие наименования анализируемых признаков. Для этого необходимо подвести курсор мыши к заголовку столбца и дважды кликнуть по нему. В результате появится окошко, в котором осуществляется настройка свойств переменной (рис. 1.7). Имя переменной указывается в поле **Имя (Name)**, установив курсор в это поле, введем изучаемый признак – «Масса клубней». Формат надписи (шрифт, его размер и т.п.) можно задать с помощью стандартных инструментов для форматирования текста, расположенных выше.

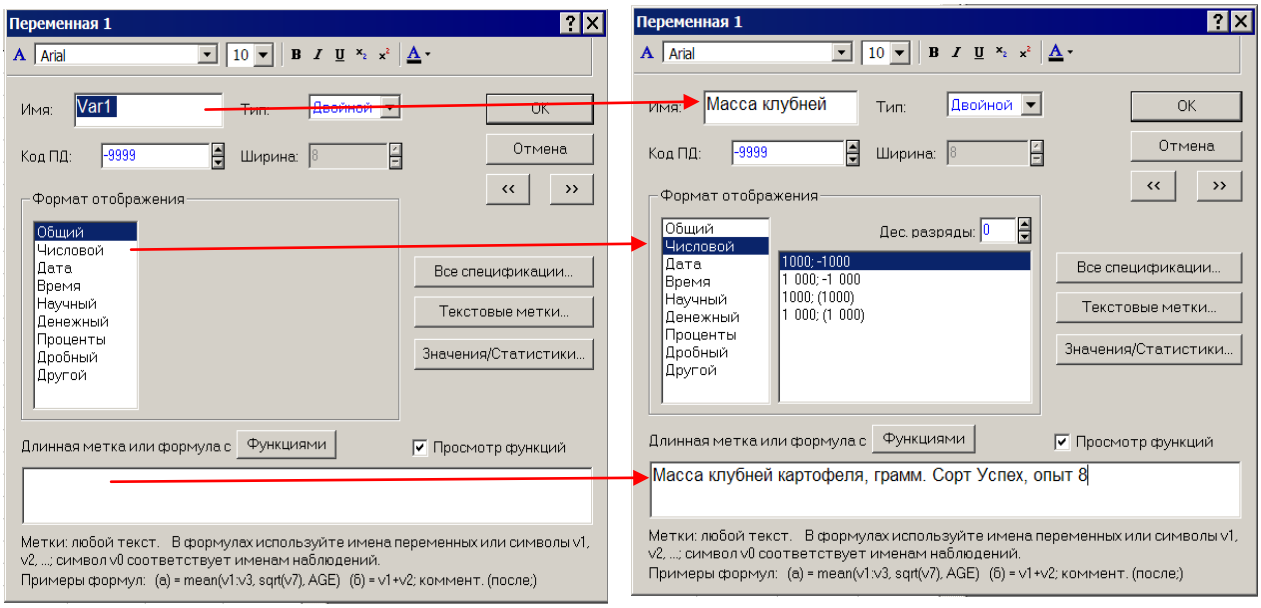


Рис. 1.7. **Изменение параметров переменной**

В поле **Тип (Type),** по умолчанию оставим **Двойной (Double)**. Так как масса клубней – это числовые данные в поле **Формат отображения (Display Format)**, выберем **Числовые (Number),** а в дополнительном поле **Десятичные знаки (Decimal Places)** – необходимое количество знаков после запятой.

В поле **Длинная метка (Long name)** можно указать дополнительную информацию по анализируемым признакам или ввести формулу, в результате чего значения переменной будут пересчитаны в соответствии с этой формулой.

Далее вводим значения массы 50 клубней в первый столбец. Сохраним созданный нами файл под названием «Масса клубней». Для этого необходимо в пункте основного меню **Файл (File)** выбрать **Сохранить (Save)** или нажать кнопку  на панели инструментов.

В программе Statistica реализована возможность импорта данных из других приложений. Например, если данные созданы в программе Excel, то можно использовать буфер обмена Windows и перенести их из Excel-файла. Для этого выделите копируемые столбцы в Excel-файла и скопируйте его в буфер обмена с помощью мыши, или нажав, сочетание клавиш «Ctrl+C». Затем вернитесь в программу Statistica, установите курсор в первую ячейку столбца и с помощью мыши или сочетания клавиш «Ctrl+V» вставьте данных из буфера.

Для получения дополнительной информации о некоторых функциях системы, следует нажать на клавишу справки (F1), когда выделена соответствующая команда или пункт меню. STATISTICA содержит Электронное руководство — справочную информацию по всем процедурам и функциям программы, доступную в контекстно-зависимом режиме при нажатии клавиши F1 или кнопки справки 

Вывод результатов статистических вычислений в программе Statistica осуществляется в так называемые рабочие книги (Workbook с расширением (.stw)). Рабочие книги помогают организовывать наборы файлов (например, таблиц результатов, графиков, текстовых/графических отчетов, пользовательских программ и т. д.), которые были созданы или использовались (например, просматривались) во время анализа набора данных. Рабочие книги

хранят список всех файлов, использовавшихся с текущим набором данных. Рабочие книги можно сохранять, изменять и экспортировать для повторного использования в других программах.

Глава 2. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

2.1 Статистические показатели (характеристики) данных агрономических исследований

В агрономических и биотехнологических исследованиях, как и во всех экспериментах для суждения о генеральной совокупности из нее отбирают выборку, и выводы о неизвестных параметрах генеральной совокупности производят на основании статистических показателей выборки (средние значения и показатели изменчивости или вариации).

К статистическим показателям (характеристикам) количественной изменчивости относятся:

Объем выборки (Valid) *N (n)* – численность выборки.

Сумма (Sum) ΣX сумма всех значений признака выборки.

Средняя арифметическая или выборочная средняя (Mean) $\bar{x} = \frac{\sum X}{n}$.

Среднее значение случайной величины представляет собой наиболее типичное, наиболее вероятное ее значение, своеобразный центр, вокруг которого разбросаны все значения признака.

В случае же если распределение отличается от нормального, количественные показатели следует описывать с использованием медианы и процентилей: Me *(Q25%; Q75%).*

Медианна (Median) – *Q50%*. Медиана – это возможное значение признака, которое делит ранжированную совокупность на две равные части, то есть медиана – средняя позиция в упорядоченном ряду значений. Она хорошо подходит как мера центральной тенденции для одномерных распределений даже в условиях выраженной ассиметрии распределения или при наличии выраженных выбросов значений.

Если медиана делит ранжированную совокупность на две равные части, то **процентили (Percentile)** – это такие значения признака, которые делят ее на 100 равных частей.

Децили (Decile) – делят ранжированную совокупность на десять равных частей.

Квартили (Quartile) – делят ранжированную совокупность на четыре равных части.

Нижний и верхний квартили (Lower, upper quartiles). Верхний квартиль – это такое значение, ниже которого располагаются 75% значений переменной. Нижний квартиль – это такое значение, ниже которого расположено 25% значений переменной.

Интерквартильный размах (Quartile range) – расстояние между нижним и верхним квартилями. Он равен разности значений 75% процентили и 25% процентили.

Мода (Mode) – это наиболее часто встречающееся значение переменной.

Геометрическое среднее (Geometric mean). Геометрическое среднее — это произведение всех значений переменной (X_i), возведенное в степень 1/n (единица, деленная на число наблюдений) $\bar{x}_{геом} = \sqrt[n]{X_1 \cdot X_2 \cdots X_n} \cdot \cdot = \sqrt[n]{\Pi X}$

Гармоническое среднее (Garmonic mean). Гармоническое среднее иногда используют для усреднения частот. Гармоническое среднее вычисляется по формуле: $\bar{x}_{гарм} = \frac{n}{\sum \frac{1}{X_i}}$, n — число наблюдений, X_i — значение наблюдения с номером i.

Дисперсия (Variance) $S^2 = \frac{\sum (X - \bar{x})^2}{n-1}$ Дисперсия является мерой изменчивости, вариации признака и представляет собой средний квадрат отклонений случаев от среднего значения признака. В отличие от других показателей вариации дисперсия может быть разложена на составные части, что позволяет тем самым оценить влияние различных факторов на вариацию признака.

Стандартное отклонение (Standard Deviation) $S = \sqrt{S^2}$. Стандартное (среднее квадратическое, основное) отклонение (или) ошибка отдельного наблюдения является абсолютной мерой изменчивости (вариации) признака. Оно показывает, на какую величину в среднем отклоняются индивидуальные значения признака от его среднего значения.

Коэффициент вариации (Coefficient of variation) $V = \frac{S}{\bar{x}} \cdot 100$ (%) – относительный показатель изменчивости.

Доверительный интервал для генеральной совокупности (Confidence interval) $\bar{x} \pm t_{05} \cdot S$ – интервал, в котором с заданной (95 или 99%) вероятностью находятся все значения генеральной совокупности.

Ошибка выборочной средней или стандартная ошибка среднего (Standard error of mean) $S_{\bar{x}} = \frac{S}{\sqrt{n}}$ – это величина, на которую отличается среднее значение выборки от среднего значения генеральной совокупности при условии, что распределение близко к нормальному.

Доверительный интервал для генеральной средней $\bar{x} \pm t_{05} \cdot S_{\bar{x}}$ – интервал, в котором с заданной (95 или 99%) вероятностью находится среднее значение генеральной совокупности.

Минимум (Minimum), максимум (maximum) X_{min} , X_{max} – минимальное и максимальное значения признака выборки.

Размах (Range) – разница между $X_{max} - X_{min}$.

Ассиметрия (Skewness) – характеризует степень смещения вариационного ряда относительно среднего значения по величине и направлению. В симметричной кривой коэффициент асимметрии равен нулю, если этот коэффициент значительно отличается от 0, распределение является ассиметричным.

Стандартная ошибка асимметрии (Standard error of Skewness)

Экцесс (Kurtosis) – характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутости кривой. В кривой

нормального распределения эксцесс равен нулю. Если эксцесс больше нуля, то кривая распределения характеризуется островершинностью, т.е. является более крутой по сравнению с нормальной. При отрицательном эксцессе кривая является более пологой по сравнению с нормальным распределением.

Стандартная ошибка эксцесса (Standard error of Kurtosis).

Пример. Произведено взвешивание 50 клубней картофеля, г.

70 64 130 76 83 12 75 60 145 85 11 112 56 42 109 94 67 98 75 150 86 49 35 114 83 89 125 76 45 9 34 85 75 63 102 99 73 65 156 35 65 35 84 134 124 76 115 59 70 82

Необходимо рассчитать основные статистические показатели количественной изменчивости. Определить доверительные интервалы и проверить на соответствие данных выборки закону нормального распределения.

Так как уже были введены данные этого примера (стр. 7-8), следует загрузить сохраненный файл исходных данных *Масса клубней.* Для выполнения необходимых расчетов воспользуемся опцией **Описательные статистики (Descriptive statistics).** В строке **Меню (Meny)** нажмем на кнопку **Анализ (Analyses)** и в ниспадающем меню выберем процедуру **Основные статистики и таблицы (Basic Statistics and Tables)**, далее в стартовой панели **Основные статистики и таблицы** выберем опцию **Описательная статистика (Descriptive statistics)**. В открывшемся диалоговом окне **Описательная статистика,** щелкнув по кнопке **Переменные,** открываем список переменных, для которых необходимо провести анализ, в нашем случае – переменную *Масса клубней* (рис. 2.1).

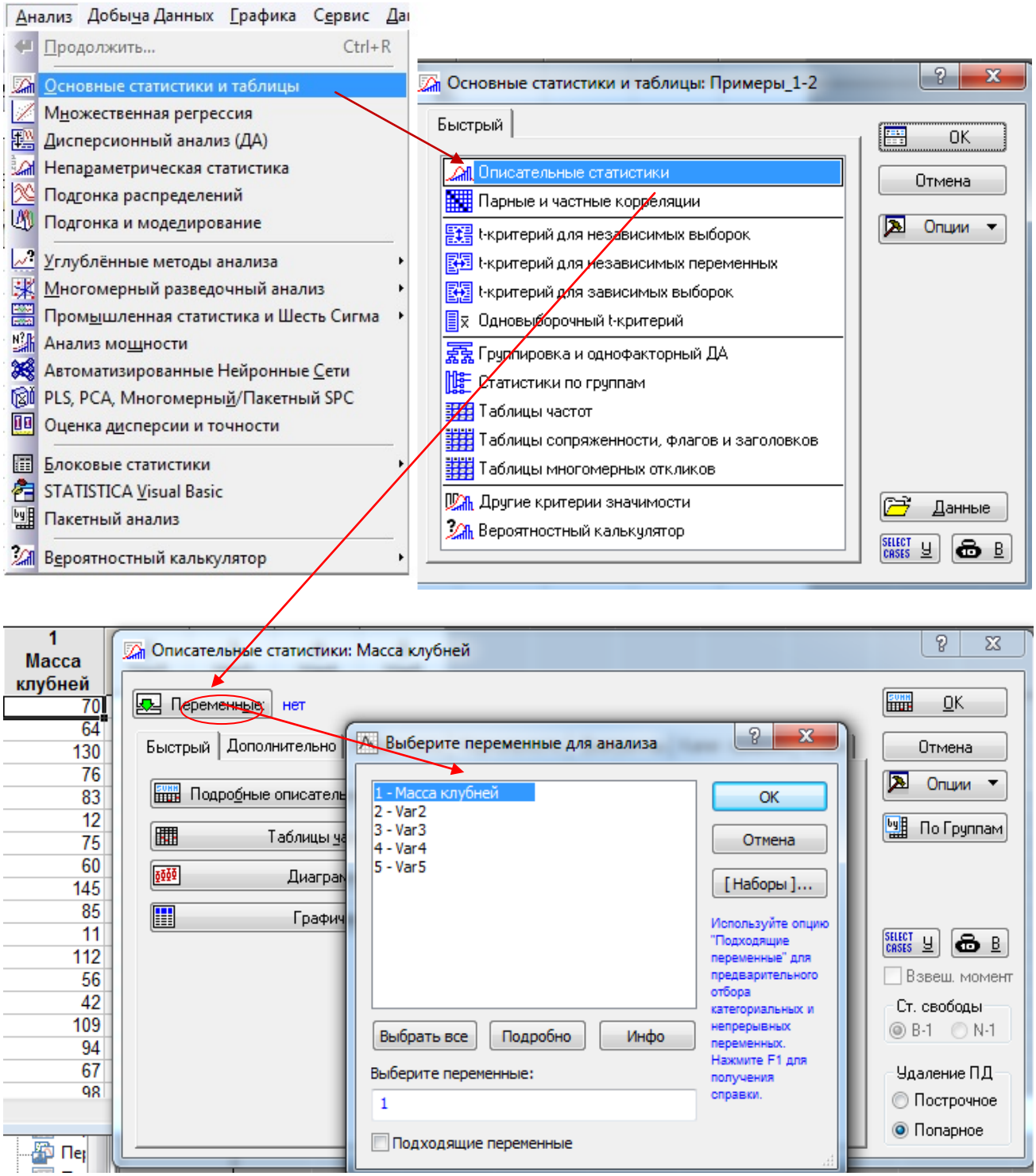


Рис. 2.1. **Последовательность ввода данных в модуле *Описательная статистика***

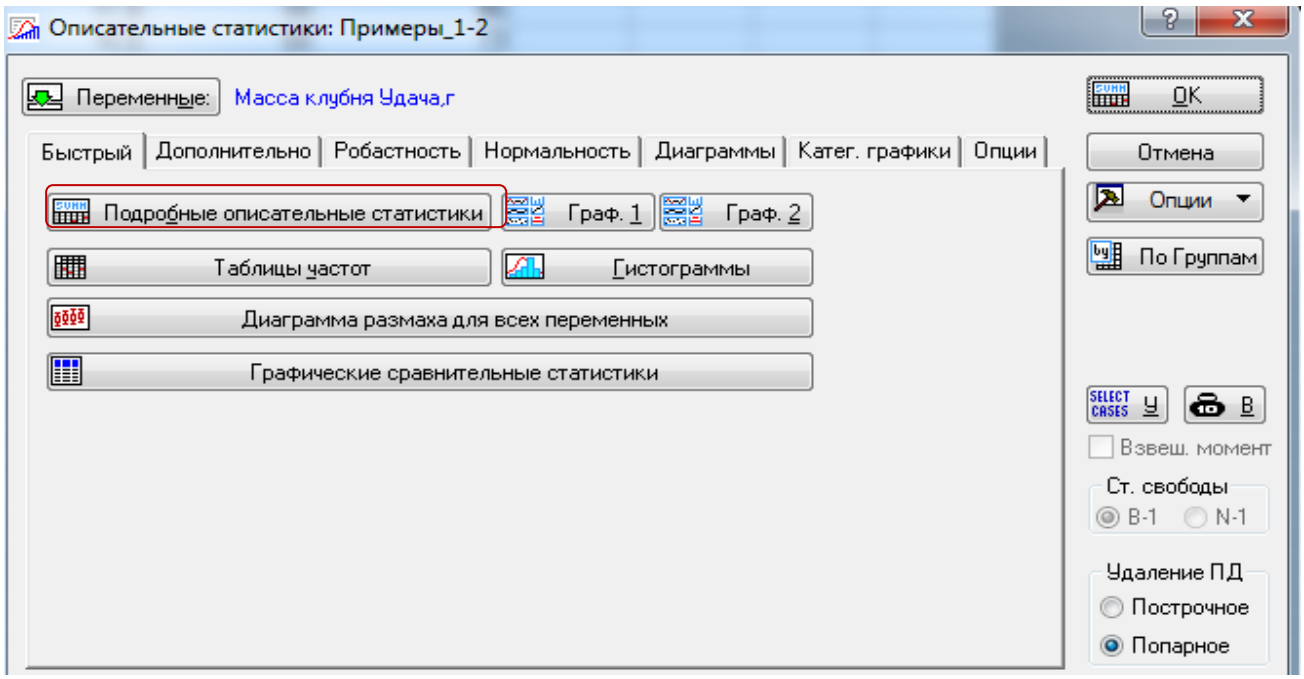


Рис. 2.2. **Диалоговое окно выбора опций описательной статистики**

После нажатия на кнопку **Ок** появляется окно выбора опций описательной статистики (рис. 2.2), в котором выберем **Подробные описательные статистики**, далее в новом диалоговом окне для выбора необходимых статистик (рис. 2.3) нажмем на вкладку **Дополнительно (Advanced)** и с помощью галочек в окошках выберем необходимые статистики. В случае выбора всех показателей нажмем на кнопку **Выбрать все** (рис. 2.3).

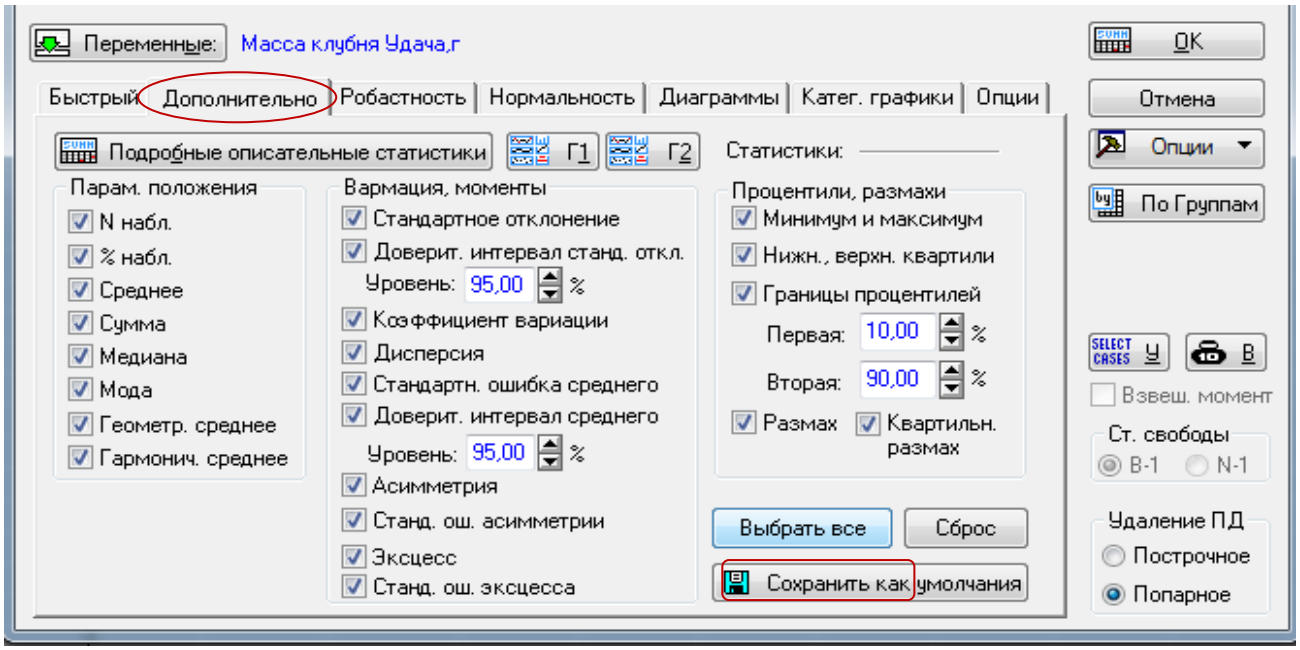


Рис. 2.3. **Окно выбора статистик**

После выбора необходимых статистик и нажатия на клавишу **Ok,** результаты вычислений размещаются в **рабочей книге (Workbook)**(рис.2.4).

Переменная	Описательные статистики (Масса клубней)									
	N набл.	Среднее	Доверит. -95,000%	Доверит. 95,000%	Медиана	Мода	Частота моды	Сумма	Минимум	Максим.
Масса клубней	50	79,12000	69,23709	89,00291	76,00000	Множест.	3	3956,000	9,000000	156,0000

Переменная	Описательные статистики (Масса клубней)										
	Нижняя Квартиль	Верхняя Квартиль	Размах	Дисперсия	Ст.откл.	Коэф.Вар.	Станд. ошибки	Асимметрия	Стд.ош. Асимметрия	Эксцесс	Стд.ош. Эксцесс
Масса клубней	60,00000	99,00000	147,0000	1209,291	34,77487	43,95206	4,917909	0,145351	0,336601	-0,110403	0,661908

Рис. 2.4. **Итоги описательной статистики**

Наибольший интерес из представленных результатов описательной статистики представляют следующие:

Средняя масса одного клубня $\bar{x} = 79{,}12$ *г.*

Медиана *Me = 76,0 г*

Дисперсия $S^2 = 1209,29$

Стандартное отклонение *S= 34,78 г.*

Коэффициент вариации *V= 43,95%*

Стандартная ошибка (ошибка выборочной средней) $S_{\bar{x}} = 4{,}92$ *г.*

95-% доверительный интервал для генеральной средней составляет $\bar{x} \pm t_{05} \cdot S_{\bar{x}} =$ *69,24 ÷ 89,00 г.*

2.2 Диаграммы размаха

Визуально средние значения и меры вариации одной или нескольких выборок (переменных) удобно представлять в виде графиков «ящик с усами» (box-whisker plot). Отдельные элементы графика («центр», «ящик», «усы») различаются для разных видов распределения. Центром такого графика является среднее выборочное значение или медианна, «ящик» показывает границы стандартного отклонения или стандартной ошибки или квартиля. «Усы» показывают границы 95 или 99% доверительного интервала для генеральной средней или всей совокупности или интерквартального размаха.

В программе Statistica предусмотрен выбор четырех форм «ящика с усами»:

1. Медиана / Квартили / Размах (Median/Quart./Range)

2. Среднее / Ошибка среднего / Стандартное отклонение (Mean/SE/SD)

3. Среднее / Стандартное отклонение / Интервал 1,96* стандартного отклонения (Mean/SD/1.96SD)

4. Среднее / Ошибка среднего / Интервал 1,96 * ошибки среднего (Mean/SE/1.96*SE)

Для построения диаграммы размахов в диалоговом окне (рис. 2.5) нажмем на вкладку **Опции (Options)** и попадаем в диалоговое окно выбора разных форм представления графиков «ящик с усами». Выберем графическое изображение доверительных интервалов для всей совокупности и генеральной средней, Для этого отметим галочками опции:

Среднее / Ст. откл./1,96* Ст.откл (Mean/SD/1.96SD)

Среднее / Ст.ош /1,96 * Ст.ош (Mean/SE/1.96*SE)

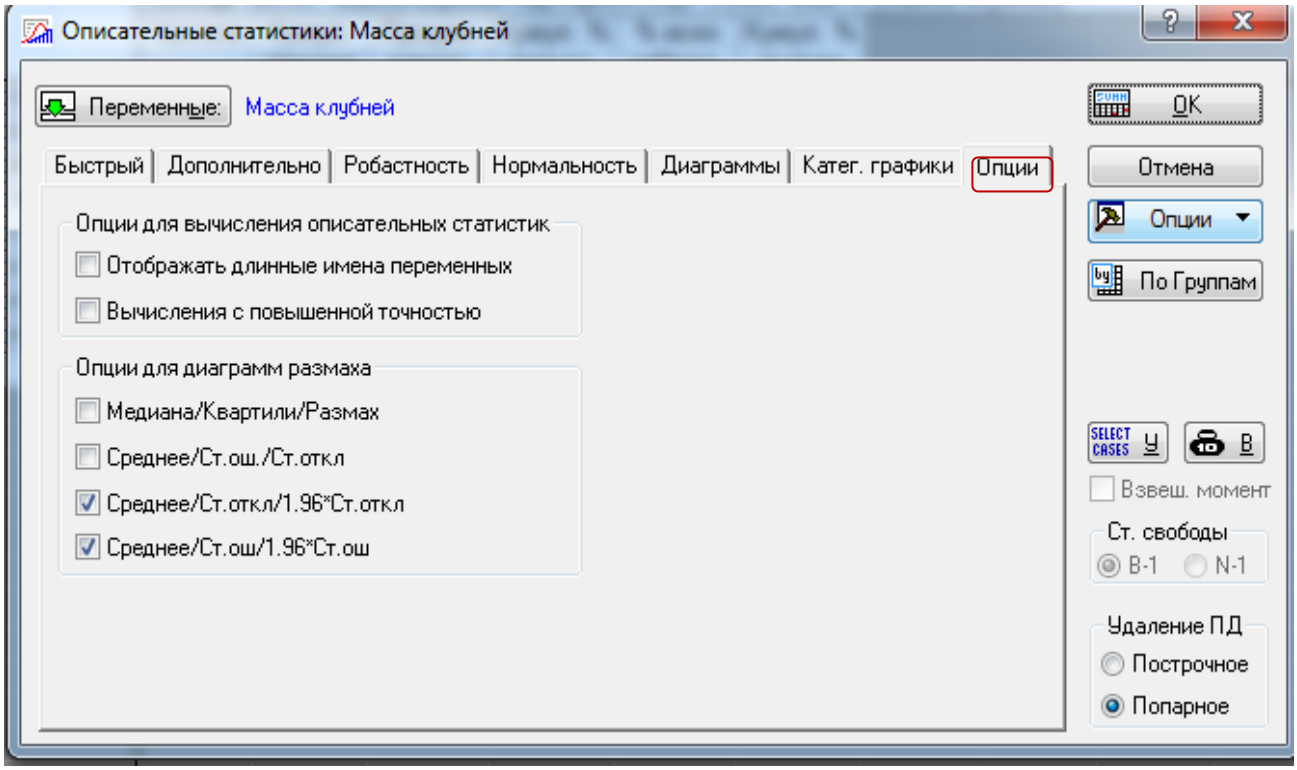


Рис. 2.5. **Выбор формы представления в виде графика «ящик с усами»**

После выбора опций для диаграммы размаха активируем (нажмем) вкладку **Быстрый (Quick)**, нажмем на клавишу **Диаграмма размаха для всех переменных** и нажмем на клавишу **Ок** (рис. 2.6).

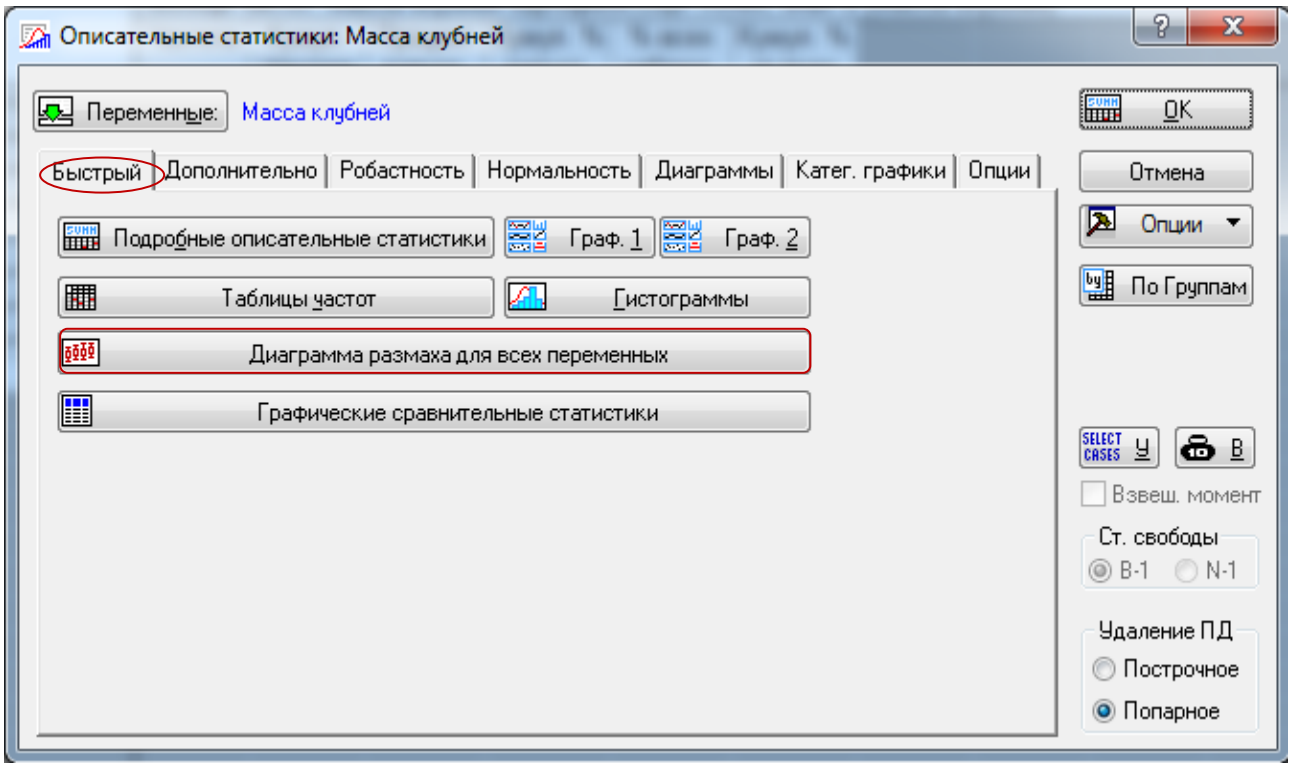


Рис. 2.6. **Выбор опций для диаграммы размаха**

В результате получаем 2 формы диаграммы размаха, которые по сути дела показывают границы доверительных интервалов: слева – 95% доверительный интервал для все совокупности, справа – 95% доверительный интервал для генеральной средней (рис. 2.7).

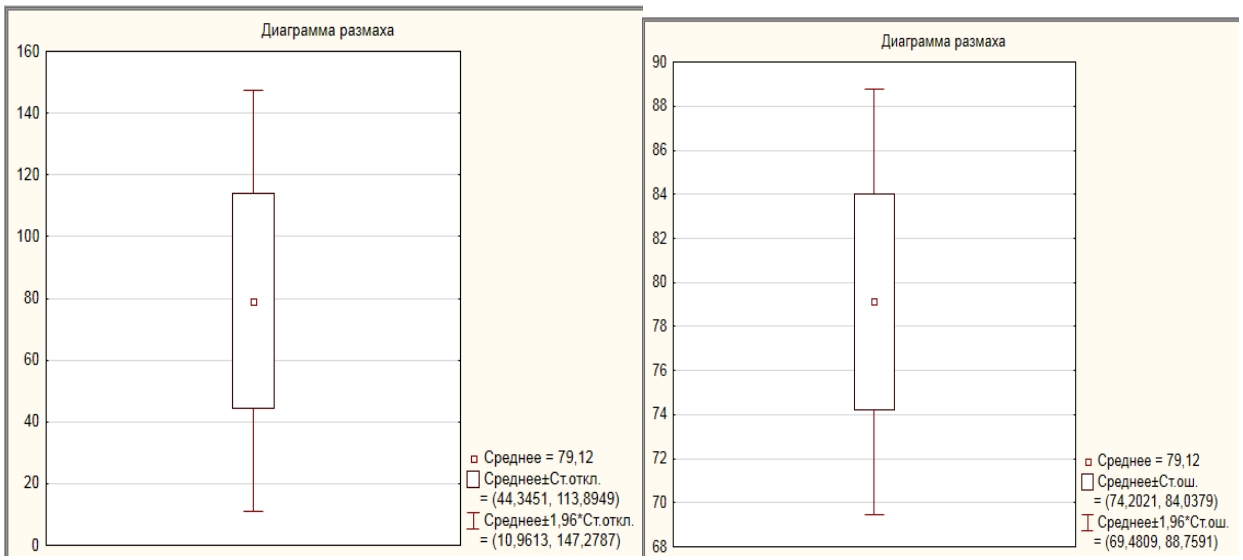


Рис. 2.7. **Диаграммы размаха «ящик с усами» (box-whisker plot) массы клубней картофеля**

Диаграммы размаха можно построить также в меню **Графика (Grafics).** Для этого загрузим данные – файл «Масса клубней», в строке меню щелкнем по кнопке **Графика (Grafics)** и в ниспадающем меню галереи графиков выберем **Графики средних с ошибками** (рис. 2.8). .Открывается новое окно, в котором можно выбрать различные типы графиков виды и представление средних значений с ошибками при 95 или 99% вероятности. Так, в этой опции можно построить «Яшик с усами», как уже было описано в опции **Описательная статистика** или представить средние с ошибками в виде гистограммы.

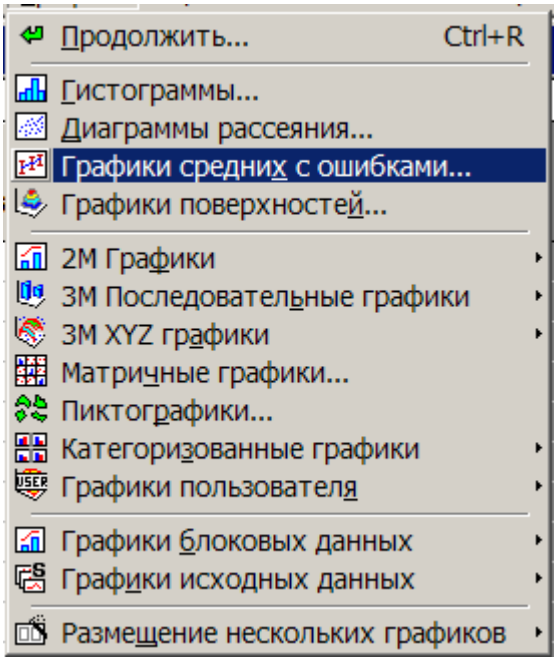


Рис. 2.8. **Меню галереи графиков**

Для представления среднего значения и показателей вариации на графике в виде гистограммы активируем вкладку **Дополнительно (Advanced)** (рис. 2.9). В диалоговом окне **Графики средних с ошибками** в поле Тип графика выберем **Столбцы,** и так как мы имеем дело с одной переменной, в правом поле укажем **Простой,** в поле **Столбец** укажем **Среднее**, в поле выбора **Доверительный интервал (Confidence interval)** укажем вероятность – **95.** Остальные опции оставим по умолчанию.

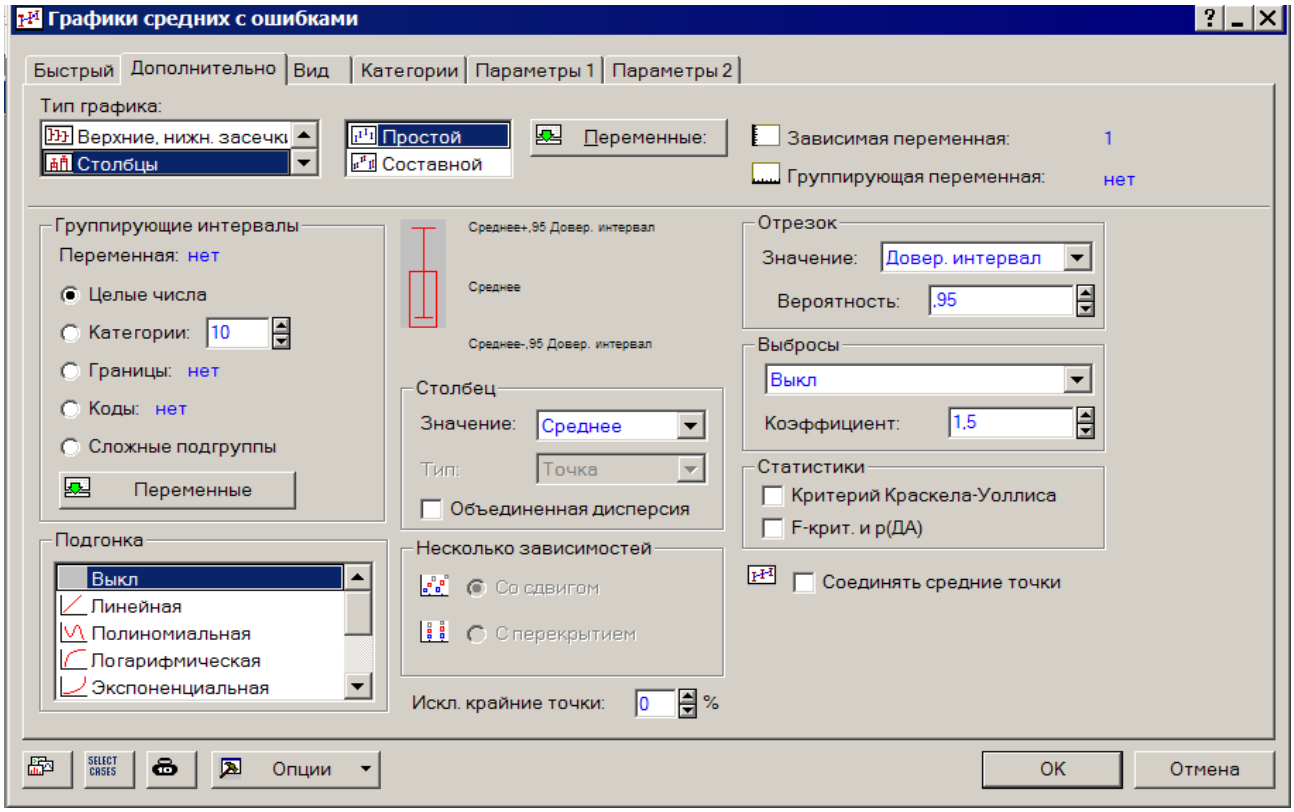


Рис. 2.9. **Диалоговое окно выбора формы графика средних с ошибками**

После нажатия на кнопку **Ok** получаем следующий столбчатый график, на котором высота гистограммы показывает среднюю массу клубней картофеля – 79,12 г., а нижний и верхний усики указывают границы 95% доверительного интервала для генеральной средней (рис. 2.10).

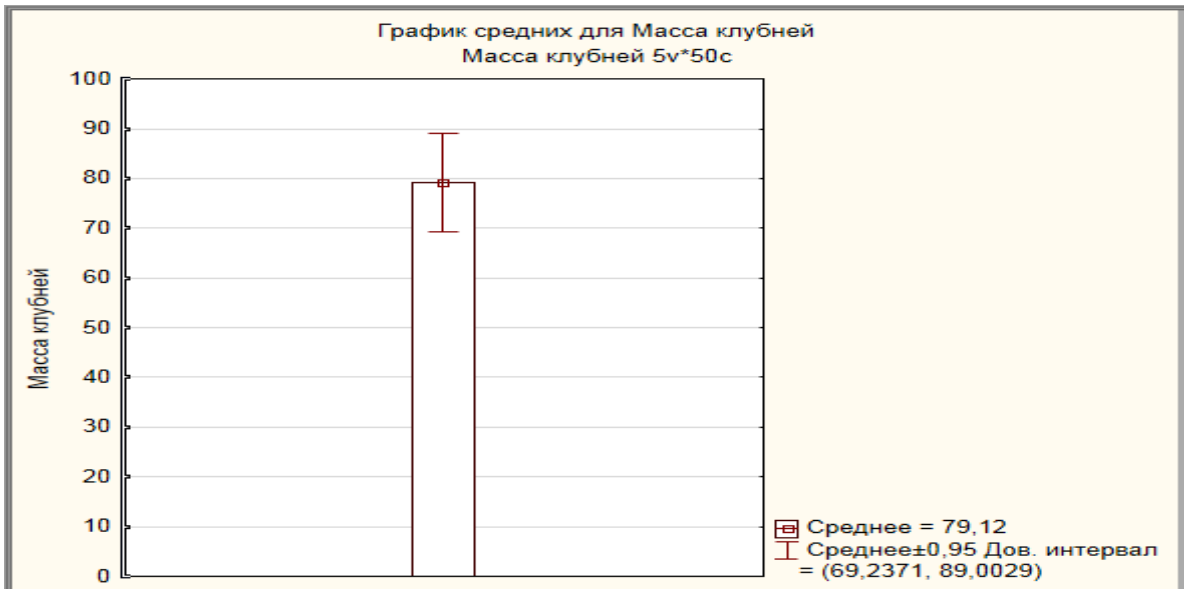
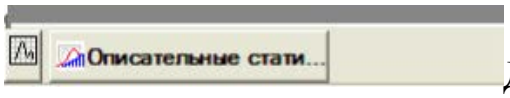


Рис. 2.10. **Гистограмма 95% доверительного интервала для средней массы клубней картофеля в генеральной совокупности**

2.3 Визуализация результатов агрономических исследований

При анализе больших по объему выборок целесообразно провести группировку, что дает возможность представить первичные данные в компактном виде в виде вариационного ряда, по которому можно выявить закономерности варьирования изучаемого признака. Вариационным рядом называется двойной ряд данных, в котором указаны значения варьирующего признака (***X***) и соответствующие им частоты (***f***). Вариационные ряды представляют в виде таблиц и графически – в виде гистограммы. В программе Statistica гистограмму можно построить как через меню **Анализ**, так и в меню **График.**

Построение вариационного ряда в модуле **Описательные статистики (Descriptive statistics).** Если вы закрыли данный модуль расчета, необходимо заново через меню **Анализ** вызвать опцию **Описательные статистики**, далее выбрать **Переменную** – Масса клубней. Если вы не закрыли данный модуль расчетов – в нижнем левом углу экрана находится значок текущего анализа.



Для продолжения необходимо щелкнуть левой кнопкой мыши по этому значку.

В появившемся диалоговом окне (рис. 2.11) активируем вкладку **Быстрый**, далее нажимаем на клавишу **Таблицы частот** и получаем следующую расширенную таблицу вариационного ряда, в которой клубни картофеля разбиты на 5 групп (классов), в первом столбце представлены интервалы каждой группы, во втором столбце – частоты каждой группы значения варьирующего признака, далее накопленные частоты (кумуляты), частоты и кумуляты в % (табл. 1).

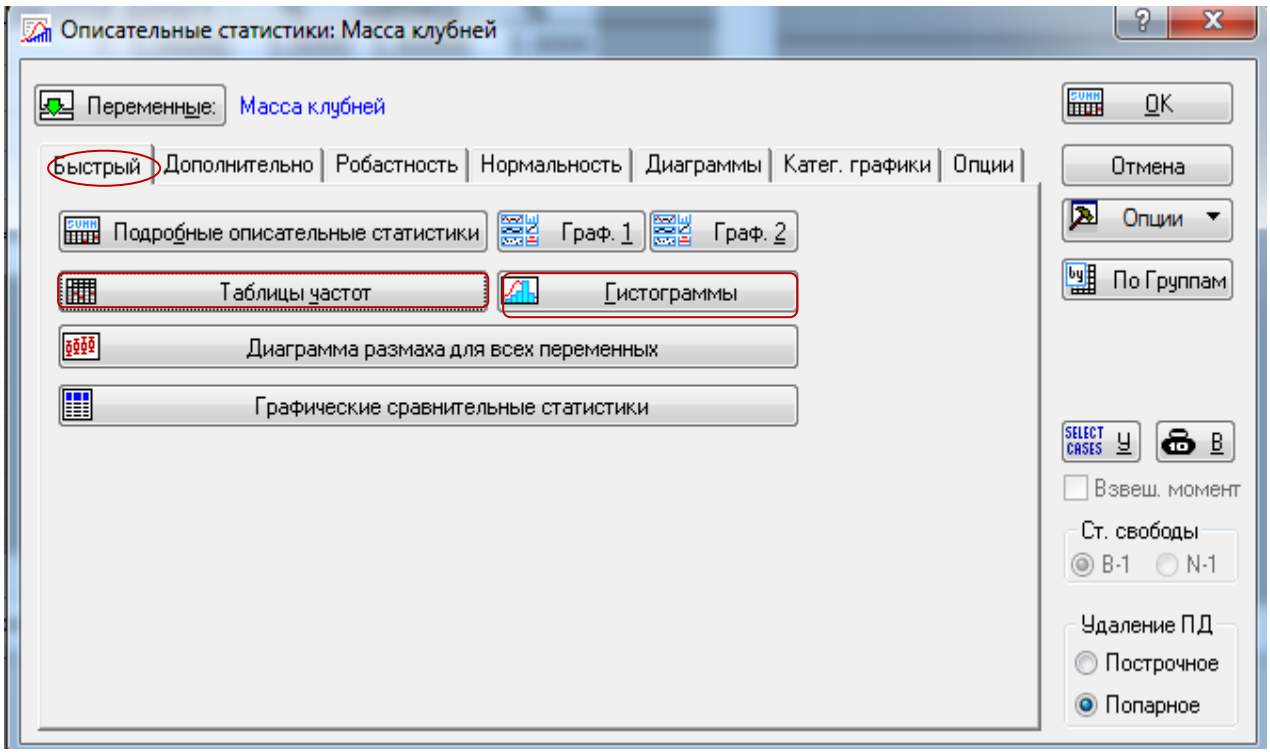


Рис. 2.11. **Выбор опций для таблицы частот и гистограммы**

Таблица 1

Таблица частот 50 клубней картофеля

Группа	Таблица частот: Масса клубней картофеля,Сорт Успех, опыт 8					
	Частота	Кумул. Частота	Процент допуст.	Кумул. % допуст.	% всех наблюд.	Кумул. % от всех
0,000000<x<=20,00000	3	3	6,00000	6,0000	6,00000	6,0000
20,00000<x<=40,00000	4	7	8,00000	14,0000	8,00000	14,0000
40,00000<x<=60,00000	6	13	12,00000	26,0000	12,00000	26,0000
60,00000<x<=80,00000	14	27	28,00000	54,0000	28,00000	54,0000
80,00000<x<=100,0000	11	38	22,00000	76,0000	22,00000	76,0000
100,0000<x<=120,0000	5	43	10,00000	86,0000	10,00000	86,0000
120,0000<x<=140,0000	4	47	8,00000	94,0000	8,00000	94,0000
140,0000<x<=160,0000	3	50	6,00000	100,0000	6,00000	100,0000

Для представления массы 50 клубней картофеля в виде гистограммы нажмем в диалоговом окне на клавишу **Гистограммы (Histograms)** и получаем следующий график (рис.2.12). Столбики гистограммы представляют фактические частоты, а красной линей показана плотность нормального распределения, что дает возможность проверить близость фактического распределения нормальному.

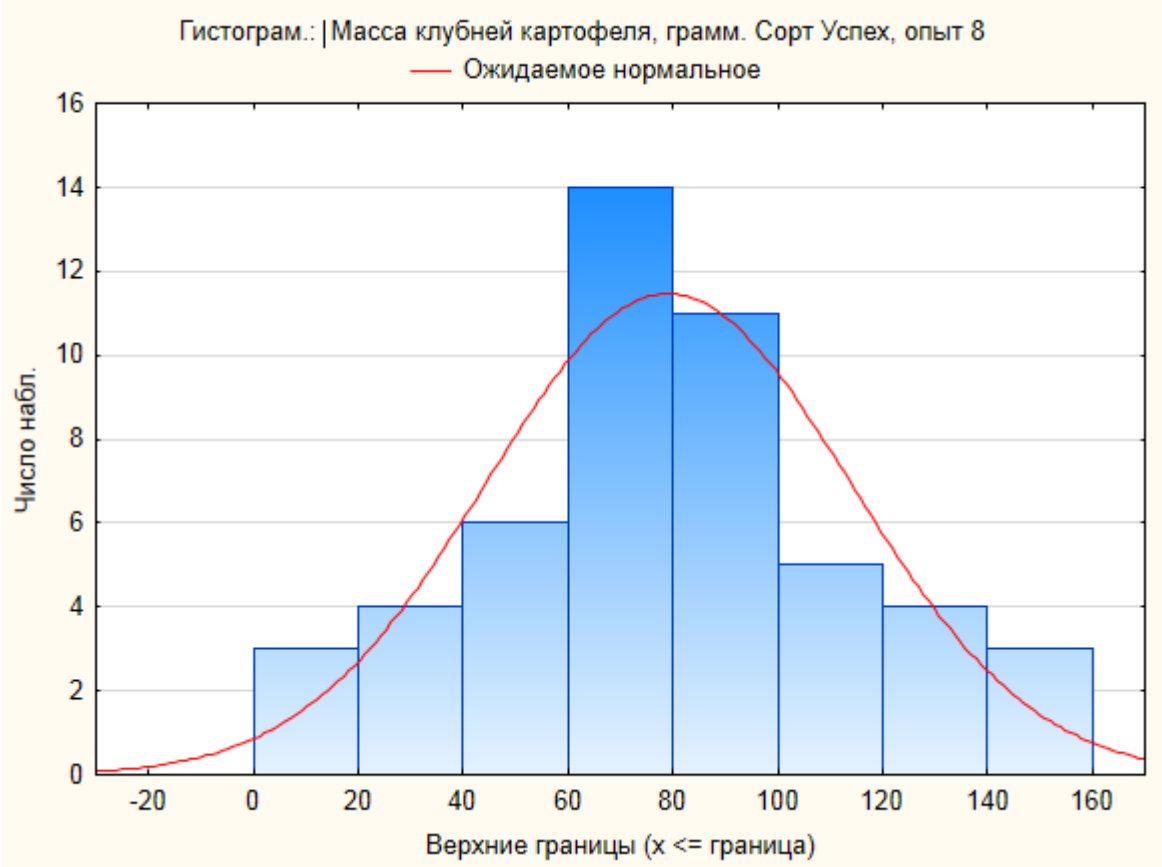


Рис. 2.12. **Гистограмма массы 50 клубней картофеля**

Контрольные вопросы:

1. Статистические показатели количественной изменчивости.
2. В чем отличие средней арифметической от медианы?
3. В каких случаях медиана предпочтительнее средней выборочной?
4. Перечислите показатели изменчивости.
5. Каким образом рассчитываются в пакете Statistica основные статистические показатели?
6. Что такое «ящик с усами»?
7. Какие статистические показатели можно показать на графике «ящик с усами»?
8. Что такое вариационный ряд? Какие бывают вариационные ряды?
9. Как построить гистограмму в программе Statistica?
10. Для чего проводится группировка данных?

Глава 3. ПРОВЕРКА СООТВЕТСТВИЯ АНАЛИЗИРУЕМЫХ ДАННЫХ НОРМАЛЬНОМУ РАСПРЕДЕЛЕНИЮ

При статистической обработке экспериментальных данных важной процедурой является проверка гипотезы о соответствии распределения изучаемого признака нормальному распределению. Нормальность наблюдаемых данных является необходимым условием для корректного описания данных и применения параметрических методов и критериев.

Важность проверки распределения эмпирических данных выборки нормальному распределению заключается в том, что эта процедура определяет дальнейшую тактику исследователя при обработке данных агрономических исследований. Если фактическое распределение соответствует нормальному, то для дальнейшей статистической обработки (оценка 2-х вариантов, дисперсионный, корреляционно-регрессионный анализы и др.) будут применяться параметрические критерии для проверки нулевой гипотезы. В случае несоответствия эмпирических данных нормальному или близкому к нему распределениям для анализа данных одной выборки следует использовать медиану и квартили, а для сравнения нескольких выборок использовать непараметрические критерии для проверки гипотез.

Предположение о нормальности можно проверить, исследуя распределение визуально с помощью гистограмм или на основе критериев нормальности.

3.1 График нормальных вероятностей

Модуль **Описательные статистики (Descriptive statistics)** предлагает несколько графических процедур для анализа распределения переменных. О нормальности распределения можно судить по графику нормальных вероятностей или так называемой «вероятностной бумаги». Такой график изображает зависимость ожидаемых нормальных частот значений признака от их фактических частот. Для построения графика в диалоговом окне **Описательные статистики (Descriptive statistics)** активируем вкладку **Диаграммы** (рис. 3.1), затем выберем опцию **Нормальные вероятностные**

графики (Normal probability plots) и получаем следующий график (рис.3.2). Чем ближе распределение к нормальному виду, тем лучше значения ложатся на прямую линию. Так как все 50 точек очень близко и симметрично расположены вокруг прямой линии, можно с очень высокой долей вероятности сделать вывод о соответствии массы 50 клубней картофеля нормальному распределению.

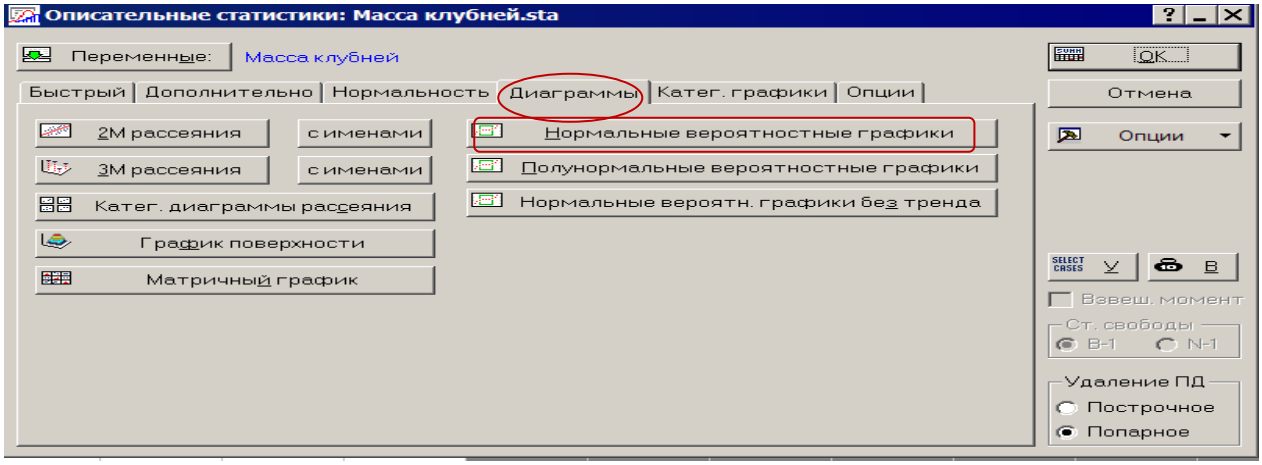


Рис. 3.1. **Диалоговое окно выбора графика вероятности**

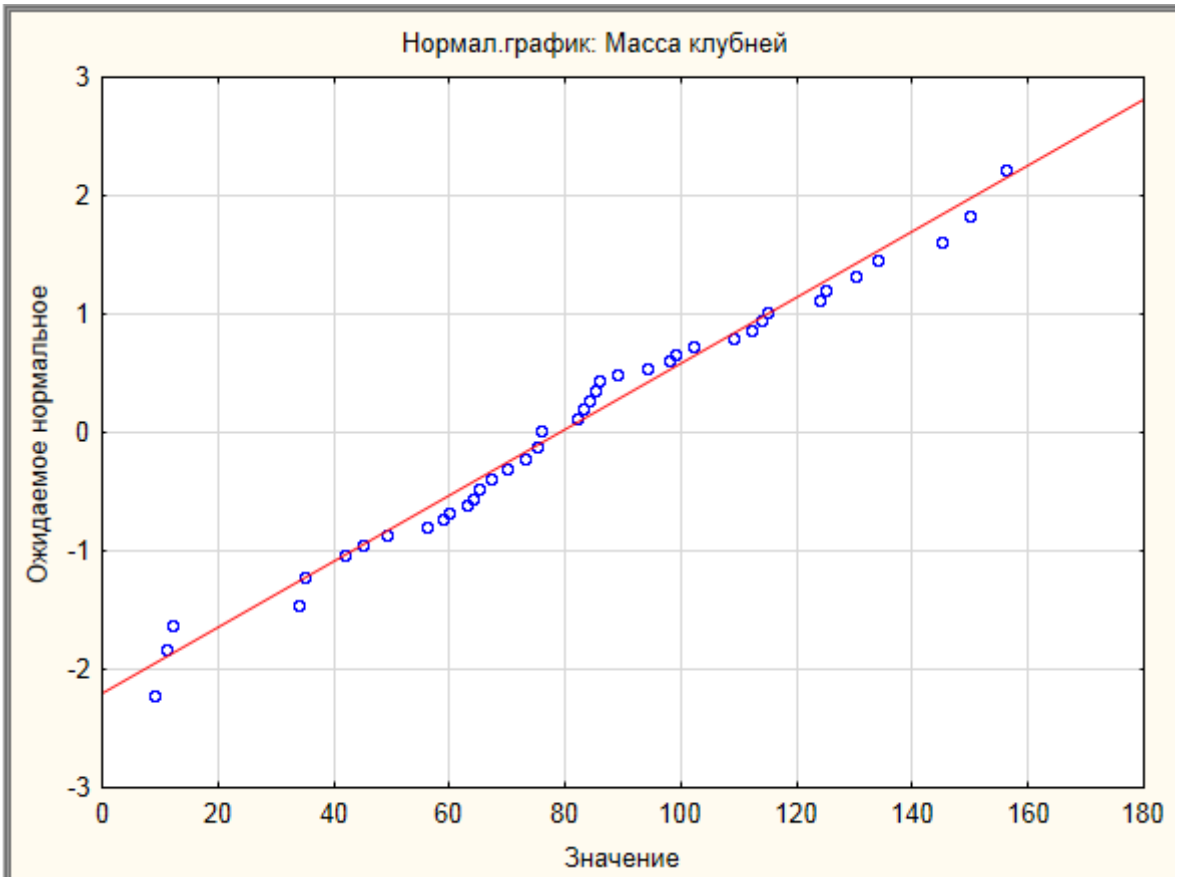


Рис. 3.2. **Нормальный вероятностный график остатков 50 клубней картофеля**

Вместе с тем этот метод оценки является фактически глазомерным, поэтому для более точного сравнения проведем проверку на нормальность распределения с использованием статистических критериев.

3.2 Критерии соответствия нормальному распределению

Для проверки соответствия изучаемого признака нормальному закону существует несколько критериев: Шапиро-Уилкса, Колмогорова-Смирнова, Лиллиефорса, Пирсона и др.

Коэффициент асимметрии (skewness) является критерием проверки на симметричность, который в случае нормального распределения равен 0. Стандартизованный коэффициент асимметрии так же равен 0.

Коэффициент эксцесса (kurtosis) является критерием проверки на эксцесс и отражает «остроту пика». В случае нормального распределения значение коэффициента эксцесса равно 3, значение стандартизованного коэффициента равно 0.

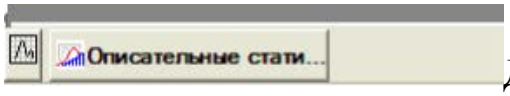
Критерий Шапиро-Уилкса - *W* рекомендуют применять при отсутствии априорной информации о типе возможного отклонения от нормальности. Если коэффициент Шапиро-Уилкса стремится к 1 при любом значении *p*, то фактическое распределение соответствует нормальному. Если коэффициент Шапиро-Уилкса стремится к 0 при $p < 0,05$, то исследуемое распределение отличается от нормального.

Критерий Колмогорова-Смирнова (*d*). Чем меньше величина этой статистики, тем ближе распределение случайной величины к нормальному.

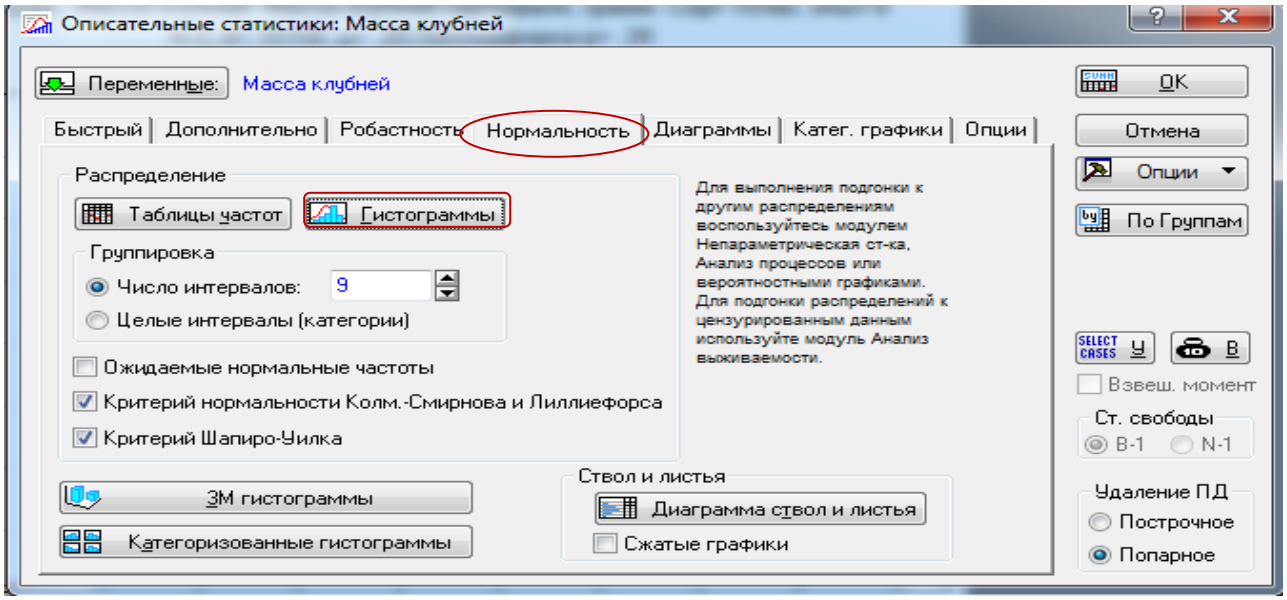
Критерий Лиллиефорса используется когда параметры нормального распределения априори неизвестны. Критические значения максимального отклонения выборочной функции от теоретической указаны в специальных таблицах.

Критерий согласия (подобия) Пирсона χ^2-*Хи-квадрат*. Если фактическое значение критерия *Хи-квадрат* меньше табличного, исследуемое распределение существенно не отличается от нормального распределения.

Проанализируем значения массы клубней картофеля на соответствие их нормальному распределению, для этого щелкнем левой кнопкой в нижнем левом углу экрана по значку:



Далее попадаем в диалоговое окно **Описательной статистики**, в котором активируем вкладку **Нормальность** и галочками отметим критерии для проверки нормальности распределения: критерий нормальности Колмогорова-Смирнова и Лиллиефорса и критерий Шапиро-Уилкса (рис. 3.3).



3.3. **Диалоговое окно выбора критериев на нормальность**

Нажмем на кнопку **Гистограммы** и получаем такую же гистограмму как на рис. 2.12, но в заголовке приведены значения критериев для проверки гипотезы о проверке на нормальность. Так как критерий Колмогорова-Смирнова $d = 0{,}10158$ $p>0{,}20$; критерий Шапиро-Уилка $W=0{,}98$, $p=0{,}53$ (для всех критериев p>0,05), можно сделать вывод о том, что распределение массы клубней картофеля с вероятностью 95% соответствует нормальному распределению.

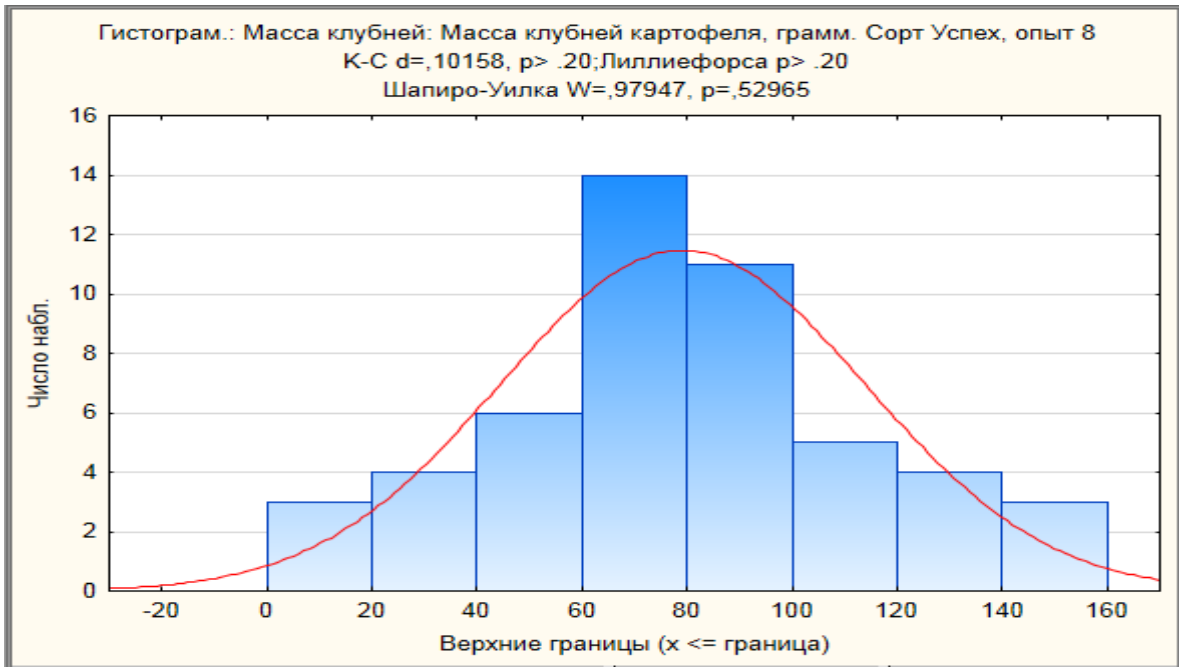


Рис. 3.4. **Гистограмма и теоретическая кривая нормального распределения 50 клубней картофеля**

3.3 Теоретические распределения. Подгонка распределения

В программе Statistica имеется процедура построения наиболее адекватной статистической модели **Подгонка распределений (Distribution Fittting),** позволяющая проверить соответствие данных фактических наблюдений теоретическим распределениям: нормальному, биноминальному, Пуассона логнормальному, гамма и др. Проверка гипотезы обычно проверяется с помощью критерия Хи-квадрат и критерия Колмогорова-Смирнова.

Для проверки соответствия массы клубней картофеля теоретическим распределениям выберем в меню **Анализ** модуль **Подгонка распределений (Distribution Fittting)** и попадаем в окно выбора типов распределения (рис. 3.5).

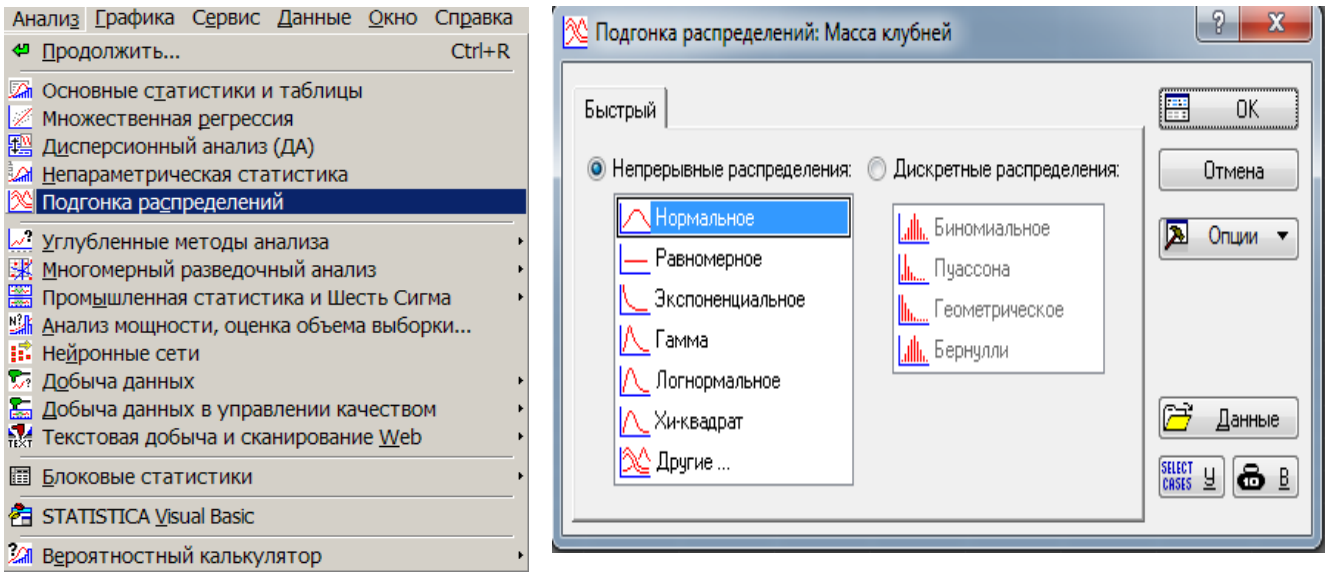


Рис. 3.5. **Окно выбора типа распределения**

Так как масса клубней картофеля относится к непрерывной изменчивости, в окне **Непрерывные распределения (Continuous Distributions)** выберем **Нормальное (Normal)** (рис. 3.5). Для данных дискретной (прерывистой) изменчивости выбирается **Биноминальное распределение (Binomial Distributions).** В появившемся окне (рис. 3.6) выберем переменную *Масса клубней*, активируем вкладку **Быстрый (Quick)** и нажмем на кнопку **График наблюдаемого и ожидаемого распределения**

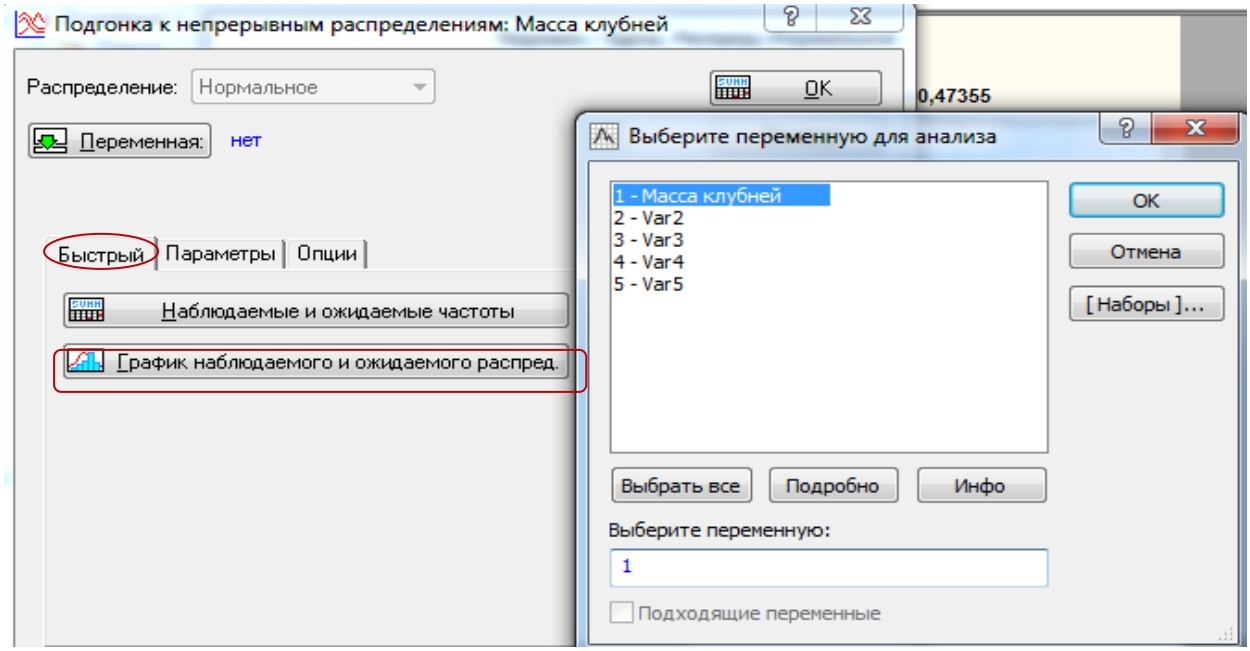


Рис. 3.6. **Диалоговое окно выбора переменной для проверки соответствия теоретическим распределениям**

В появившемся окне выбора опций нажмем на вкладку **Параметры (Parametrs)** и в окошках устанавливаем параметры для построения гистограммы: число классов (*7, k=√n±2*), нижнюю (X_{min} = *9*) и верхнюю (X_{max} = *156*) границу классов. Далее нажмем на вкладку **Опции** и в появившемся окне отмечаем нужные опции для расчета критерия Колмогорова-Смирнова критерия Пирсона как показано на рис. 3.7. Переходим на вкладку **Быстрый(Quick)** и нажмем на клавишу **График наблюдаемого и ожидаемого распределения** и на кнопку **Ok**.

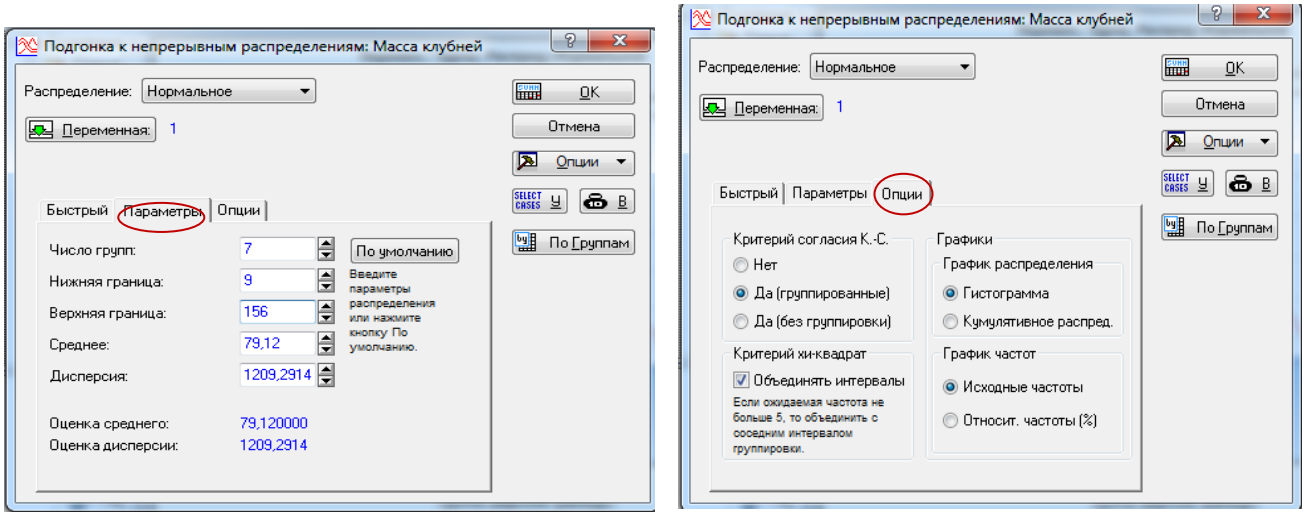


Рис. 3.7. **Диалоговое окно выбора параметров и опций**

После нажатия на кнопку **Ок** получаем график (рис.3.8).

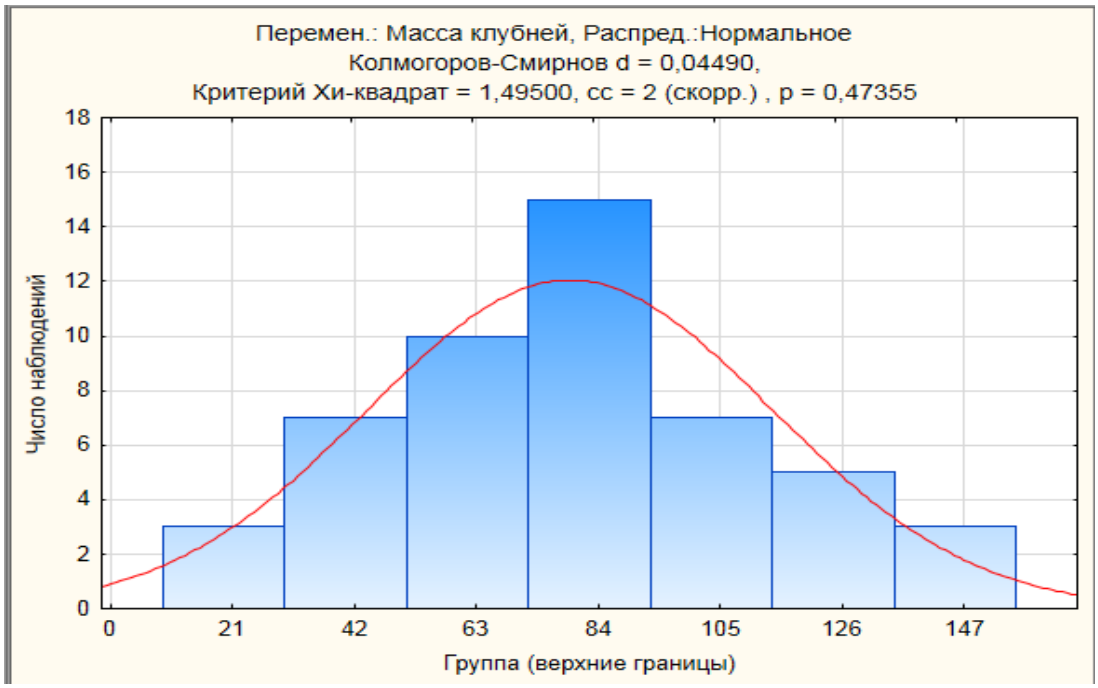


Рис. 3.8. **Гистограмма распределения 50 клубней картофеля**

На рис. 3.8 показана гистограмма распределения 50 клубней картофеля. Ожидаемая кривая нормального распределения при текущих значениях среднего ($\bar{x} = 79,12$ *г*) и дисперсии ($S^2 = 1209,29$) изображена красным цветом. Первичная визуальная оценка показывает, что фактическое распределение в целом симметрично и не отличается от нормального распределения. Для проверки нулевой гипотезы о соответствии распределения клубней картофеля нормальному проанализируем рассчитанные критерии. Результаты выбранных тестов на нормальность автоматически располагаются в заголовке этого графика. При ***p > 0,05*** можно заключить, что анализируемое распределение не отличается от нормального. Так как *критерий Колмогорова-Смирнова* ***d=0,04490***, а *критерий Хи-квадрат*=***1,495*** при ***p>0,05*** (***p=0,47355***), с вероятностью 95% принимается нулевая гипотеза: масса клубней картофеля соответствует нормальному распределению.

Контрольные вопросы:

1. Эмпирические (наблюдаемые) и теоретические (ожидаемые) распределения.
2. Что такое нормальное распределение? Каковы закономерности кривой нормального распределения?
3. С какой целью проводится проверка данных на соответствие их нормальному распределению?
4. Для чего служит нормальный вероятностный график?
5. С помощью каких критериев оценивается соответствие нормальному распределению?
6. Как проверить соответствие эмпирического ряда распределения нормальному?
7. Как по критерию Шапиро-Уилкса проверить гипотезу о нормальности распределения?
8. Какие инструменты используются в программе Statstica для проверки гипотезы о соответствии эмпирических рядов нормальному?

Глава 4. СРАВНЕНИЕ ДВУХ ВАРИАНТОВ (ПРОВЕРКА НУЛЕВОЙ ГИПОТЕЗЫ)

В агрономических исследованиях часто необходимо оценить существенность или надежность различий между двумя группами данных, обычно между двумя или несколькими вариантами. Все статистические методы направлены на проверку нулевой гипотезы, иными словами, они оценивают ее правомочность, т.е. справедливость. Для проверки нулевой гипотезы применяются статистические критерии, которые делятся на параметрические и непараметрические. К параметрическим относятся критерии, включающие в формулу расчета параметры распределения, т.е. среднее и дисперсии (критерии Стьюдента, Фишера). К непараметрическим относят критерии, основанные на рангах и не включающие в формулу расчета параметры распределения (критерии Колмогорова-Смирнова, Уилкоксона, Манна-Уитни, знаков и др.).

Существенность различий оценивается, как правило, сравнением фактических значений выбранных критериев с критическим значением. Надежность вывода о различиях оценивается по другому показателю – «уровню значимости». В биологических, в том числе и агрономических исследований, считается достаточным уровень значимости $\alpha = 0,05$ (это означает, что в 95 случаях из 100 оцениваемое нами явление произойдет). В программе Statistica, **уровень значимости α обозначается *p*-level** (т.е. уровень значимости для проверки нулевой гипотезы)**. Если $p \geq 0,05$ => Но - принимается, если $p < 0,05$ => Но – отвергается.** Более низкий ***p*-*уровень*** соответствует более высокому уровню доверия. Если в качестве критического значения вместо *0,05* взять значение *0,01*, то надежность результатов возрастает.

Самой распространенной задачей в агрономических и биологических исследованиях является сравнение арифметических средних двух вариантов. Классическим методом, позволяющим ее решать, является ***t-тест Стьюдента,*** или просто «***t-тест***». Нулевая гипотеза, проверяемая в ходе данного теста, заключается в том, что оба варианта (выборки) происходят из одной

генеральной совокупности; другими словами, что наблюдаемые различия между средними значениями сравниваемых выборок случайны и не вызваны действием изучаемого фактора. Критерий Стьюдента относится к группе параметрических методов анализа, поэтому для корректного его применения необходимо, чтобы обе выборки подчинялись закону нормального распределения. В случае невыполнения данного условия оценить различие между средними возможно с помощью подходящего непараметрического критерия. Способы проверки нормальности распределения описаны в предыдущей главе.

Ниже описывается, как проверяется нулевая гипотеза с использованием критерия Стьюдента в программе Statistica. При этом выборки могут быть независимыми (несопряженными) или зависимыми (сопряженными). Кроме того, данные могут относиться к непрерывной или прерывистой изменчивости.

4.1 Сравнение средних независимых выборок при количественной изменчивости

Сравнение средних независимых выборок называется процедурой **t-критерий для независимых выборок (t-test for independent samples)**. При сравнении двух вариантов объемы выборок могут равными ($n_1 = n_2$) или разными ($n_1 \neq n_2$).

Пример 1. Содержание белка (%) в зерне при испытании двух сортов яровой пшеницы Иргина и Комета. $n_1 = 20$, $n_2 = 18$.

Сорт	Содержание белка в зерне, %																			
Иргина	18,6	17,3	16,9	20,0	17,9	18,3	17,4	18,4	19,6	18,0	19,0	19,3	18,6	19,5	18,0	17,0	19,2	18,4	19,5	18,0
Комета	17,8	16,6	17,0	19,5	16,5	17,0	17,1	16,4	18,5	17,3	19,0	18,1	17,6	17,3	16,9	16,9	17,4	16,7		

Необходимо определить существенны ли различия в содержании белка между сортами?

Создадим файл исходных данных с 2-мя переменными *Иргина* и *Комета*, как указано в подразделе 1.2. Ввод и редактирование данных.

В разделе основного меню **Анализ (Statistics)** выберем модуль **Основные статистики и таблицы (Basic Statistics / Tables),** открывается диалоговое окно (рис. 4.1). Эта процедура используется для установления достоверной статистической разницы между средними значениями выборок ($d = \bar{x}_2 - \bar{x}_1$) на основе t-критерия Стьюдента.

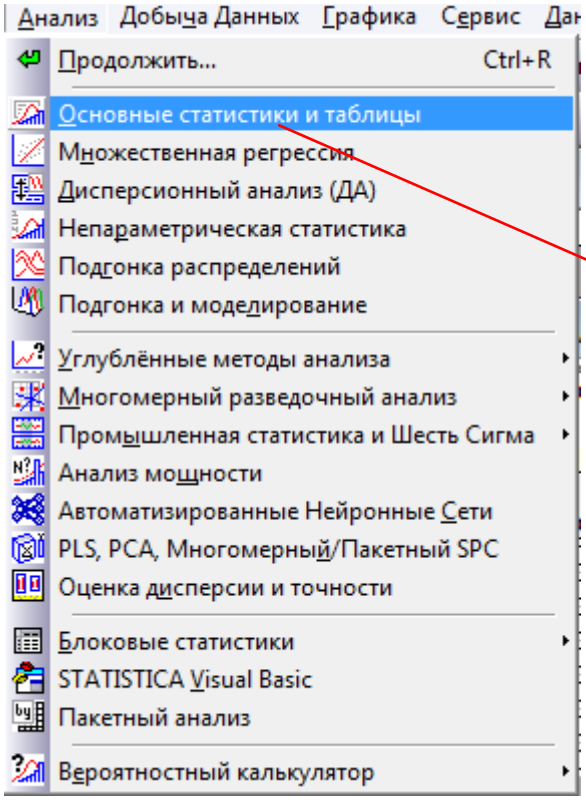


Рис. 4.1. **Меню *Анализ***

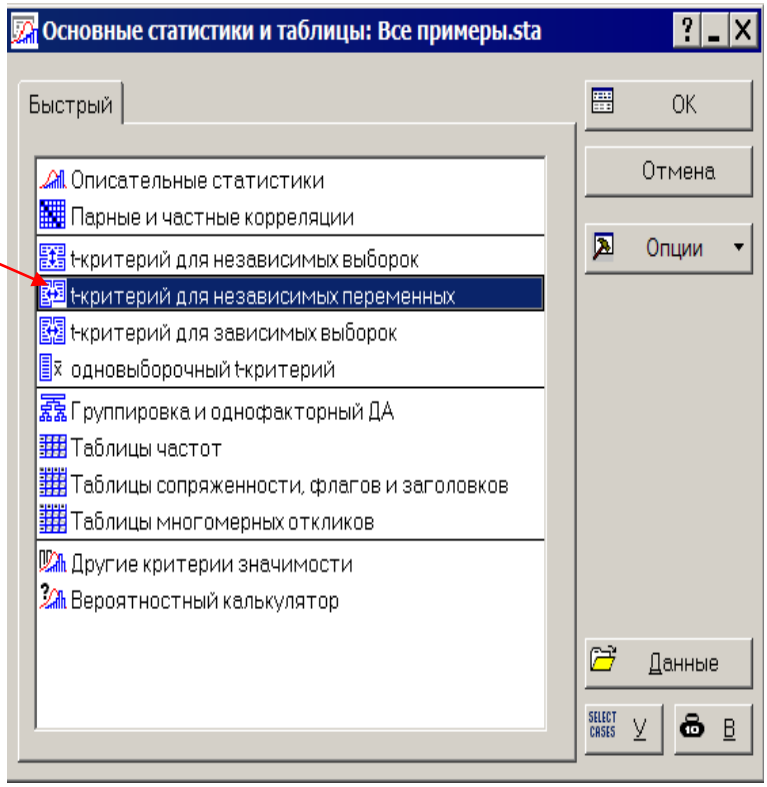


Рис.4.2. **Стартовая панель *Основные статистики и таблицы***

Выберем процедуру **t-критерий для независимых переменных (t-test for independent variables)**(рис. 4.2).

В появившемся диалоговом окне необходимо указать, какие переменные мы собираемся анализировать. Щелкнем по кнопке **Переменные (группы) (Variable)**, чтобы открыть стандартное диалоговое окно для выбора переменных. Появляется окно выбора переменных, в первом списке двойным щелчком указываем переменную Иргина, во втором – Комета (рис. 4.3).

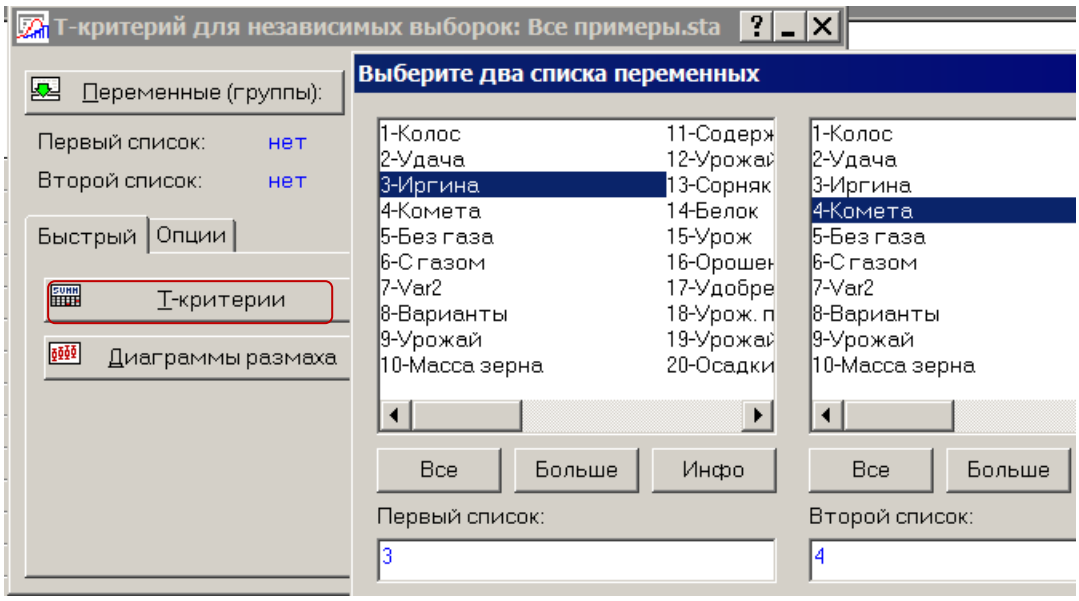


Рис. 4.3. **Окно выбора переменных**

После появления в диалоговом окне выбранных переменных, нажмем на кнопку **Ок**, чтобы вернуться в диалоговое окно **Т-критерий для независимых выборок.** В появившемся диалоговом окне нажимаем на кнопку **Т-критерии** и в рабочем книге получаем итоговую таблицу (рис. 4.4), в которой представлены результаты статистического анализа: среднее содержание белка у сорта Иргина – *18,44*, у сорта Комета – *17,42%*, фактическое значение критерия $t_{фак}$ = ***3,56.***

Т-критерий независимых выборок (Все примеры)
Замечание: Переменные рассм. как независимые выборки

Среднее Группа 1	Среднее Группа 2	t-знач.	cc	p	N набл. Группа 1	N набл. Группа 2	Ст.откл. Группа 1	Ст.откл. Группа 2	F-отн. дисперс.	p дисперс.	Среднее 1 Среднее 2	Доверит. -95,000%	Доверит. +95,000%
18,44500	17,42222	3,559352	36	0,001066	20	18	0,901154	0,865384	1,084378	0,872686	1,022778	0,440006	1,605549

Рис. 4.4. **Итоговая таблица результатов сравнения 2-х сортов пшеницы по содержанию белка**

По результатам итоговой таблицы проверку нулевой гипотезы (***H_0 :d = 0***) можно провести 3 способами.

1. Так как фактическое значение критерия Стьюдента (*t-знач = 3,56*) в нашем примере *t-знач.* больше табличного (*t-критическое двухстороннее = 2,10* при числе степеней свободы *df=20+18-2=36*), H_0 отвергается – между сортами

по содержанию белка имеются существенные различия. И наш вывод будет справедлив не только с 95% вероятностью, но и с вероятностью 99%.

2. Уровень p-значимости для *t-критерия* равен вероятности ошибочно отвергнуть гипотезу о равенстве средних двух выборок, когда в действительности эта гипотеза имеет место, в нашем случае *p=0,001066*. То есть, отвергая нулевую гипотезу о равенстве содержания белка у сортов Иргина и Комета на уровне 0,001, мы рискуем ошибиться на 0,1%.

В программе Statistica для всех типов задач, если нулевая гипотеза отвергается с заданным в опциях уровнем значимости, то цифры в таблице будут показаны красным цветом.

3. Существенность (значимость) различий между средними можно оценить также по доверительному интервалу для генеральной разности $\boldsymbol{d \pm t_{05} \cdot S_d}$ = 0,44÷1,61. Для нашего примера разность между средними *d=1,022778*, нижняя граница доверительного интервала для генеральной разности *0,44006* (предпоследняя колонка), верхняя граница – *1,605549* (последняя колонка). Так как нижняя граница доверительного интервала для генеральной разности *>0*, нулевая гипотеза с вероятностью 95% отвергается. $\boldsymbol{H_0: d \neq 0}$.

При сравнении средних, как и анализе других данных, чрезвычайно полезны визуальные методы, в частности диаграммы размаха. Для представления средних значений на графике нажмем на кнопку **Диаграммы размаха** и в рабочей книге получаем 2 «ящика с усами» (рис. 4.5), где усы показывают границы доверительных интервалов для генеральных средних. Так как нижняя граница доверительного интервала для генеральной средней сорта Иргина не пересекается с верхней границей доверительного интервала для генеральной средней сорта Комета, нулевая гипотеза о равенстве генеральных средних отвергается. Это свидетельствует о том, что различия между этими средними значимы с вероятностью 95%.

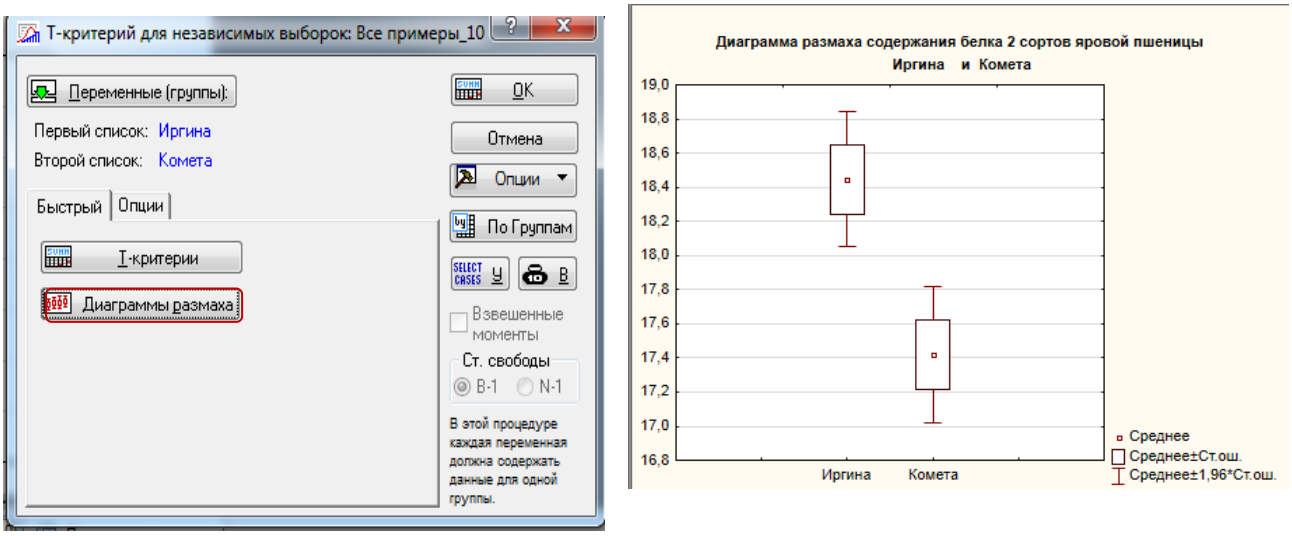


Рис. 4.5. **Диаграмма размаха содержания белка сортов Иргина и Комета**

Таким образом, оценка существенности разности средних по критерию t-Стьюдента, уровню значимости (***p***), доверительному интервалу для генеральной разности и диаграмме размаха (доверительным интервалам для генеральных средних) показывают, что нулевая гипотеза с вероятностью 95% отвергается и принимается альтернативная гипотеза: между сортами яровой пшеницы Иргина и Комета отмечается существенные различия по содержанию белка.

4.2 Сравнение средних зависимых выборок при количественной изменчивости

Пример 2. При изучении 2-х способов хранения груши в полиэтиленовых пакетах в одних и тех же камерах холодильника процент сохранившихся плодов составил,%:

Способы хранения	Число пар наблюдений, n									
	1	2	3	4	5	6	7	8	9	10
Без газовой среды	56	68	74	75	80	56	63	66	75	64
Газовая среда	85	73	75	95	78	85	76	74	83	60

Следует определить существенны ли различия в сохранности плодов груши при разных способах их хранения?

Создадим файл исходных данных с 2-мя переменными *Без газа* и *С газом*, как указано в подразделе **1.2. Ввод и редактирование данных**.

Так как плоды груши хранились в одних и тех же холодильных камерах, переменные данного примера относятся к зависимым или сопряженным выборкам, поэтому будем оценивать не разность средних, а среднюю разность

$$\overline{d} = \frac{\sum d}{n}$$

В стартовой панели **Основные статистики и таблицы** выберем опцию **t-критерий для зависимых выборок**

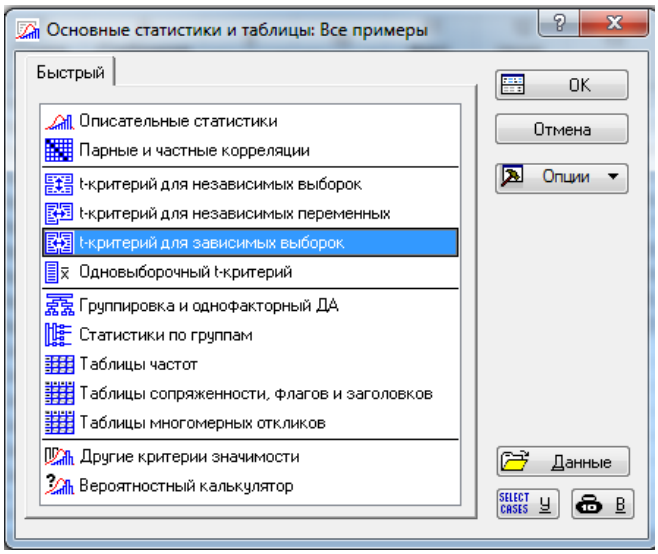


Рис. 4.6. **Стартовая панель основные статистики**

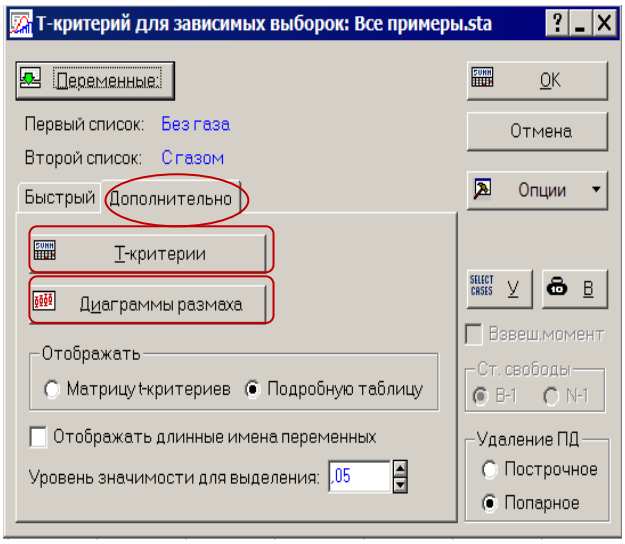


Рис. 4.7. **Диалоговое окно выбора переменных**

В диалоговом окне выбора переменных в первом списке задаем переменную *Без газа*, а во втором – переменную *С газом*, активируем вкладку Дополнительно и нажимаем на кнопку Т-критерии. После нажатия на кнопку **Ок** в рабочей книге получаем итоговую таблицу, в которой наибольший интерес представляют средняя разность $\overline{d} = -10{,}9$, $p=0{,}0161$, $t_ф = 2{,}95$; $t_{05} = 2{,}26$ находим в таблице Стьюдента при числе степеней свободы $cc = 9$ (рис.4.7).

Переменная	Среднее	Стд.от.	N	разн.	Стд.от. разн.	t	cc	p
Без газа	67,70000	8,233401						
С газом	78,60000	8,934328	10	-10,9000	11,66619	-2,95459	9	0,016100

Т-критерий для зависимых выборок (Все примеры.sta)
Отмечены разности, значимые на уров. p < ,05000

Рис. 4.7. **Итоговая таблица результатов сравнения 2-х вариантов хранения груш**

Нажав в диалоговом окне (рис. 4.7) на кнопку **Диаграммы размаха** в рабочей книге получаем диаграмму размаха сохранности плодов груши (рис. 4.8).

Рис. 4.8. **Диаграмма размаха сохранности плодов груши**

На основании данных итоговой таблицы ($t_ф > t_{05}$, $p<0,05$, все значения окрашены красным цветом) и диаграммы размаха (границы доверительных интервалов не пересекаются) с вероятностью 95% можно сделать вывод о том, что сохранность плодов груши в пакетах с газом существенно выше, чем в пакетах без газа, в то время как с вероятностью 99% ($p>0,01$), существенных различий нет.

4.3 Сравнение выборок с использованием непараметрических критериев

В агрономических исследованиях бывают ситуации, когда сравниваемые данные не подчиняются закону нормального распределения, они выражены в условных единицах, баллах, в порядковой шкале. Применять к таким данным

параметрические критерии неправомерно, поэтому для таких случаев используют непараметрические или порядковые критерии. Непараметрические методы – это методы, свободные от предположений о виде функции распределения. Непараметрические критерии часто называют ранговыми или порядковыми критериями, так как в расчетах вместо исходных значений используют ранги (порядок).

Для проверки нулевой гипотезы в агрономических исследованиях можно использовать следующие непараметрические критерии:

- непараметрические критерии для независимых выборок (Ван-дер-Вандена, Манна-Уитни Колмогорова- Смирнова и др.);
- критерии различия для зависимых выборок (Вилкоксона, знаков и др.);
- оценка степени зависимости между переменными (Спирмена, Кендалла).

Рассмотрим, как проверяется нулевая гипотеза с использованием непараметрических критериев для независимых и зависимых выборок.

Непараметрический критерий Манна-Уитни для независимых выборок

Пример 1. В опыте изучали пораженность 2-х сортов овса (Кречет и Скакун) стеблевой ржавчиной, причем пораженность оценивали в баллах: 0 баллов – отсутствие болезни, 4 балла – максимальная пораженность.

Сорта овса	Пораженность стеблевой ржавчиной, в баллах									
Кречет	2	2	1	4	3	3	4	3	2	3
Скакун	1	0	2	1	0	0	3	3	2	1

Определить, существенны ли различия в пораженности стеблевой ржавчиной у 2 сортов?

Ввиду того, что пораженность оценивалась в баллах, данные не подчиняются нормальному распределению, применение параметрического критерия t- Стьюдента не правомерно, оценку существенности будем проводить с использованием непараметрического критерия Манна-Уитни.

Для корректного применения критерия Манна-Уитни необходимо должным образом сформировать файл исходных данных нашего примера. В

частности, файл должен содержать группирующую переменную, имеющую, по крайней мере, два разных кода для однозначной идентификации принадлежности каждого наблюдения к определенной группе. Для этого создаем две переменные: переменную *Сорта овса* с наименованием в строках Кречет и Скакун, и переменную *Стеблевая ржавчина*, в строках которой напротив каждого сорта вводим соответствующие им баллы пораженности ржавчиной, как представлено на рис.4.10.

После создания файла исходных данных щелкнем в строке **Меню Анализ** и выберем модуль **Непараметрическая статистика (Nonparametric Statistics).** Так как мы сравниваем 2 сорта, которые не связаны общим условием, и они не сопряжены, в появившемся окне выбираем опцию **Сравнение двух независимых групп** (рис. 4.9).

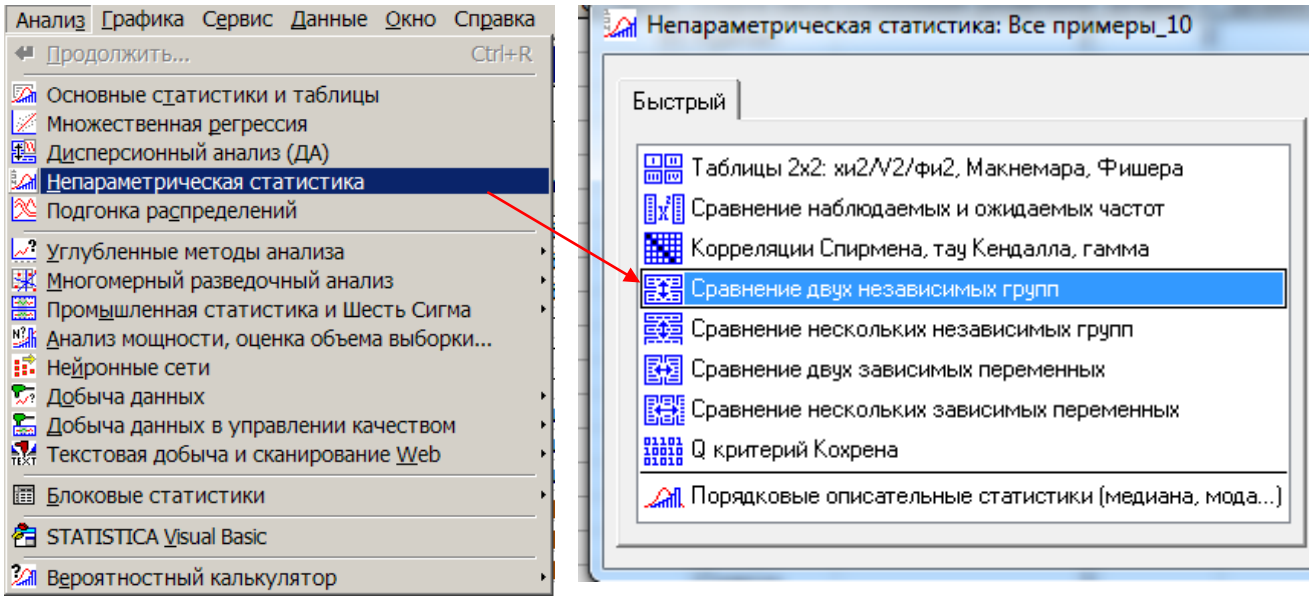


Рис. 4.9. **Стартовая панель модуля *Непараметрическая статистика (Nonparametric Statistics)***

Далее появляется окно для выбора переменных и непараметрического критерия. В качестве зависимой переменной выбираем переменную **Стеблевая ржавчина**, в качестве группирующей переменной – Сорта овса (рис. 4.10). В окошках автоматически указывается код для первой группы «Кречет», для второй группы – «Скакун». Для проверки нулевой гипотезы выберем **U-**

критерий Манна-Уитни (Mann-Whitney U test), который представляет мощную непараметрическую альтернативу t-критерию для независимых выборок.

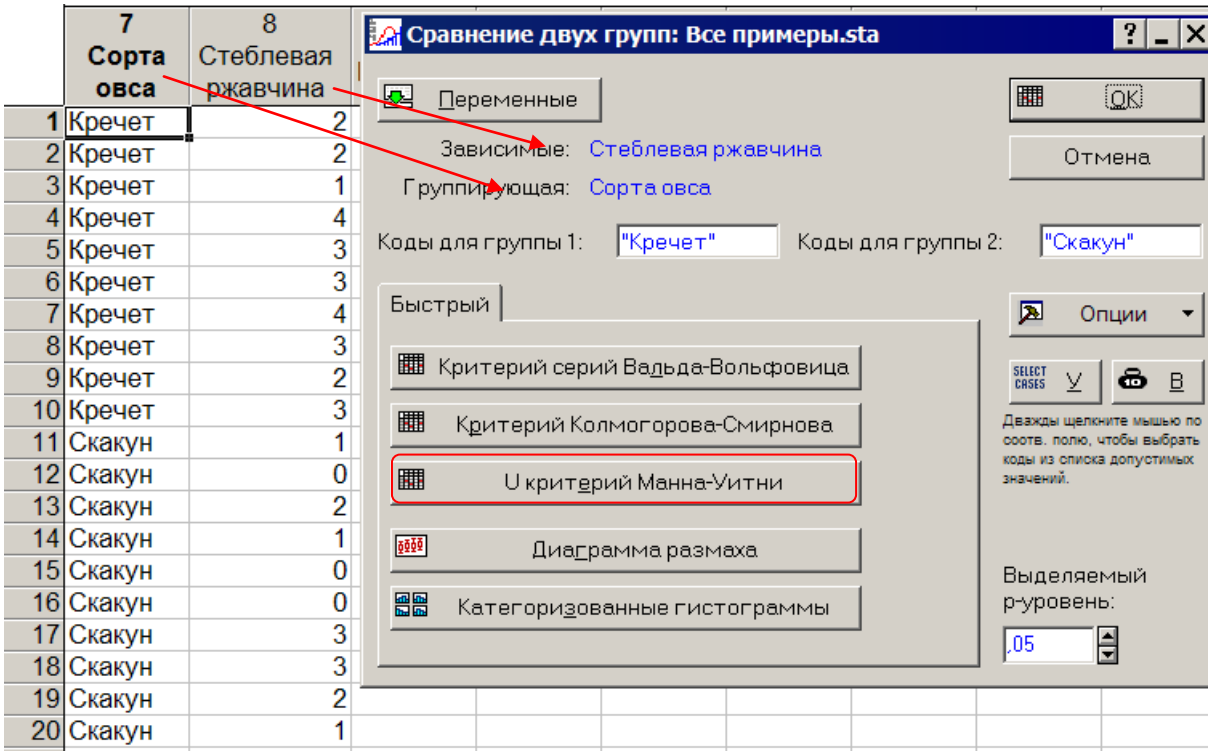


Рис.4.10. **Переменные (исходные данные) и диалоговое окно для выбора переменных и непараметрического критерия**

После нажатия на кнопку **Ок** получаем итоговую таблицу с фактическим значением критерия Манна-Уитни и *p-уровнем* значимости (рис. 4.11). Сумма рангов для сорта Кречет равна *136,5*, сорта Скакун – *73,5*, критерий Z Манна-Уитни – *2,38*. По уровню значимости ***p = 0,017258*** (*p<0,05*) можно сделать вывод, что ***с*** вероятностью 95% пораженность стеблевой ржавчиной сорта Скакун существенно ниже, чем у сорта Кречет, однако с вероятностью 99% (*p>0,01*) различия не существенны.

	Манна-Уитни U критерий (Все примеры.sta) По перем. Сорта овса Отмеченные критерии значимы на уровне p <,05000				
Перем.	**Сум.ранг Кречет**	Сум.ранг Скакун	U	Z	p-уров.
Стеблевая ржавчина	136,5000	73,50000	18,50000	2,381176	0,017258

Рис. 4.11. **Итоговая таблица сравнения по критерии. Манн-Уитни**

Для графического представления результатов нажмем на кнопку **Диаграмма размаха** и получим диаграмму размаха, по которой можно проверить нулевую гипотезу об отсутствии различий между вариантами. Так как данные примера не подчиняются закону нормального распределения, в диалоговом окне **Диаграмма размаха** выберем тип графика **Медиана/кварт/размах** и нажимаем **Ок.** На диаграмме размаха для каждой переменной показана медиана – маленький квадратик, квартальный размах – ящик (границы 25% и 75% процентилей), минимальное и максимальное значение – усы.

Если судить по диаграмме размаха (рис.4.12), то так как верхняя граница ящика для сорта Скакун (75% процентиль) не соприкасается с нижней границей сорта Кречет (25% процентиль), можно сделать дополнительный вывод о существенности различий на 05% уровне значимости по пораженности этих сортов овса стеблевой ржавчиной.

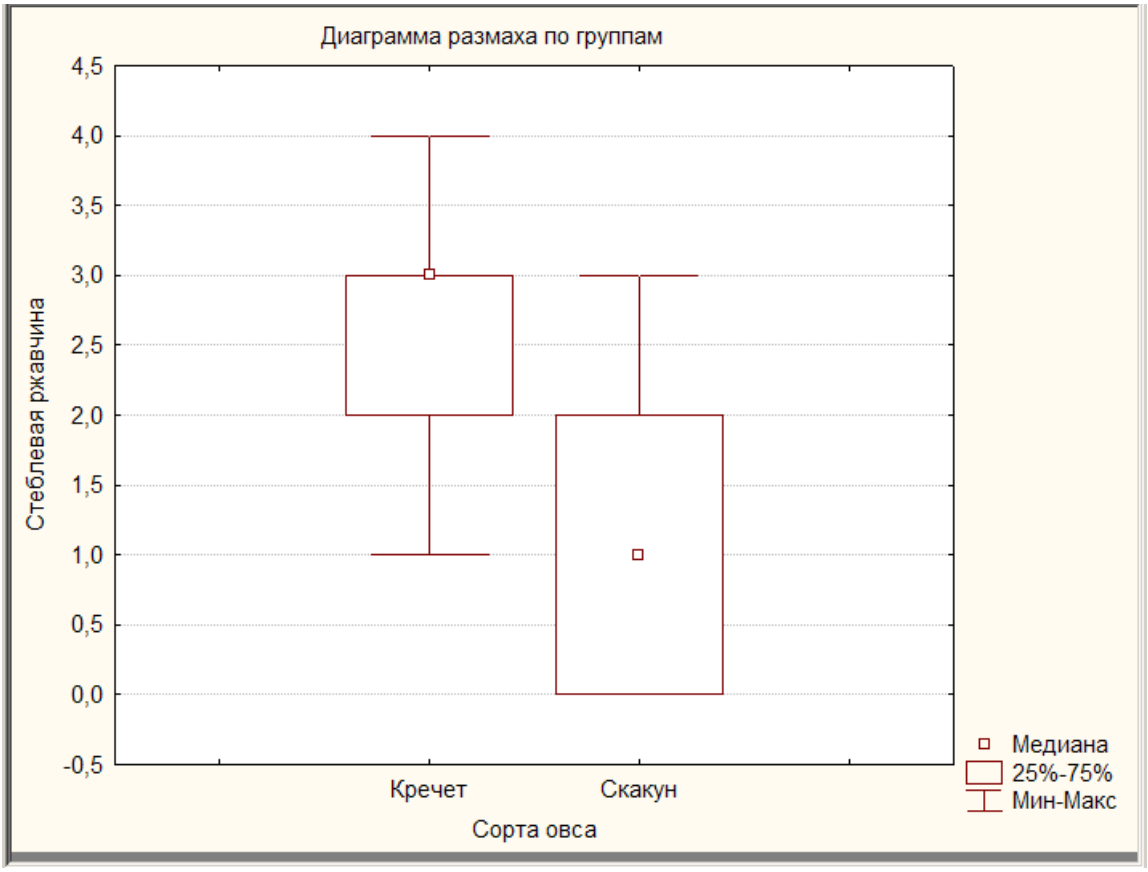


Рис. 4.12. **Диаграмма размаха пораженности стеблевой ржавчиной 2-х сортов овса**

Непараметрические критерии для зависимых выборок

Непараметрические критерии: *критерий Вилкоксона и критерий знаков* являются альтернативой t-критерию для зависимых выборок.

Критерий *знаков* применяется в ситуациях, когда исследователь проводит два измерения (например, при разных условиях) одних и тех же субъектов и желает установить наличие или отсутствие различия результатов. Для применения этого критерия не требуется знания о природе или форме распределения. Подсчитывается количество положительных разностей (∑"+") между значениями переменной (A) и значениями переменной (B). При нулевой гипотезе (отсутствие эффекта обработки) число положительных разностей будет примерно таким же, как и отрицательных.

При использовании критерия *W – Вилкоксона* для каждой пары наблюдений рассчитываются разности (*d*), которые затем ранжируются. Требования к критерию *Вилкоксона* более строгие, чем к критерию *знаков*. Однако если они удовлетворены, то критерий *Вилкоксона* имеет большую мощность, чем критерий *знаков*.

Пример 2. Поражение листьев яблони при инокуляции штаммами Venturia inaequalis (парша) определялось по диаметру пятна, мм:

Штамм	№ листа											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
№1	0	0	5	4	3	4	4	5	6	3	3	4
№2	0	0	7	6	5	3	7	8	8	5	3	6

Сформируем файл исходных данных, содержащий две переменные: *Штамм1* и *Штамм2* с соответствующими значениями диаметра пятна по принципу 2-х списков, как показано на рис. 4.14.

В стартовой панели **Непараметрическая статистика (Nonparametric Statistics)** выберем опцию **Сравнение двух зависимых переменных** (рис. 4.13).

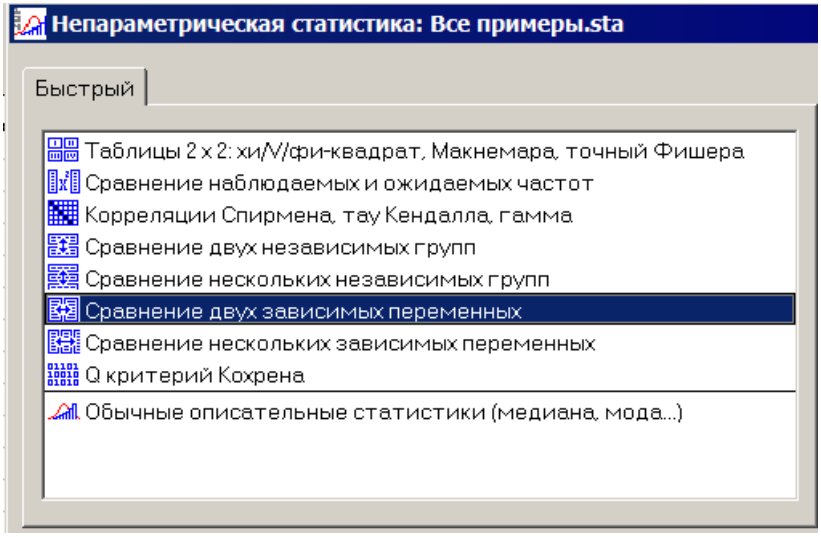


Рис. 4.13. **Стартовая панель модуля *Непараметрическая статистика***

После выбора опции на экране появится диалоговое окно, в котором необходимо указать переменные из двух списков. Каждая переменная первого списка сравнивается с каждой переменной второго списка. В окне ввода переменных (рис. 4.14) укажем сравниваемые штаммы 1 и 2, нажмем на **Критерий знаков (Sign test),** затем на клавишу **Ok**, далее аналогично выберем **Критерий Вилкоксона (Wilcoxon test).**

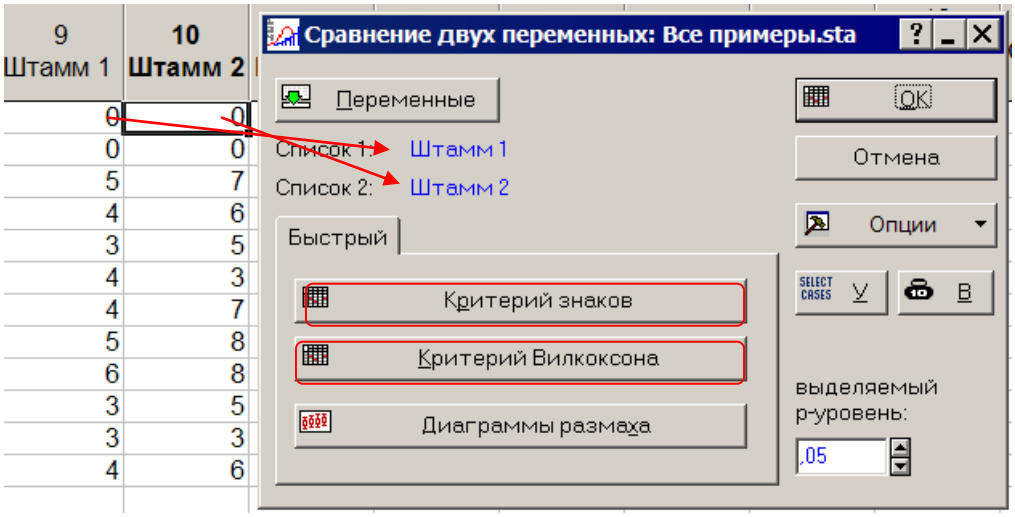


Рис. 4.14. **Переменные (исходные данные) и диалоговое окно выбора критериев для анализа**

После нажатия на кнопку **Ок** в результате расчетов появятся таблицы с фактическим значением критерия знаков (*Z=2,00; p=0,0455*) и фактическим значением критерия Вилкоксона (*Z=2,547; p=0,010*).

Пара перем.	Критерий знаков (Все примеры.sta) Отмеченные критерии значимы на уровне p <,05 Число несовп.	Процент v < V	Z	p-уров.
Штамм 1 & Штамм 2	9	88,88889	2,000000	0,045500

Пара перем.	Критерий Вилкоксона (Все примеры.sta) Отмеченные критерии значимы на уровне p <,05 Число набл.	T	Z	p-уров.
Штамм 1 & Штамм 2	12	1,000000	2,547100	0,010863

Рис. 4.15. **Фактические значения критериев знаков и Вилкоксона**

Так как в обоих случаях все показатели выделены красным цветом ($p<0,05$), с вероятностью 95% можно сделать вывод о значимых (существенных) различиях в вредоносности изучаемых штаммов. И в то же время с вероятностью 99% эти различия не существенны, так как $p<0,01$.

4.4 Критерий сопряженности – критерий Пирсона – Хи-квадрат

Критерий Пирсона χ^2 (критерий Хи-квадрат) рассчитывается по формуле: $\chi^2_ф = \sum \frac{(f-F)^2}{F}$ или $\chi^2_ф = \sum \frac{(H-O)^2}{O}$, где f, H – наблюдаемые или фактические частоты, F,O – ожидаемые (теоретические) частоты). Он применяется для проверки нулевой гипотезы при анализе подсчета численности качественных признаков. В биологических исследованиях – это определение сопряженности или независимости изучаемых вариантов, соответствие между фактическими (наблюдаемыми) и ожидаемыми (теоретическими) распределениями – анализ расщепления в генетических исследованиях и др.

Нулевая гипотеза: предположение о несущественности различий между изучаемыми вариантами принимается, если $\chi^2_ф < \chi^2_{табл}$ при заданном уровне значимости ($p>0,05$) и наоборот принимается альтернативная гипотеза о существенности различий, если $\chi^2_ф \geq \chi^2_{табл}$ ($p<0,05$)

Условия для применения Хи-квадрат:

- анализ только качественной изменчивости
- общий объем выборки не менее 50 наблюдений
- при численности классов крайних групп менее 5 наблюдений, их объединяют.

Хи-квадрат для проверки нулевой гипотезы о соответствии между эмпирическими и теоретическими распределениями

Пример 1. При дигибридном скрещивании сортов гороха во втором поколении произошло расщепление 381 горошин на: гладкие и жёлтые (f_1 = 221), гладкие и зелёные (f_2 = 67), морщинистые и жёлтые (f_3 = 74), морщинистые и зелёные (f_4 = 19). Соответствует ли наблюдаемое расщепление закону Менделя (соотношение 9:3:3:1)?

Исходя из закона Менделя рассчитаем вероятности (P по H_o) отдельных фенотипов, которые составят: гладко-жёлтые $P_1 = {}^9/_{16}$, гладко-зелёные $P_2 = {}^3/_{16}$, морщинисто-жёлтые $P_3 = {}^3/_{16}$ и морщинисто-зелёные $P_4 = {}^1/_{16}$. Перемножив общую сумму горошин на рассчитанные вероятности, получим значения теоретических частот: F_1= 381*${}^9/_{16}$=214,4; F_2= 381*${}^3/_{16}$=71,4; F_3= 381*${}^3/_{16}$=71,4; F_4 381*${}^1/_{16}$=23,8.

В таблице исходных данных программы Statistica создадим две переменные «факт. частоты» и «теорт. частоты» и внесем в них значения частот (рис. 4.16).

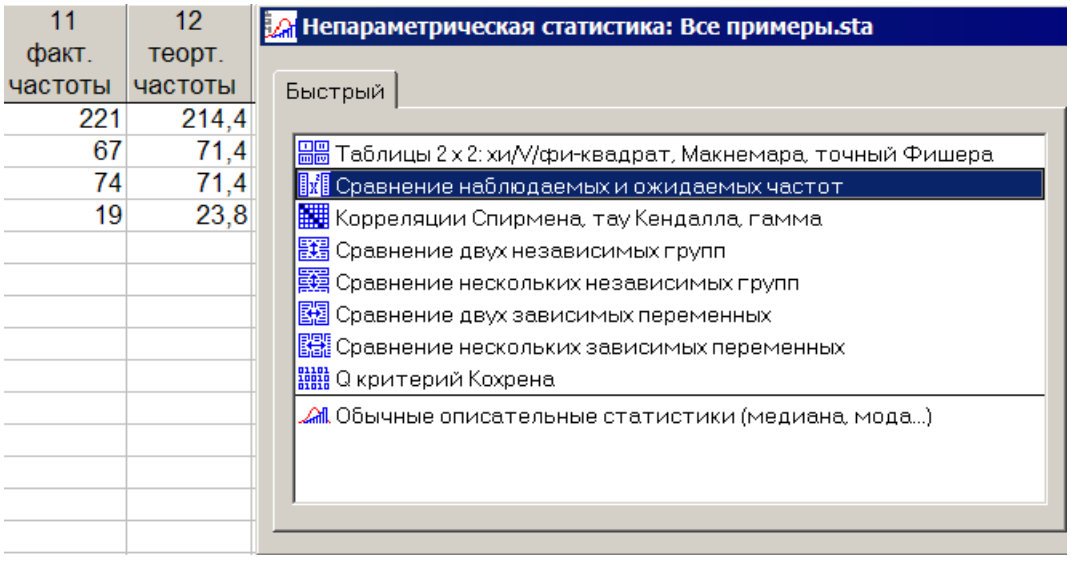
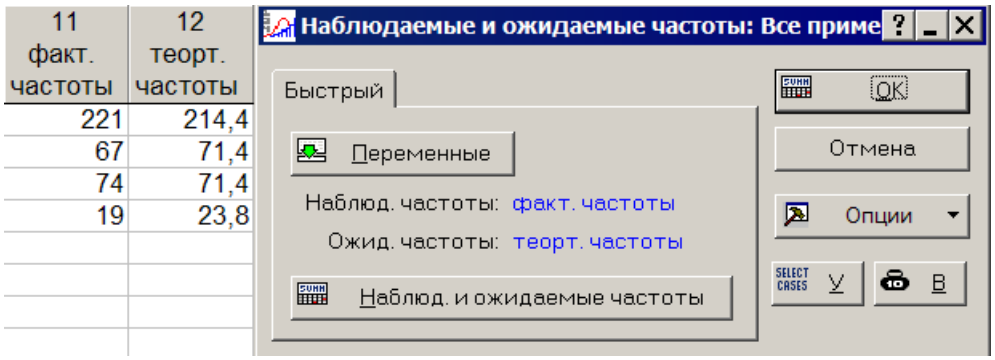


Рис. 4.16. **Стартовая панель модуля *Непараметрическая статистика***

В стартовой панели **Непараметрическая статистика (Nonparametric Statistics)** выберем опцию **Сравнение наблюдаемых и ожидаемых частот (Observed versus expected XI),** затем в диалоговом окне укажем анализируемые переменные и нажмем на клавишу **Ok.**



В итоговой таблице (рис. 4.17) показан алгоритм расчет фактического значения критерия Хи-квадрат $\chi^2_\phi = 1{,}537$ $p < 0{,}67$.

Наблюд.	Наблюдаемые и ожидаемые частоты (Все приме Хи-квадрат = 1,537065 cc = 3 p < ,673743 **Наблюд. факт. частоты**	Ожидаем. теорт. частоты	H - O	(H-O)**2 /O
C: 1	221,0000	214,4000	6,60000	0,203172
C: 2	67,0000	71,4000	-4,40000	0,271148
C: 3	74,0000	71,4000	2,60000	0,094678
C: 4	19,0000	23,8000	-4,80000	0,968067
Сумма	381,0000	381,0000	-0,00000	1,537065

Рис. 4.17. **Итоговая таблица расчета критерия Хи-квадрат**

На основании проведенного теста можно сделать следующий вывод: так как фактическое значение критерия $\chi^2_\phi = 1{,}537$ меньше табличного $\chi^2_{05} = 7{,}81$ *и* $p > 0{,}05$, нулевая гипотеза о соответствии наблюдаемого расщепления ожидаемому (теоретическому) по закону Менделя с вероятностью 95% принимается.

Хи-квадрат для проверки нулевой гипотезы о независимости

Пример 2. Изучали нормы расхода инсектицида Спинтор на гибель колорадского жука. При норме расхода Спинтора 0,1 л/га из 150 жуков погибло 64, а при норме – 0,2 л/га из 150 жуков погибло 102 жука. Необходимо определить существенно ли различие в гибели колорадского жука от нормы расхода Спинтора?

Выдвигаем нулевую гипотезу на 5% уровне значимости: гибель колорадского жука не зависит от норм расхода инсектицида Спинтор.

Для сравнения 2-х групп, относящихся к альтернативной качественной изменчивости составим так называемую таблицу исходных данных 2х2 с двумя строками, в которые записываем нормы расхода Спинтора и с двумя столбцами, в которых указываем количество погибших и живых жуков.

Нормы расхода Спинтора, л/га	Количество жуков		Сумма
	погибшие	живые	
0,1	64	86	150
0,2	102	48	150

Для расчета критерия Хи-квадрат в программе Statistica в стартовой панели **Непараметрическая статистика** выберем опцию **Таблицы 2х2 Хи/V** и в появившемся диалоговом окне в окошечки вводим частоты нашего примера (рис.4.18)

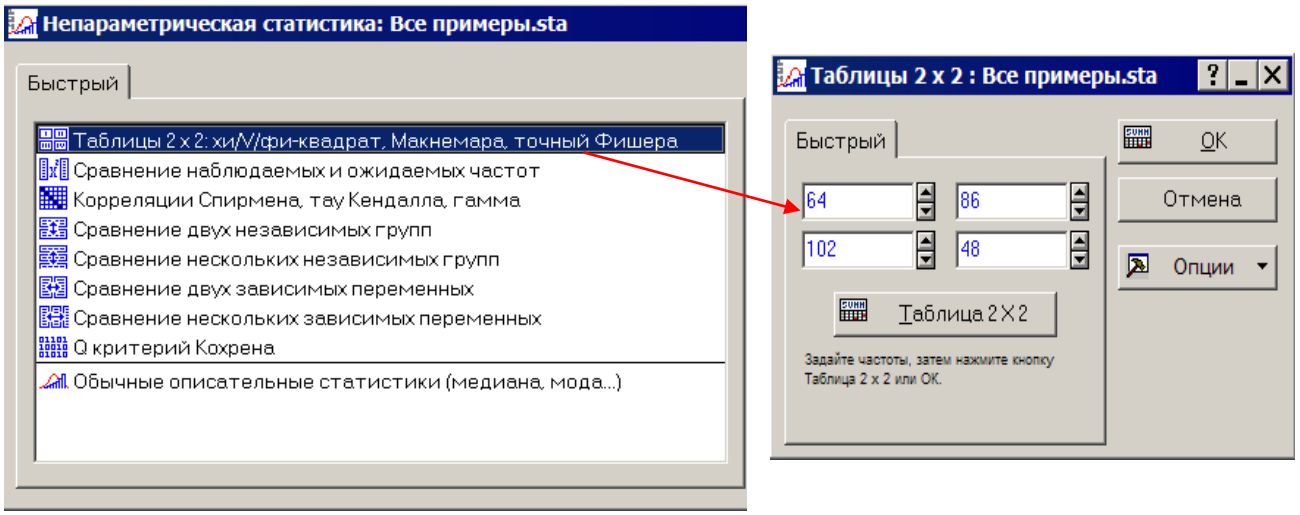


Рис. 4.18. **Стартовая панель модуля *Непараметрическая статистика и заполнение таблицы 2х2***

После выбора в стартовой панели процедуры **Таблицы 2х2 хи/V/фи-квадрат** открывается диалоговое окно, в котором вводим частоты в таблицу 2х2, состоящую из двух строк и двух столбцов и можем вычислить различные статистики, позволяющие оценить зависимость между двумя альтернативными качественными переменными. После нажатия на клавишу **Ок** получаем

итоговую таблицу, в которой приведены результаты расчета фактического значения критерия *Хи-квадрат* (рис. 4.19).

	Таблица 2x2 (Все примеры.sta)		
	Столб. 1	Столб. 2	Сумма строкам
Частоты, строка 1	64	86	150
Процент от общего	21,333%	28,667%	50,000%
Частоты, строка 2	102	48	150
Процент от общего	34,000%	16,000%	50,000%
Сумма по столбцам	166	134	300
Процент от общего	55,333%	44,667%	
Хи-квадрат (ст.св.=1)	19,47	p= ,0000	
V-квадрат (ст.св.=1)	19,41	p= ,0000	
Поправка Йетса	18,46	p= ,0000	
Фи коэффициент	,06492		
Фишера p, односторонний		p= ,0000	
двусторонний		p= ,0000	
Макнемара Хи-квадрат (A/D)	2,01	p= ,1564	
Хи-квадрат (B/C)	1,20	p= ,2740	

Рис. 4.19. **Итоговая таблица с результатами**

Так как $\chi^2_\phi = 19{,}47$ больше табличного значения ($\chi^2_{05} = 3{,}84$; $\chi^2_{01} = 6{,}63$) $p < 0{,}01$, нулевая гипотеза о независимости гибели жуков от норм расхода отвергается, и с вероятностью не только 95%, но с 99% мы можем считать, что увеличение нормы расхода инсектицида Спинтор с 0,1 до 0,2 л/га приводит к существенному увеличению гибели жуков колорадского жука.

В итоговой таблице помимо стандартного критерия Хи-квадрат Пирсона и скорректированного Хи-квадрат (V-квадрат) рассчитаны следующие критерии: Фи-квадрат коэффциент, критерий Хи-квадрат Макнемара и точный Фишера.

Контрольные вопросы:

1. Нулевая гипотеза в агрономических исследованиях. Какие критерии служат для проверки нулевой гипотезы?
2. Уровень значимости и доверительная вероятность в агрономических исследованиях.
3. Способы проверки нулевой гипотезы агрономических исследований.
4. Параметрический критерий t-Стьдента для проверки нулевой гипотезы.

5. Как оценить существенность разности средних 2-х независимых вариантов?
6. Как оценить существенность средней разности 2-х зависимых вариантов?
7. Сравнение средних с помощью диаграммы размаха?
8. Как сравнить варианты, если данные агрономических исследований не подчиняются закону нормального распределения?
9. Непараметрические критерии для проверки нулевой гипотезы.
10. Критерий хи-квадрат (Пирсона) для проверки соответствия фактических рядов распределения нормальному.
11. Каковы условия применения критерия хи-квадрат?

Глава 5. ДИСПЕРСИОННЫЙ АНАЛИЗ ДАННЫХ АГРОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ

Дисперсионный анализ (английская аббревиатура ANOVA – анализ варианс) является одним из распространенных методов статистической обработки результатов агрономических исследований. Рекомендуется использовать для сравнения 3-х и более вариантов опыта.

Сущностью дисперсионного анализа является одновременное разложение суммы квадратов и числа степеней свободы на составляющие компоненты, которые соответствуют структуре эксперимента и оценка действия и взаимодействия изучаемых вариантов по *F- критерию*. С помощью дисперсионного анализа можно разложить общую вариацию (изменчивость) в опыте (эксперименте) на составляющие компоненты в зависимости от условий проведения экспериментов на вариацию, определяемую действием изучаемого фактора и вариацию, вызываемую случайными (неконтролируемыми) в данном опыте условиями.

Общее варьирование изучаемого признака будет состоять из варьирования между выборками (вариантами) и варьирования внутри выборок. Вариация между выборками (вариантами) представляет ту часть общей изменчивости, которая обусловлена действием изучаемых факторов, а вариация

внутри выборок характеризует действие случайных факторов, т.е. ошибку эксперимента.

Оценка значимости действия изучаемых факторов проводится по *критерию Фишера – F,* который представляет собой отношение дисперсии (среднего квадрата) вариантов к дисперсии ошибки $F_ф = \frac{S_v^2}{S_e^2}$

Если $F_{факт} < F_{табл.,}$ то нулевая гипотеза (предположение: все средние значения по вариантам являются оценками одной генеральной средней и между ними нет существенных различий) принимается и на этом расчеты заканчиваются, если $F_{факт} \geq F_{табл.}$, то нулевая гипотеза отвергается и необходимо дополнительно оценить существенность частных различий – определить между какими средними имеются значимые различия.

В зависимости от условий проведения опытов применяют различные схемы (модели) дисперсионного анализа, в которых указывается на какие конкретно суммы квадратов и степени свободы расчленяют общее варьирование.

К сожалению, в программе Statistica полностью, не прибегая к дополнительным расчетам, можно провести дисперсионный анализ только данных однофакторных и многофакторных опытов, заложенных по схеме полной рандомизации (неорганизованных повторений). Процедура не позволяет вычленить дисперсию блоков (повторений), рядов, столбцов, поэтому она не предназначена для обработки данных опытов, проведенных методом рандомизированных повторений (блоков), латинского квадрата, расщепленных делянок и блоков.

5.1 Дисперсионный анализ данных однофакторных экспериментов с полной рандомизацией вариантов (лабораторный, вегетационный, полевой опыты с полной рандомизацией вариантов)

Пример 1. В вегетационном опыте изучали урожайность 5 сортов озимой пшеницы, г/сосуд. Опыт проведен в 4-х кратной повторности. Следует определить, существенны ли различия между вариантами (сортами) опыта?

Сорта	Повторность			
	1	*2*	*3*	*4*
1. Безостая 1,st	35,4	37,1	36,6	35,8
2. Купава	39,6	40,4	38,3	39,5
3. Августа	35,6	34,8	35,8	36,5
4. Гарант	41,6	39,9	39,3	45,6
5. Немчиновская-57	46,2	45,4	43,8	49,9

Для проведения дисперсионного анализа создадим в программе Statistica новый файл с данными из двух переменных (рис. 5.1), первой переменной дадим наименование изучаемого фактора – **Сорта** и в строки (наблюдения) этой переменной впишем наименование изучаемых вариантов (сортов), причем исходя из заданной повторности, каждый из вариантов повторяется 4 раза. Вместо наименования вариантов их можно закодировать. Во второй столбец занесем урожайные данные каждого сорта и назовем вторую переменную – **Урожай.** В опыте изучается 5 вариантов в 4-х кратной повторности, общее число наблюдений составляет 20. После создания файла исходных данных сохраним его. На рис.5.1 представлен вид окна с файлом данных. В нашем случае файл сохранен под именем: ***Дисперсионный анализ.sta***

	Сорта	Урожай
1	Безостая 1	35,4
2	Безостая 1	37,1
3	Безостая 1	36,6

4	Безостая 1	35,8
5	Купава	39,6
6	Купава	40,4
7	Купава	38,3
8	Купава	39,5
9	Августа	35,6
10	Августа	34,8
11	Августа	35,6
12	Августа	36,5
13	Гарант	41,6
14	Гарант	39,9
15	Гарант	39,3
16	Гарант	45,6
17	Немчиновская-57	46,2
18	Немчиновская-57	45,4
19	Немчиновская-57	43,8
20	Немчиновская-57	49,9

Рис. 5.1. **Вид окна файла исходных данных**

В меню **Анализ (Statistics)** выберем модуль **Дисперсионный анализ (Anova)**, далее – вид анализа **Однофакторный ДА (One-way ANOVA)**.

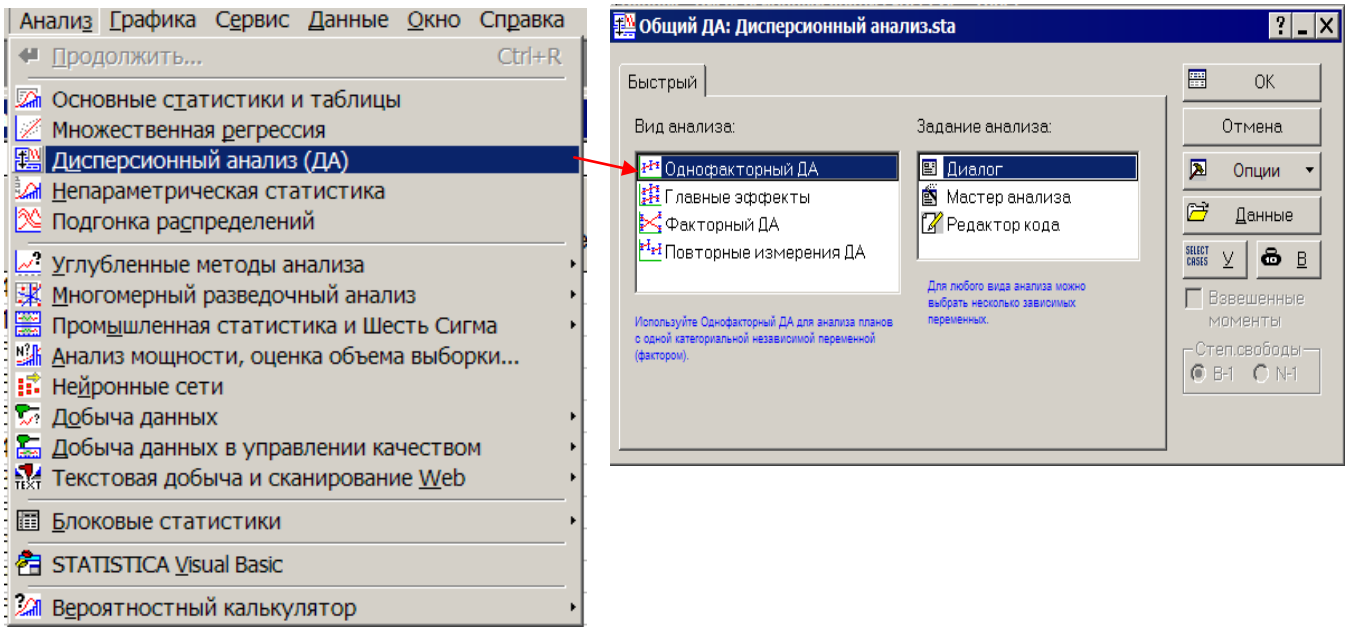


Рис.5.2. **Стартовая панель модуля Дисперсионный анализ**

Во вновь открывшемся окне (рис. 5.3) в качестве зависимой переменной укажем переменную «**Урожай**», а в качестве независимой переменной выступает категориальный предиктор, в нашем случае – «**Сорта**».

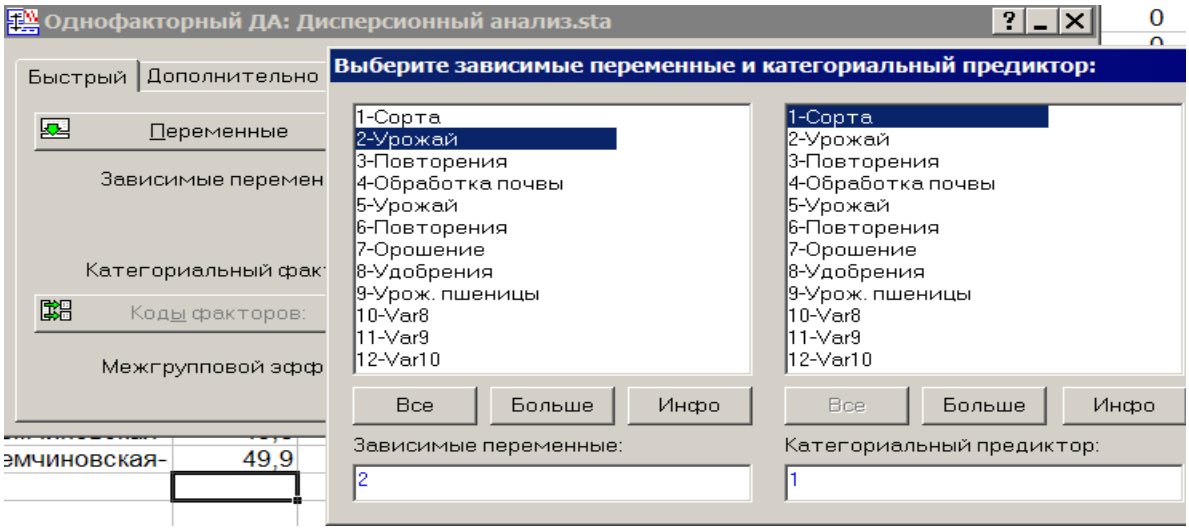


Рис. 5.3. **Диалоговое окно выбора переменных**

В этом же окне нажмем на вкладку **Дополнительно (Advanced),** выберем ортогональную модель дисперсионного анализа – **Тип III** и оставим по умолчанию указанные ниже параметры.

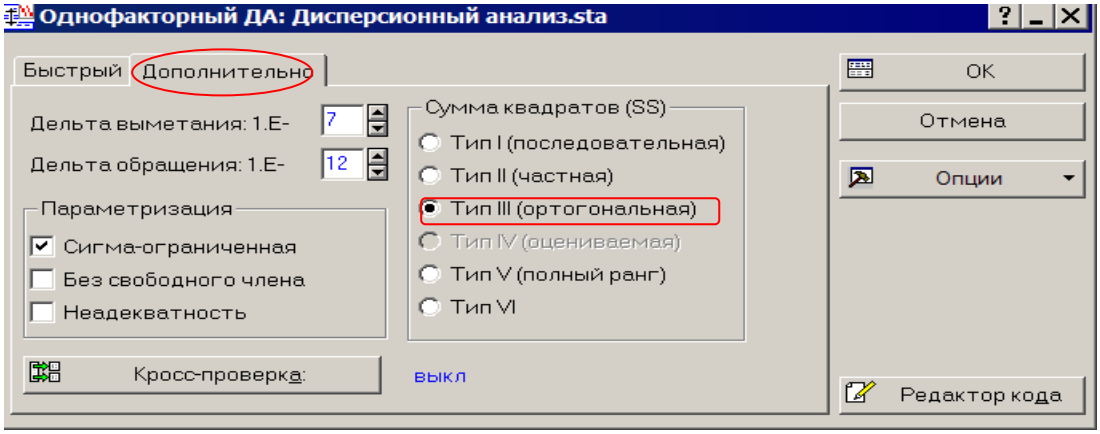


Рис. 5.4. **Диалоговое окно выбора разных моделей дисперсионного анализа**

После нажатия на клавишу **Ок** откроется окно выбора разных форм результатов дисперсионного анализа однофакторного опыта (рис. 5.4). Автоматически программа откроет его на вкладке **Быстрый (Quick)** и результаты анализа можно получить на этом этапе. Однако для развернутого представления результатов дисперсионного анализа (проверка на предпосылки Anova, множественные сравнения и др.) в нижней части окна изменяем или

соглашаемся с указанными по умолчанию (*0,05*) значениями уровня значимости доверительного интервала и нажмем на кнопку **Больше (More Results).** После нажатия на эту кнопку появляется второе диалоговое окно с расширенным набором вкладок для выбора всевозможных форм результатов дисперсионного анализа (рис. 5.5).

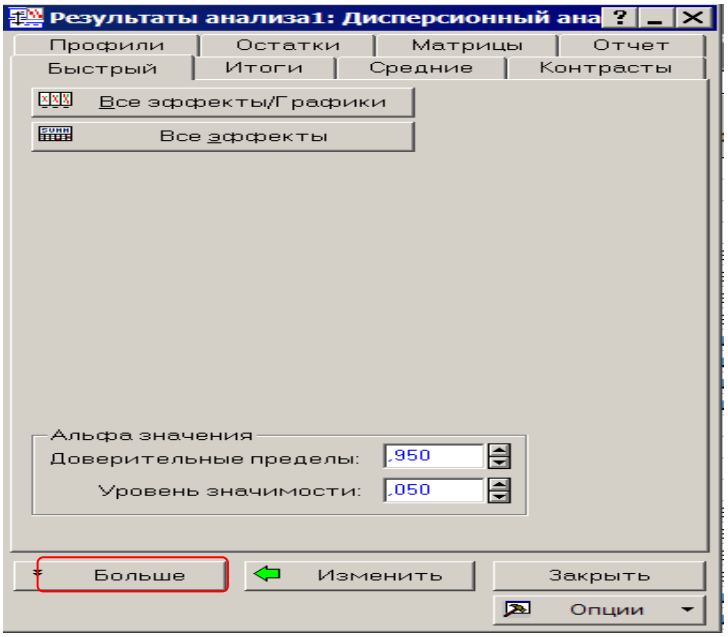


Рис. 5.5. **Диалоговое окно выбора результатов дисперсионного анализа**

5.1.1 Проверка гипотезы на однородность дисперсий по критериям Бартлетта и Левина (предпосылки дисперсионного анализа)

Для проверки условий проведения анализа предпосылкам Anova в открывшемся окне дополнительных результатов дисперсионного анализа активируем вкладку **Предположения (Assumptions)** и нажмем на **клавишу Кохрена C, Хартли, Бартлетта (Cochran C, Hartly, Bartlett),** (рис. 5.6.)

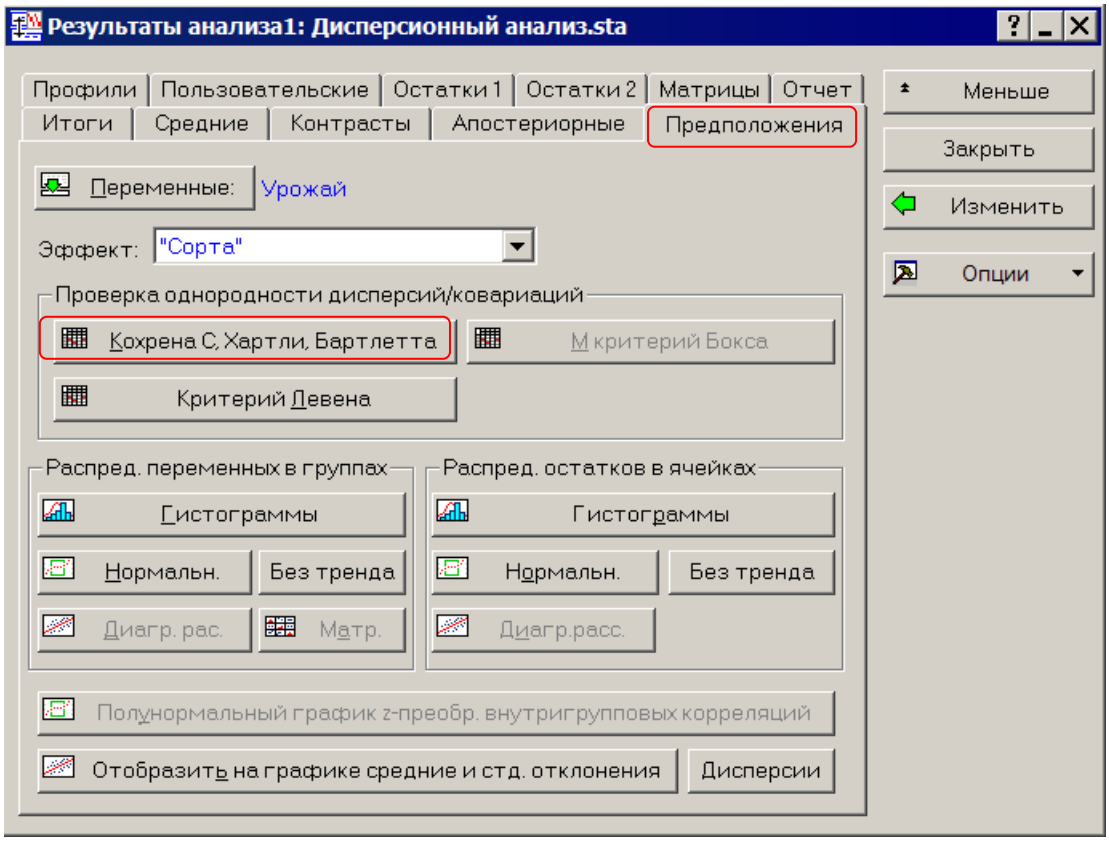


Рис. 5.6. **Диалоговое окно выбора теста для проверки однородности дисперсий**

После выбора критериев для проверки однородности дисперсий появляется таблица с фактическими значениями выбранных критериев и значение уровня значимости (*p*).

	Критерии однородности дисперсий (Дисперсионный анал Эффект: "Сорта"				
	Хартли F-макс	Кохрена С	Бартлетт Хи-квад.	ст.св.	p
Урожай	16,70466	0,486788	9,389386	4	0,052070

Так как уровень вероятность по всем критериям $p>0,05$, нулевая гипотеза об однородности дисперсий в представленном опыте принимается, что дает нам основание для продолжения дисперсионного анализа и применения параметрических критериев.

После проверки гипотезы на предпосылки дисперсионному анализу возвращаемся к окну с дополнительными результатами дисперсионного анализа (рис. 5.6), откроем вкладку **Итоги (Summary)** и нажмем на клавишу **Одномерные результаты.**

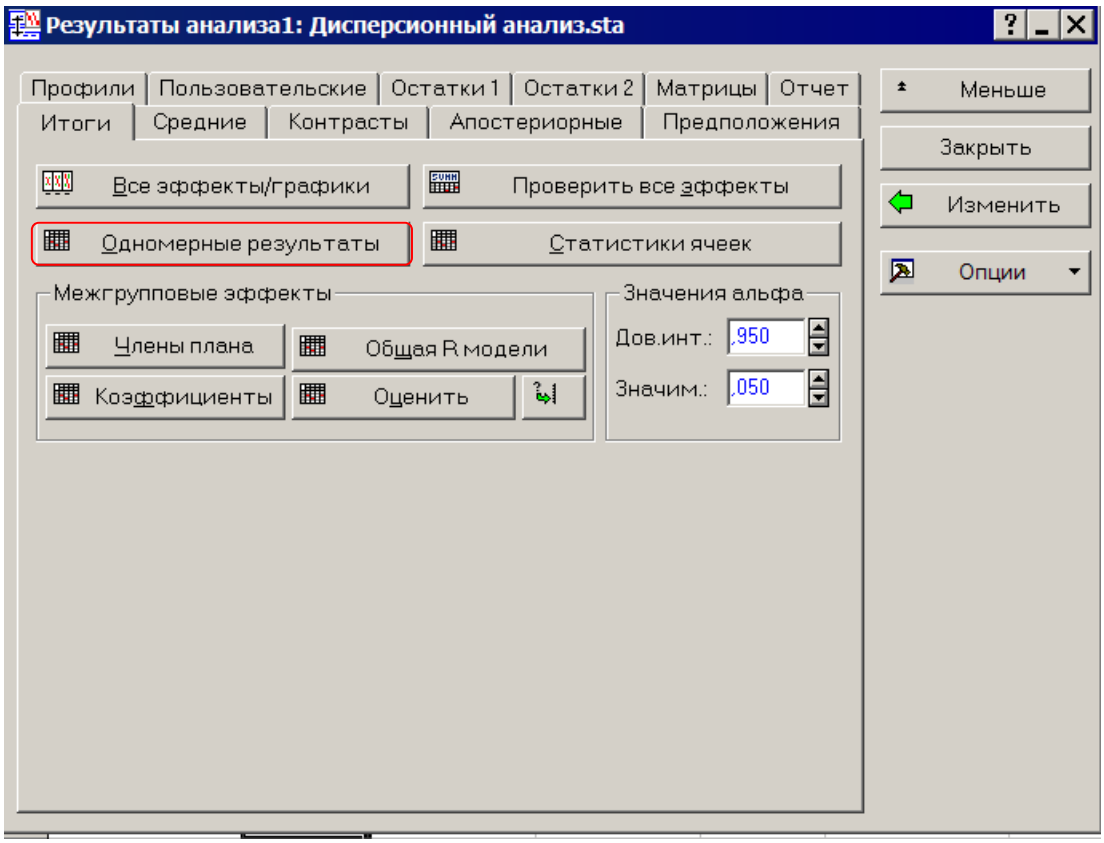


Рис.5.6. **Диалоговое окно выбора результатов дисперсионного анализа**

При нажатии на кнопку **Одномерные результаты** получаем основную таблицу дисперсионного анализа с результатами сумм квадратов (SS), числа степеней свободы и дисперсий (MS) по источникам вариации в опыте, в нашем случае: изучаемый фактор – Сорта и случайный – Ошибка (рис. 5.7). Значения по строке Св.член в нашем анализе в расчет не принимаются. Самым важным показателем в представленной таблице являются критерий Фишера – ***F*** и вероятность значимости ***p,*** по которым проверяется нулевая гипотеза**.** В специальной таблице находим табличное значение критерия Фишера (F_{05} = ***3,06;*** F_{01} = ***4,89***) при числе степеней свободы для вариантов (*с.с.v = 4*) и ошибки (*с.с.е = 15*) и сравниваем с F_{ϕ} . Так как F_{ϕ} =***22,99***> F_{01} = ***4,89*** и ***p = 0,000003***< ***0,01***, нулевая гипотеза о равенстве средних значений изучаемых вариантов отвергается и с вероятностью 99% можно считать, что в опыте есть существенные различия по урожайности между сортами. Мы отвергаем нулевую гипотезу об отсутствии различий в целом между средними на уровне ***0,000003.*** Иными словами, отвергая гипотезу о равенстве средних в опыте, мы рискуем ошибиться практически с нулевой вероятностью.

Эффект	Одномерные результаты для каждой ЗП (Дисперсионный анализ.sta) Сигма-ограниченная параметризация Декомпозиция III типа **Степени свободы**	Урожай SS	Урожай MS	Урожай F	Урожай p
Св. член	1	31752,48	31752,48	9588,549	0,000000
Сорта	4	304,56	76,14	22,992	0,000003
Ошибка	15	49,67	3,31		
Всего	19	354,23			

Рис. 5.7. **Таблица дисперсионного анализа**

Если $F_{факт} < F_{табл.,}$ то нулевая гипотеза $H_o : \bar{x}_2 - \bar{x}_1 = 0$(предположение: все средние значения по вариантам являются оценками одной генеральной средней и между нет существенных различий) принимается и на этом этапе все расчеты заканчиваются. Но так как по результатам наших расчетов $F_{факт} \geq F_{табл.,}$ то нулевая гипотеза отвергается и необходимо дополнительно оценить существенность частных различий по величине HCP_{05} *или* HCP_{01} и определить между какими средними имеются значимые (существенные) различия.

5.1.2 Множественные сравнения разности средних между вариантами

F- критерий используется как общий критерий, подтверждающий или опровергающий значимое влияние изучаемого фактора на результаты опыта в целом. Несмотря на то, что в нашем опыте в целом доказано существенное влияние сортов на урожайность озимой пшеницы, это автоматически не означает, что между всеми вариантами (сортами) есть значимые различия. Поэтому следующим важным этапом дисперсионного анализа является установление существенности частных различий, т.е. попарное сравнение средних значений урожайности озимой пшеницы по вариантам опыта. Для этого в программе Statistica используется процедура множественных сравнений. Сравнение групповых средних можно провести при помощи 7-ми различных тестов:

– оценки существенности разности **– Фишера НЗР(Fisher LSD), Бонферонни; Шеффе(Scheff), Тьюки (Tukey);**

– оценки размахов – **Ньюмана-Кеулса (Дункана (Duncan)'**

– сравнения опытных вариантов с контрольным или стандартным вариантом – **Дуннета**

Наиболее распространенным является тест **Фишера НЗР** – наименьшая значимая разность **(Fisher LSD** – less significance distance). Этот тест сравнения в отечественной литературе известен как ***НСР*** – наименьшая существенная разность, он рассчитывается по формуле, и строго говоря, в нем используется не критерий Фишера, а критерий *t - Стьюдента*

$$S_d = \sqrt{\frac{2S_e^2}{n}} \qquad HCP_{05} = t_{05} \cdot S_d \qquad HCP_{01} = t_{01} \cdot S_d$$

С помощью *НСР* и других тестов оценивается существенность разности средних (***d***). Если $|d| \geq HCP_{05}$, то между этими средними наблюдаются существенные различия на 05% уровне значимости. Оценка средних по величине *НСР*, рассчитанной с помощью критерия *t- Стьюдента,* при числе вариантов больше 3 – 4-х приводит к некоторому завышению существенных различий по сравнению с другими тестами, о чем свидетельствуют значения HCP_{05}, при расчете которой использовались эти критерии. Если при сравнении двух соседних средних (*r=2*), где *r* – ранг средних значений с использованием любых вышеприведенных критериев получаем одинаковые значения HCP_{05}, то при сравнении средних, удаленных на 3-4 и более мест с использованием критериев *Тьюки, Дуннета и Дункана* значение *НСР* несколько возрастает. То есть, эти критерии являются более строгими по сравнению с критерием *t-Стьюдента* для доказательства нулевой гипотезы.

Ниже приводятся расчеты НСР, основанные на критериях *Стьюдента, Тьюки, Дуннета и Дункана*.

Степени свободы для остатка (c.c.v = 4; c.c.e. = 15)

$$S_{\bar{x}} = \sqrt{\frac{S_e^2}{n}} = \sqrt{\frac{3{,}31}{4}} = 0{,}91 \qquad S_d = \sqrt{\frac{2S_e^2}{n}} = \sqrt{\frac{2 \cdot 3{,}31}{4}} = 1{,}29$$

Сравнение любых средних между собой:

- *Тест Фишера (критерий Стьюдента) – $HCP_{05} = t_{05} \cdot Sd = 2{,}13 \cdot 1{,}29 = 2{,}7$*
- *Критерий Тьюки – $HCP_{05} = q_{05} \cdot S_{\bar{x}} = 3{,}01 \cdot 0{,}91 = 2{,}7$ (r=2)*

Критерий Тьюки – $НСР_{05} = q_{05} \cdot S = 3,67 \cdot 0,91 = 3,3$ (r=3)

<u>Сравнение опытных вариантов со стандартом(контролем):</u>

- *Критерий Дуннета – $НСР_{05} = D_{05} \cdot Sd = 2,44 \cdot 1,29 = 3,1$ (r=2)*

На заключительном этапе дисперсионного анализа традиционно составляется итоговая таблица средних значений по вариантам опыта с результатами разностей (отклонений) между этими средними и по величине *НСР* проводится оценка существенности различий.

Ниже приводится полная матрица разности средних всех вариантов между собой и оценка этих разностей с использованием *НСР* – критерия Стьюдента. В шапке таблицы цифрами 1,2,3,4,5 обозначены номера вариантов (сортов) и средняя урожайность этих сортов, в ячейках таблицы на пересечении строки и столбца показана разность средних.

Итоги сравнения любых средних между собой

Сорта	Средние по сортам				
	{1} 36,2	{2} 39,4	{3} 35,6	{4} 41,6	{5} 46,3
	Разности средних, *d*				
1. Безостая	–	3,2*	-0,6	5,4**	10,1**
2. Купава	3,2*	–	-3,8*	2,4	6,9**
3. Августа	-0,6	-3,8**	–	6,0**	10,7**
4. Гарант	5,4**	2,4	6,0**	–	4,7**
5. Немчиновская-57	10,1**	6,9**	10,7**	4,7**	–

$НСР_{05} = t_{05} \cdot Sd = 2,13 \cdot 1,29 = 2,7$

$НСР_{01} = t_{01} \cdot Sd = 2,95 \cdot 1,29 = 3,8$

Звездочками обозначены существенные различия:

* – на 5% уровне значимости

** – на 1% уровне значимости

<u>Сравнение средних в ранжированном ряду</u>

Среди многоранговых критериев критерий Дункана является одним из распространенных. *НСР* на основе критерия Дункана по своему значению

совпадает с общепринятой *НСР*, рассчитанной с использованием критерия Стьюдента при сравнении соседних средних, а при увеличении удаленности средних друг от друга в ранжированном ряду его значение заметно возрастает.

1. Для расчета $НСР_{05}$ на основе критерия Дункана в формулу $НСР_{05} = t_{05} \cdot Sd = 2,13 \cdot 1,29 = 2,7$ вводим множитель *С*, который зависит от ранга (удаленности средних, *r*) и числа степеней свободы, в нашем примере с.с.=15.

-Критерий Дункана – $НСР_{05} = С\ (t_{05} \cdot Sd) = 1,0 \cdot 2,13 \cdot 1,29 = 2,7\ (r=2)$

-Критерий Дункана – $НСР_{05} = С\ (t_{05} \cdot Sd) = 1,05 \cdot 2,13 \cdot 1,29 = 2,9\ (r=3)$

- Критерий Дункана – $НСР_{05} = С\ (t_{05} \cdot Sd) = 1,08 \cdot 2,13 \cdot 1,29 = 3,0\ (r=4)$

-Критерий Дункана – $НСР_{05} = С\ (t_{05} \cdot Sd) = 1,10 \cdot 2,13 \cdot 1,29 = 3,0\ (r=5)$

2. Построим ранжированный ряд средних для проверки существенности разности:

Варианты	Августа	Безостая 1	Купавна	Гарант	Немчиновская-57
Средние	35,6	36,2	39,4	41,6	46,3

3. Множественные сравнения средних.

Сначала находим разность между максимальным значением средней величины и минимальным значением средней (в нашем примере $d = 46,3 - 35,6 = 10,7;\ r = 5$; $НСР_{05}$ по Дункану = 3,0, разность существенна). Далее проведем сравнение между максимальным значением средней и второй средней ($d = 46,3 - 36,2 = 10,1;\ r = 4$; $НСР_{05}$ по Дункану = 3,0), затем – с третьей и четвертой средними.

Согласно правилу Дункана, если в ранжированном ряду средних разность между максимальным значением средней и минимальной средней несущественна, то различия между любыми средними, находящимися в промежутке между ними, будут незначимыми и поэтому нет необходимости их проверять на существенность. В нашем примере разность между средней урожайностью сорта Немчиновка – 57 и сорта Август существенна, поэтому будем попарно сравнивать промежуточные средние между собой. При этом,

если между средними разность несущественна, эти средние соединяют линией или обозначают общими буквами.

4. Обозначение несущественных различий:

– помощью линий: Августа; Безостая 1; Купавна; Гарант; Немчиновская-57

35,6 36,2 39,4 41,6 46,3

– помощью букв: Августа; Безостая 1; Купавна; Гарант; Немчиновская-57

35,6 a 36,2 ab 39,4 c 41,6 cd 46,3 e

Средние, соединенные одной линией или обозначенные общими буквами, различаются не существенно.

Итак, по результатам оценки средних урожайных данных на основании многорангового критерия Дункана можно выделить 3 группы: в первую группу входят сорта Август и Безостая 1, во вторую – Купавна и Гарант, и в третью – Немчиновская-57. Внутри каждой группы между сортами нет существенных различий, в то время как между сортами, которые относятся к разным группам отклонения существенны на 5% уровне значимости.

В программе Statistica в таблицах, где представлены результаты сравнения вариантов между собой приводятся не разности между средними значениями, а вероятности для апостериорных тестов (критериев), причем, если ***p<0,05***, то они выделены красным цветом, что указывает на существенность разности между этими средними.

В развернутом диалоговом окне выбора разных форм результатов дисперсионного анализа активируем вкладку **Апостериорные сравнения средних Post-hot comparisons of means** (рис. 5.8), в поле **Отображать** выберем **Значимые разности**, а затем последовательно нажмем на кнопки представленных тестов (Фишера НЗР, Бонферонни, Тьюки и т д.).

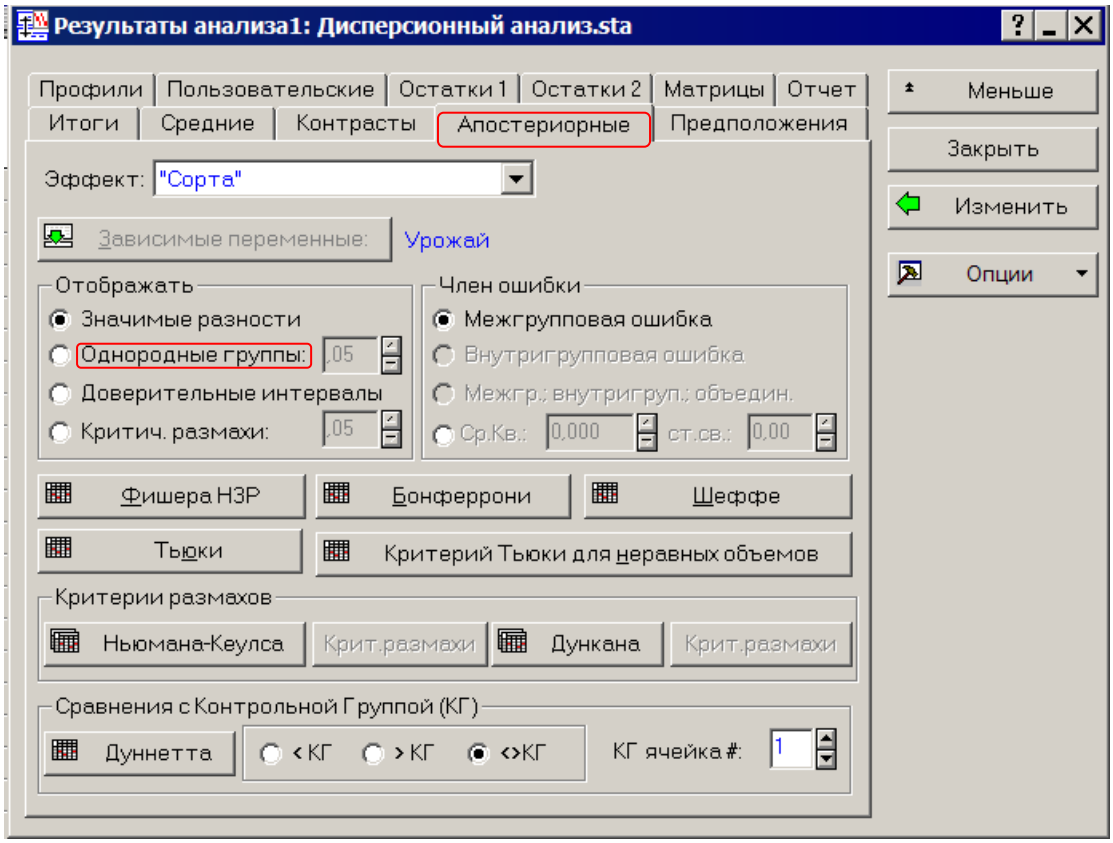


Рис. 5.8. **Диалоговое окно выбора критериев для оценки существенности разности средних (апостериорное сравнение)**

Ниже приводятся результаты сравнения средних урожайных данных по 5 сортам озимой пшеницы с использованием 7-ми тестов (рис. 5.8.1). В шапке всех таблиц цифрами 1,2,3,4,5 обозначены номера вариантов (сортов) и средняя урожайность этих сортов. В ячейке таблицы на пересечении строки и столбца показан уровень значимости ***p*** для проверки нулевой гипотезы о равенстве двух средних, находящихся на пересечении столбца и строки. Красным цветом выделены существенные различия между сравниваемыми сортами на 5% уровне значимости, черным цветом обозначены несущественные различия.

НЗР крит.; перем. Урожай (Дисперсионный анализ.sta)
Вероятности для апостер. критериев
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	{1} 36,225	{2} 39,450	{3} 35,625	{4} 41,600	{5} 46,325
1	Безостая 1		0,024203	0,647709	0,000809	0,000001
2	Купава	0,024203		0,009488	0,115476	0,000082
3	Августа	0,647709	0,009488		0,000318	0,000001
4	Гарант	0,000809	0,115476	0,000318		0,002265
5	Немчиновская-57	0,000001	0,000082	0,000001	0,002265	

Крит. Бонферрони; перем. Урожай (Дисперсионный анализ.sta)
Вероятности для апостер. критериев
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	{1} 36,225	{2} 39,450	{3} 35,625	{4} 41,600	{5} 46,325
1	Безостая 1		0,242031	1,000000	0,008094	0,000011
2	Купава	0,242031		0,094876	1,000000	0,000821
3	Августа	1,000000	0,094876		0,003184	0,000005
4	Гарант	0,008094	1,000000	0,003184		0,022653
5	Немчиновская-57	0,000011	0,000821	0,000005	0,022653	

Крит. Шеффе; перемен. Урожай (Дисперсионный анализ.sta)
Вероятности для апостер. критериев
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	{1} 36,225	{2} 39,450	{3} 35,625	{4} 41,600	{5} 46,325
1	Безостая 1		0,233170	0,993884	0,015436	0,000034
2	Купава	0,233170		0,117199	0,605157	0,001998
3	Августа	0,993884	0,117199		0,006796	0,000017
4	Гарант	0,015436	0,605157	0,006796		0,037090
5	Немчиновская-57	0,000034	0,001998	0,000017	0,037090	

Крит. Тьюки ДЗР; перем. Урожай (Дисперсионный анализ.sta)
Приближенные вероятности для апостер. критериев
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	{1} 36,225	{2} 39,450	{3} 35,625	{4} 41,600	{5} 46,325
1	Безостая 1		0,141430	0,989351	0,006261	0,000154
2	Купава	0,141430		0,061979	0,479089	0,000791
3	Августа	0,989351	0,061979		0,002615	0,000151
4	Гарант	0,006261	0,479089	0,002615		0,016550
5	Немчиновская-57	0,000154	0,000791	0,000151	0,016550	

Крит.Ньюмана-Кеулса; перем. Урожай (Дисперсионный анализ.sta)
Приближенные вероятности для апостер. критериев
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	{1} 36,225	{2} 39,450	{3} 35,625	{4} 41,600	{5} 46,325
1	Безостая 1		0,024343	0,647857	0,002319	0,000189
2	Купава	0,024343		0,024358	0,115612	0,000375
3	Августа	0,647857	0,024358		0,001747	0,000151
4	Гарант	0,002319	0,115612	0,001747		0,002413
5	Немчиновская-57	0,000189	0,000375	0,000151	0,002413	

Крит. Дункана; перем. Урожай (Дисперсионный анализ.sta)
Приближенные вероятности для апостер. критериев
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	{1} 36,225	{2} 39,450	{3} 35,625	{4} 41,600	{5} 46,325
1	Безостая 1		0,024343	0,647857	0,001160	0,000063
2	Купава	0,024343		0,012254	0,115612	0,000187
3	Августа	0,647857	0,012254		0,000583	0,000038
4	Гарант	0,001160	0,115612	0,000583		0,002413
5	Немчиновская-57	0,000063	0,000187	0,000038	0,002413	

Крит. Дуннетта; перем. Урожай (Дисперсионный анализ.s
Вероятности для апостер. критериев (2-стор.)
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	{1} 36,225
1	Безостая 1	
2	Купава	0,075796
3	Августа	0,970529
4	Гарант	0,002855
5	Немчиновская-57	0,000009

Рис. 5.8.1. **Сравнение средних по 7-ми критериям**

Для группировки средних в зависимости от их значений выберем **Однородные группы,** (рис. 5.8) и тогда средние значения будут распределены в однородные группы. В однородные группы попадают средние значения, между которыми различия несущественны, в то время как различия между разнородными группами существенны. Так, в первую группу входят сорта Августа и Безостая 1, разность между средними у этих сортов незначима ($d = 36,225 – 35,625 = 0,6 < НСР_{05}=2,7$), во вторую группу – сорта Купава и Гарант ($d=2,15 < НСР_{05}=2,7$), в третьей группе сорт Немчиновская-57 (рис. 5.8.2)

НЗР крит.; перем. Урожай (Дисперсионный анализ.sta)
Однородные группы, alpha = ,05000
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	Урожай Среднее	1	2	3
3	Августа	35,62500	****		
1	Безостая 1	36,22500	****		
2	Купава	39,45000		****	
4	Гарант	41,60000		****	
5	Немчиновская-57	46,32500			****

Крит. Шеффе; перемен. Урожай (Дисперси
Однородные группы, alpha = ,05000
Ошибка: Межгр. MS = 3,3115, cc = 15,000

N ячейки	Сорта	Урожай Среднее	1	2	3
3	Августа	35,62500	****		
1	Безостая 1	36,22500	****		
2	Купава	39,45000	****	****	
4	Гарант	41,60000		****	
5	Немчиновская-57	46,32500			****

Рис. 5.8.2. **Распределение вариантов в однородные группы**

Результаты группировки сортов, приведенные на рис. 5.8.2, полностью совпадают с оценкой указанных сортов по критерию Дункана, описанной на стр. 66.

5.1.3 Графическое изображение средних с доверительными интервалами

Для построения графика средних значений по сортам с доверительными интервалами вернемся к панели **Результаты анализа**, в которой активируем вкладку **Средние** и в появившемся окне нажмем на клавишу **График (Plot)** рядом с кнопкой **Наблюдаемые, невзвешенные (Observed, unweighted)** и **График (Plot)** рядом с кнопкой **Наблюдаемые, взвешенные (Observed, weighted)** (рис. 5.9).

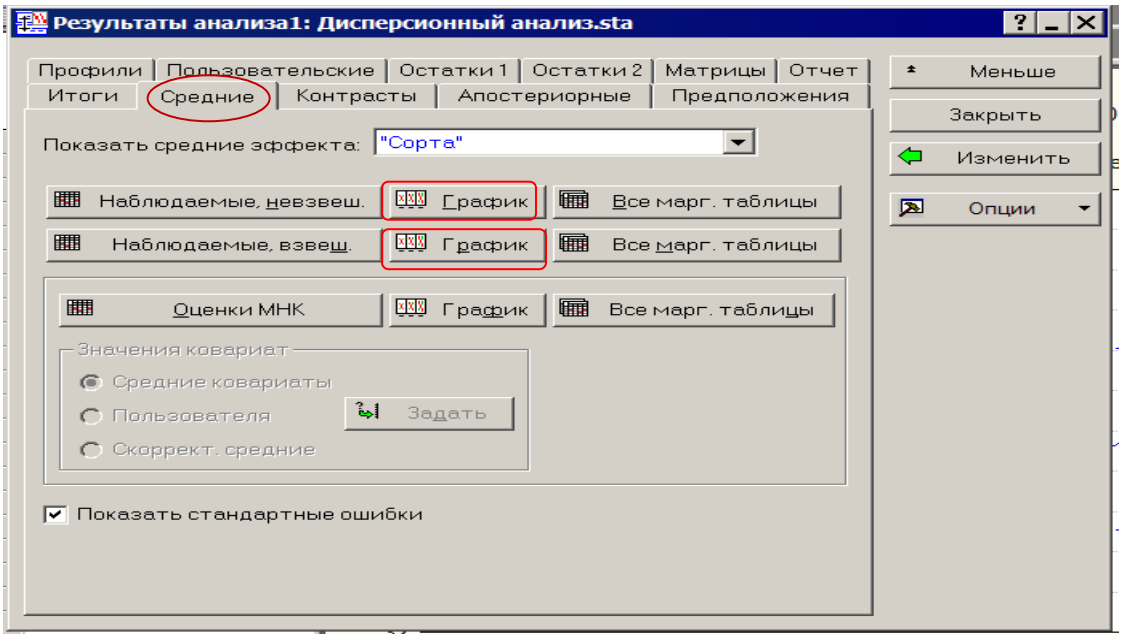


Рис. 5.9. **Диалоговое окно для выбора графиков средних значений по сортам**

В результате получаем два графика со средними урожайными данными по изучаемым сортам с их доверительными интервалами. На первом графике (рис. 5.10) величина доверительного интервала для всех сортов усреднена, она рассчитана по следующей формуле t_{05} $S_{\bar{x}}$ $S_{\bar{x}} = \sqrt{\frac{mS}{n}} = \sqrt{\frac{3{,}31}{4}} = 0{,}91$, где mS– средний квадрат остатка или остаточная дисперсия, которую берем из таблицы дисперсионного анализа (рис. 5.7).

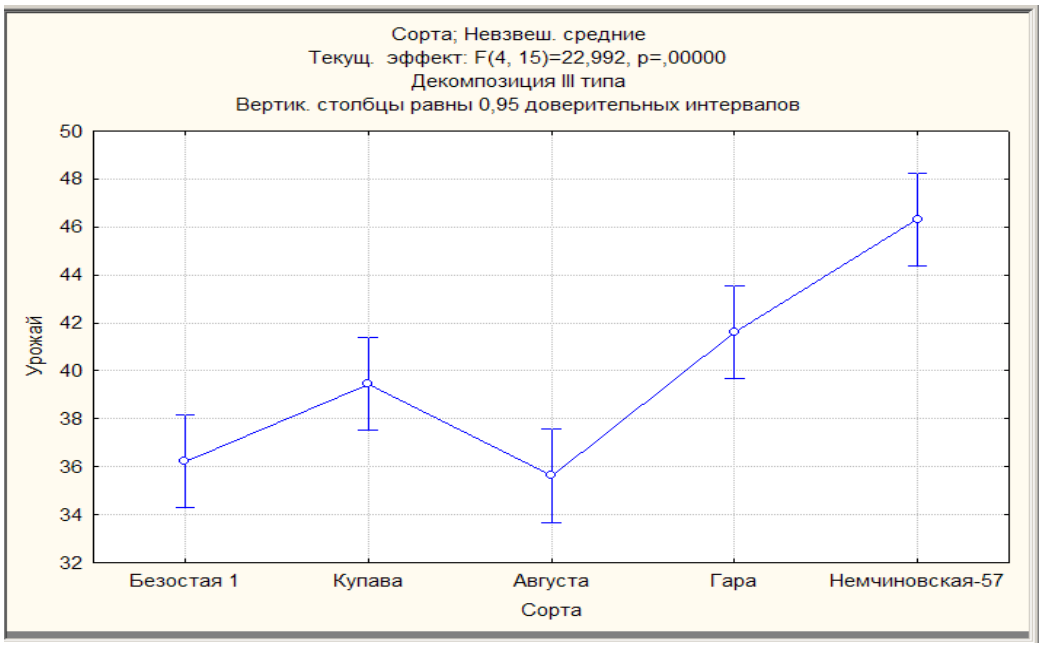


Рис. 5.10. **График средних с усредненными доверительными интервалами**

На втором графике величина доверительного интервала для каждого варианта рассчитана с учетом изменчивости внутри каждого варианта (рис. 5.11). Величины доверительных интервалов для сортов Гарант и Немчиновская 67 значительно больше, так как у этих сортов размах варьирования примерно в два раза выше, чем у других сортов, о чем свидетельствуют исходные данные.

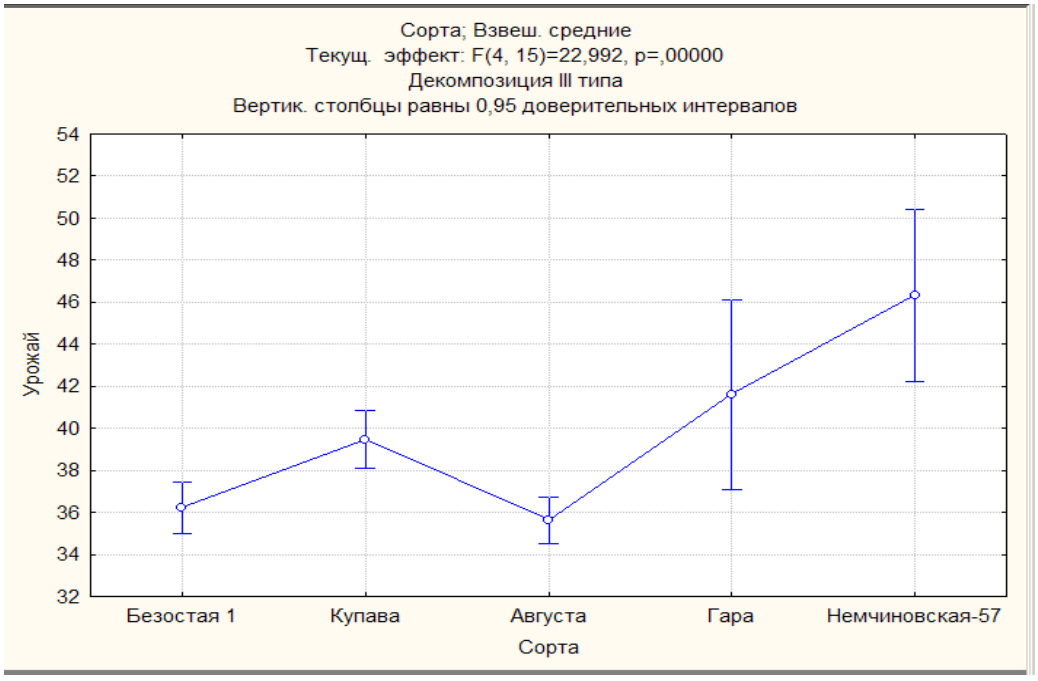


Рис. 5.11. **График средних с индивидуальными доверительными интервалами**

О существенности разности средних можно судить по представленным графикам, где вертикальные столбцы показывают величину 95% доверительного интервала для каждой средней. Если нижняя граница доверительного интервала одной средней заходит на верхнюю границу доверительного интервала другой средней, значит, различия между этими средними не значимы, а если интервалы не пересекаются, то различия существенны.

Так, в подтверждении результатов множественных сравнений средних между собой на основании графического представления можно сделать вывод о несущественности разности между 1,3 и 2, 4 вариантами

Из двух представленных графиков наиболее точные результаты получаем при использовании диаграммы размаха с взвешенными средними, где доверительные интервалы вычислены для каждой средней с учетом внутригрупповой вариации каждого варианта.

5.2 Дисперсионный анализ данных однофакторных экспериментов с рандомизированными (организованными) повторениями (блоками)

Пример 2. В полевом опыте изучается влияние 4-х вариантов обработки почвы на урожайность ячменя, опыт проведен методом организованных повторений (блоков). Повторность опыта четырехкратная.

Варианты опыта	Повторения			
	I	II	III	IV
1.Обычная	39,8	41,3	40,1	39,6
2. Глубокая	40,3	39,5	38,1	37,3
3.Дисковая	35,6	36,8	38,3	34,4
4. Фрезерная	42,5	43,2	42,7	41,8

В программе Statistica создадим файл исходных данных с тремя переменными: *Повторения, Обработка почвы, Урожай* и введем данные однофакторного опыта, заложенного методом организованных повторений в следующем виде:

Повторения	Обработка почвы	Урожай
I	Обычная	39,8
II	Обычная	41,3
III	Обычная	40,1
IV	Обычная	39,6
I	Глубокая	40,3
II	Глубокая	39,5
III	Глубокая	38,1
IV	Глубокая	37,3
I	Дисковая	35,6
II	Дисковая	36,8
III	Дисковая	38,3
IV	Дисковая	34,4
I	Фрезерная	42,5
II	Фрезерная	43,2
III	Фрезерная	42,7
IV	Фрезерная	41,8

В программе Statistica в модуле **Однофакторный дисперсионный анализ (One-way ANOVA)** обрабатываются только лишь данные экспериментов с 2-мя переменными – зависимая переменная (количественные значения признака изучаемого фактора) и категориальный или независимый предиктор (градации или варианты изучаемого фактора). К таким экспериментам в агрономических исследованиях относятся данные однофакторных вегетационных или полевых опытов, заложенных методом полной рандомизации. Повторения как третья переменная в опции ANOVA однофакторного опыта программой Statistica игнорируются. Поэтому для обработки данных полевого однофакторного опыта с размещением вариантов методом организованных (рандомизированных) повторений можно использовать в меню **Дисперсионный анализ (Anova)** опцию **Факторный анализ (Faktorial Anova)** (рис. 5.12) с последующими дополнительными расчетами

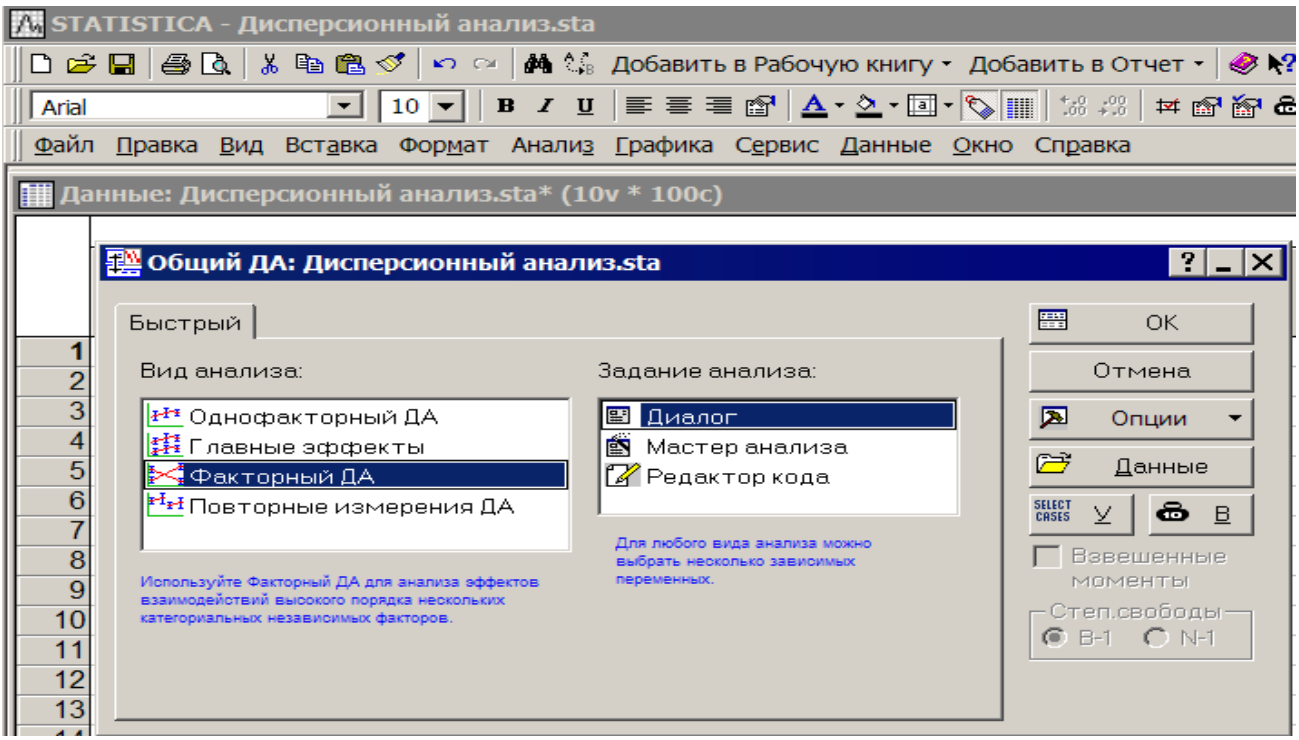


Рис. 5.12. **Диалоговое окно выбора модели дисперсионного анализа**

В появившемся окне (рис. 5.13) выберем переменные: в окошке **Зависимые переменные** – *Урожай*, в окошке **Независимые предикторы** – *Повторения и Обработка почвы.*

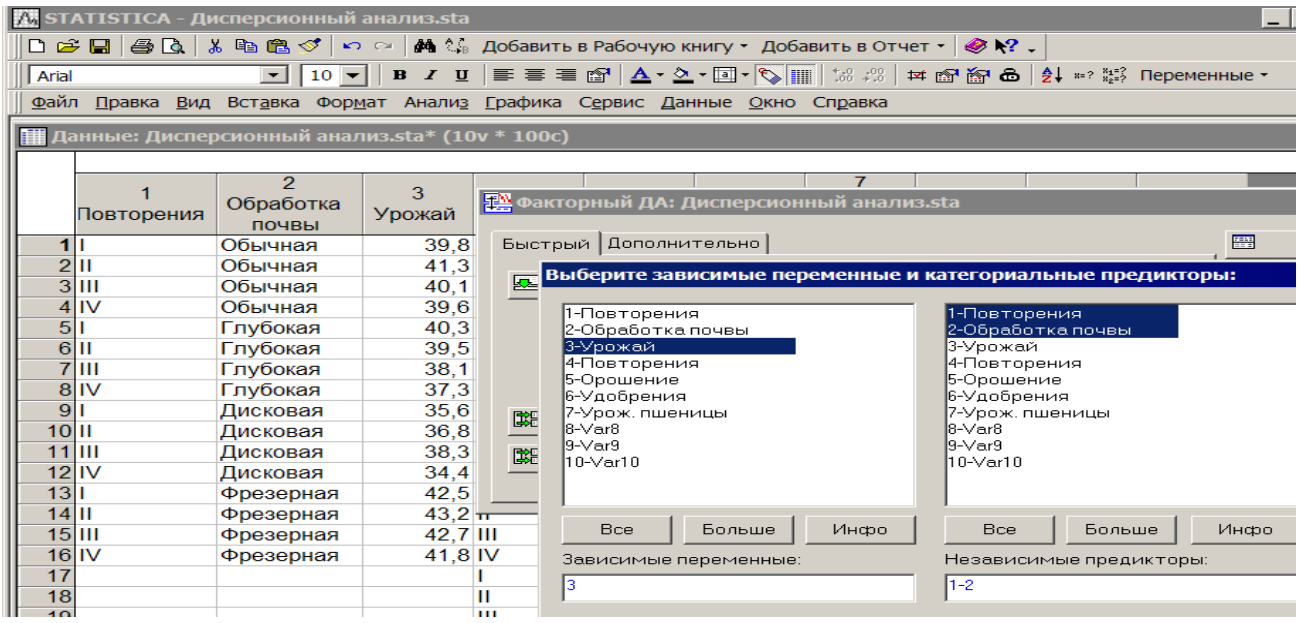


Рис. 5.13. **Диалоговое окно для ввода переменных**

Далее в расширенной панели вывода результатов дисперсионного анализа нажмем на кнопку **Одномерные результаты** (Рис. 5.14)

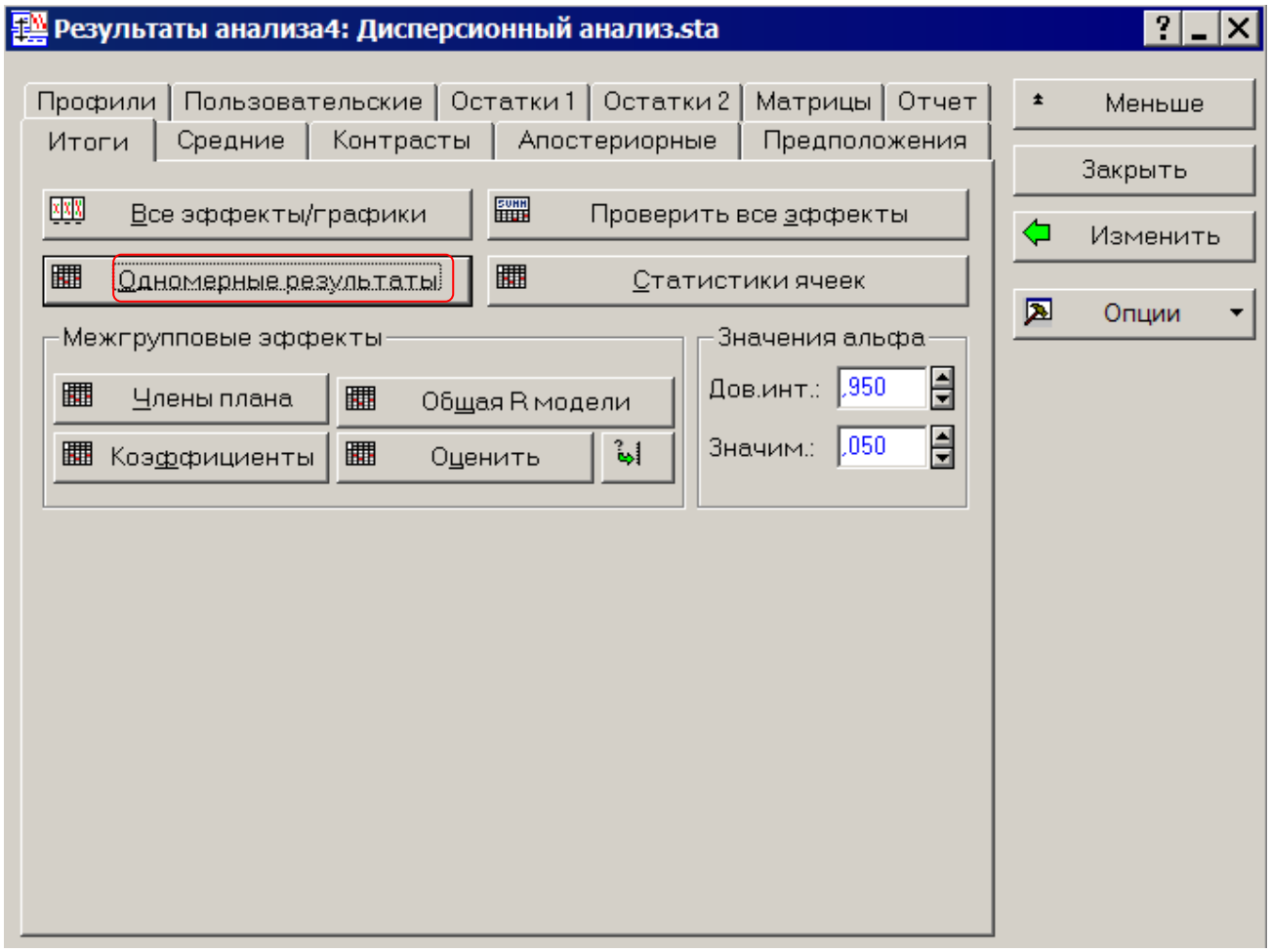


Рис. 5.14. **Расширенная панель вывода результатов дисперсионного анализа**

После нажатия на кнопку **Одномерные результаты** в рабочей книге получаем основную таблицу дисперсионного анализа (рис. 5.15), в которой общая сумма квадратов и степени свободы разложены на следующие компоненты:

Суммы квадратов $SS_{общая} = SS_{повторения} + SS_{обработка\ почвы} + SS_{повт*обработка\ почвы}$

$$99{,}28 = 8{,}30 + 82{,}70 + 8{,}28$$

Степени свободы $15 = 3 + 3 + 9$

Эффект	Одномерные результаты для каждой ЗП Сигма-ограниченная параметризация Декомпозиция гипотезы **Степени свободы**	Урожай SS	Урожай MS	Урожай F
Св. член	1	24908,73	24908,73	
Повторения	3	8,30	2,77	
Обработка почвы	3	82,70	27,57	
Повторения*Обработка почвы	9	8,28	0,92	
Ошибка	0			
Всего	15	99,28		

Рис. 5.15. **Таблица дисперсионного анализа**

В данном случае $SS_{повт*обработка\ почвы}$ = 8,28 и средний квадрат отклонений (дисперсия) = 0,92 характеризуют остаточную сумму квадратов и остаточную (случайную) дисперсию в однофакторном опыте, заложенным методом организованных повторений.

Критерий Фишера и НСР можно рассчитать вручную на калькуляторе или воспользоваться опциями **Формулы** в пакете Statistica

$$F_ф = \frac{S_v^2}{S_e^2} = \frac{mS_v}{mS_e} = \frac{27{,}57}{0{,}92} = 29{,}97$$

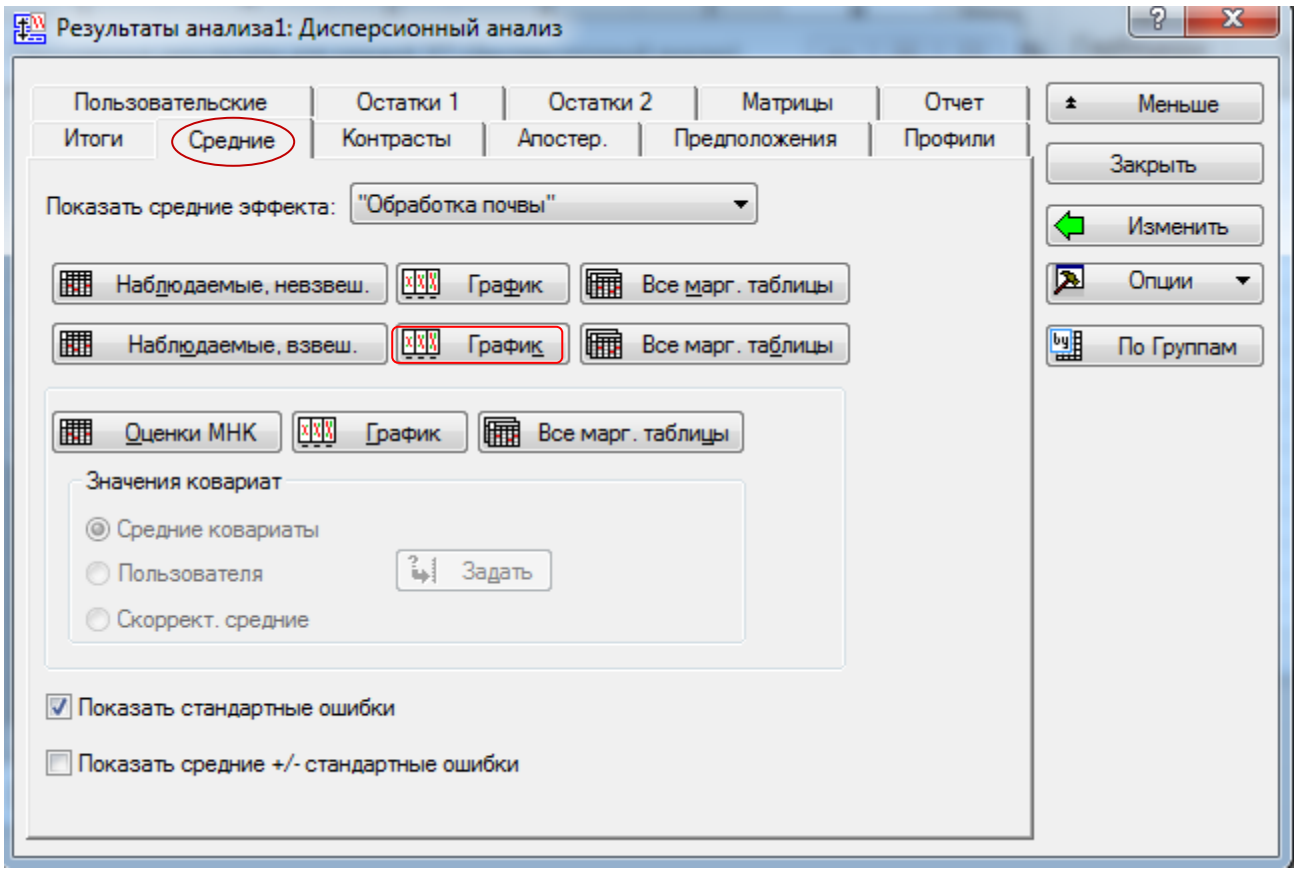
F_{05} = 3,86 F_{01} =6,99. *при* *c.c.v* =3, *c.c.e* =9.

Так как $F_ф > F_{01}$ $H_0 \neq 0$, нулевая гипотеза с вероятностью 99% отвергается – в опыте в целом есть существенные различия, поэтому необходимо рассчитать *НСР*.

$$S_d = \sqrt{\frac{2S_e^2}{n}} = \sqrt{\frac{2 \cdot 0{,}92}{4}} = 0{,}68 \qquad НСР_{05} = t_{05} \cdot * S_d = 2{,}26 \cdot 0{,}68 = 1{,}53 \text{ ц/га}$$

t_{05} = 2,26 при *df (cce)* = 9 степенях свободы для остатка

Для графического представления средних значений с 95% доверительными интервалами в расширенном диалоговом окне результатов дисперсионного анализа нажмем на вкладку **Средние (Means)** и в появившемся диалоговом окне в окошке **Показать средние эффекта** выберем «Обработка почвы» и нажмем на клавишу **График** напротив **Наблюдаемые, взвешенные** (рис. 5.16).



5.16. **Диалоговое окно выбора графика средних значений**

Получаем диаграмму размаха средних значений урожайности ячменя в зависимости от разных обработок почвы (рис. 5.17).

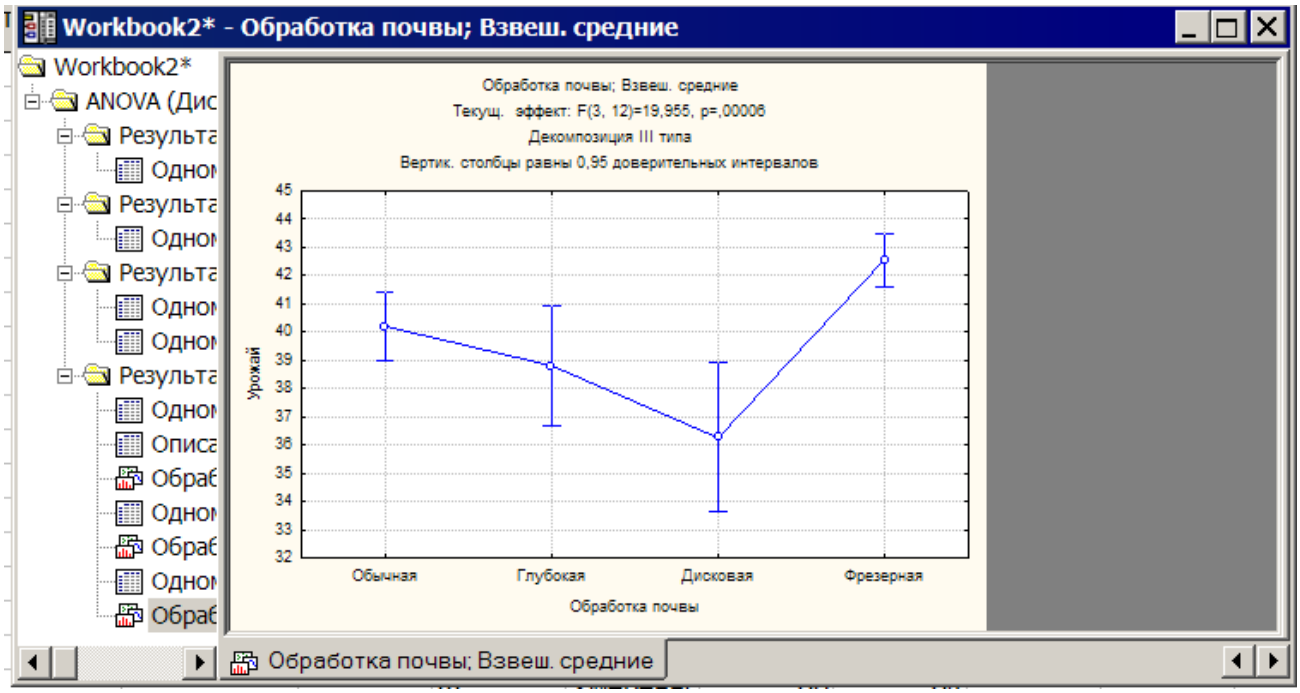


Рис. 5.17. **График средних с индивидуальными доверительными интервалами**

Вертикальными отрезками на графике указаны значения 95% доверительных интервалов, по которым визуально можно судить о существенности действия обработок почвы на урожайность ячменя, и тем самым проверять нулевую гипотезу в отношении каждой пары вариантов. Так, из графика видно, что самым лучшим вариантом в опыте является фрезерная обработка почвы – нижняя граница доверительного интервала генеральной средней у данного варианта выше верхней границы доверительных интервалов обычной, глубокой и дисковой обработок почвы. Дисковая обработка приводит к существенному уменьшению урожайности, снижение урожайности при глубокой обработки не существенно по сравнению с обычной обработкой почвы.

5.3 Дисперсионный анализ многофакторных опытов

Пример 3. В двухфакторном полевом опыте 3х4, заложенным методом полной рандомизации изучается влияние орошения и доз азота на урожайность озимой пшеницы, ц/га.

Фактор А – 3 уровня орошения (***0*** *– без орошения,* ***1*** *– умеренное,* ***2*** *– обильное*)

Фактор В – 4 дозы азота, кг/га (***0*** *– без азота,* ***1*** *– 60 ,* ***2*** *– 90 ,* ***3*** *– 120*)

Повторность опыта, *n =4*. Общее число делянок, *N = 3*4*4= 48.*

Фактор А Орошение	Фактор В Удобрения	Повторность			
		I	*II*	*III*	*IV*
Без орошения 0	*0*	30	41	41	34
	60	42	41	44	46
	90	38	41	39	37
	120	40	40	42	44
Умеренное 1	*0*	50	50	52	52
	60	60	60	64	65
	90	67	68	67	69
	120	63	59	70	68
Обильное 2	*0*	52	49	57	56
	60	72	69	73	72
	90	64	66	62	64
	120	72	71	70	73

Создадим в программе Statistica файл исходных данных с 3-мя переменными: *Орошение, Удобрение и Урож. пшеницы*, 48 наблюдениями и занесем данные как показано на рис.5.18.

	Орошение	Удобрение	Урож. пшеницы
1	О	0	30
2	О	0	41
3	О	0	41
4	О	0	34
5	Умеренное	0	50
6	Умеренное	0	50
7	Умеренное	0	52
8	Умеренное	0	52
9	Обильное	0	52
.	.	.	.
.	.	.	.
.	.	.	.
46	Обильное	120	71
47	Обильное	120	70
48	Обильное	120	73

Рис. 5.18. **Окно исходных данных**

Щелкнем по кнопке **Анализ (Statistics),** выберем опцию **Дисперсионный анализ (Anova)**, в открывшемся окне – вид дисперсионного анализа **Факторный ДА (Faktorial Anova),** раздел **Диалог (Quick specs dialog),** после нажатия на **Ок** попадем в окно выбора **Переменных (Variabels)** (рис. 5.19).

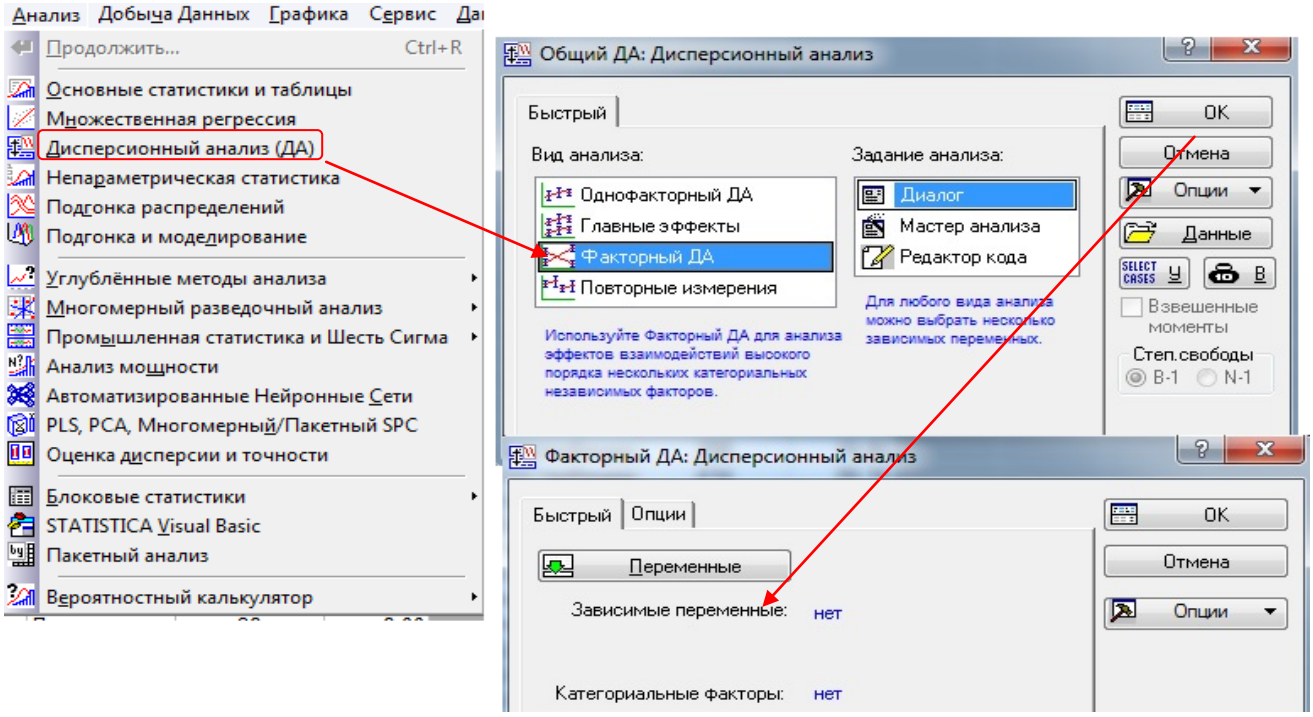


Рис. 5.19. **Диалоговое окно выбора модуля ANOVA**

В появившемся окне в качестве зависимых переменных укажем *Урожайность пшеницы*, категориальных предикторов изучаемые факторы – *Орошения и Удобрения* (рис.5.20).

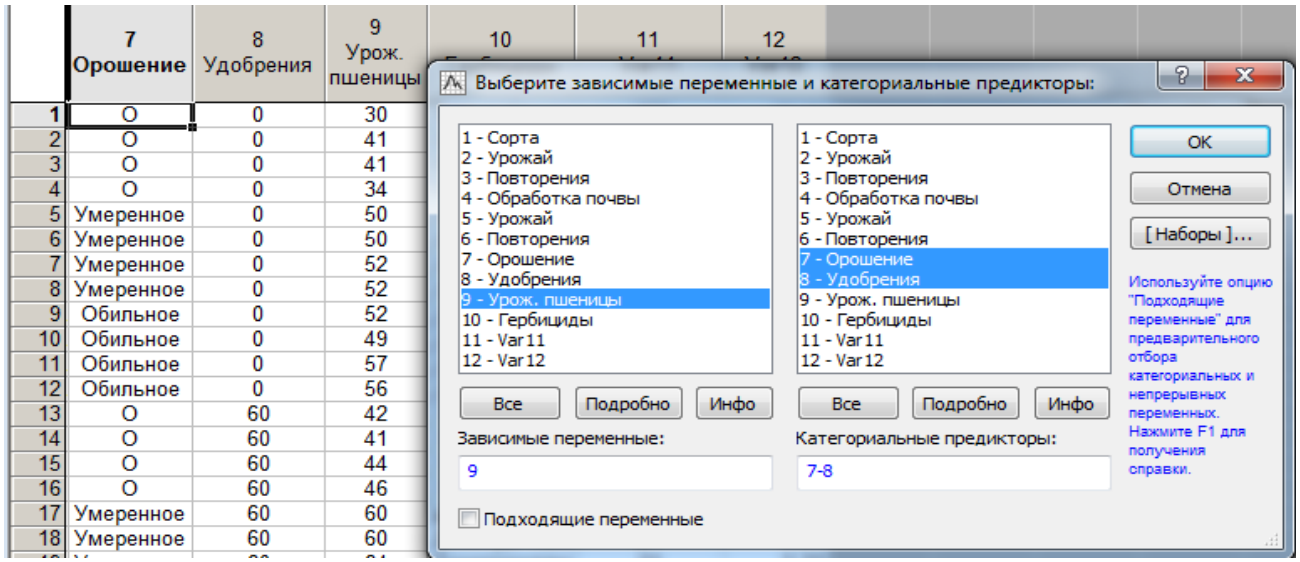


Рис. 5.20. **Диалоговое окно выбора переменных**

Далее в краткой панели результатов дисперсионного анализа нажмем на клавишу **Больше,** открывается расширенное диалоговое окно вывода результатов дисперсионного анализа в которой последовательно выбираем нужные нам формы результатов дисперсионного анализа.

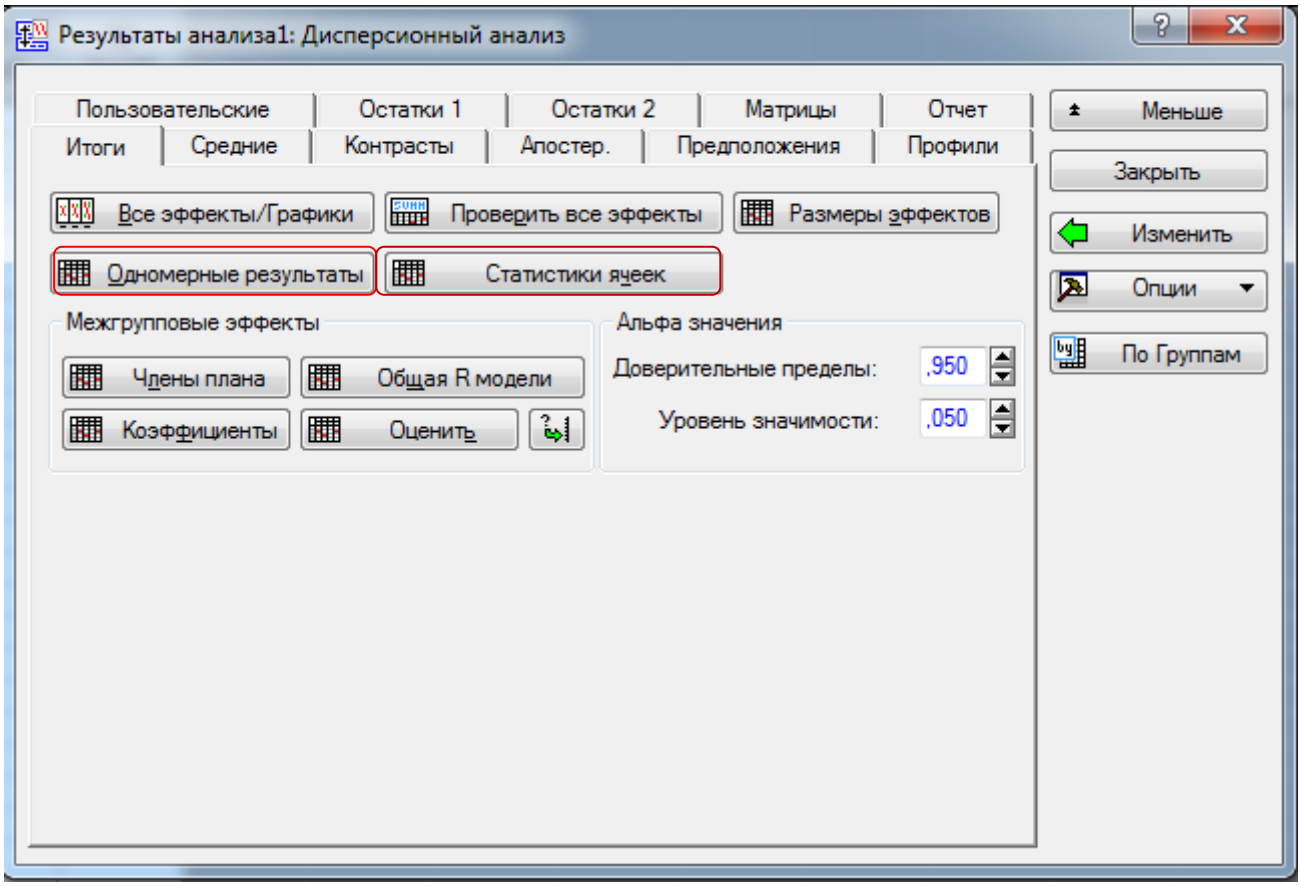


Рис.5.21. **Расширенное диалоговое окно результатов дисперсионного анализа**

В расширенном диалоговом окне результатов дисперсионного анализа (рис. 5.21) нажмем на кнопку **Одномерные результаты** и получим в рабочей книге основную таблицу дисперсионного анализа для двухфакторного анализа, в которой представлены суммы квадратов (*SS*), степени свободы, дисперсии (*MS*), критерий Фишера и уровни вероятности (*p)* для оценки значимости главных эффектов фактора А (орошение), фактора В (удобрение) и их взаимодействия АВ (рис. 5.22). Так как эти показатели выделены красным цветом и *p* намного меньше 0,05, с вероятностью не только 95%, но и с вероятностью 99% отмечаем существенное влияние орошения, удобрения и их взаимодействия на урожайность озимой пшеницы.

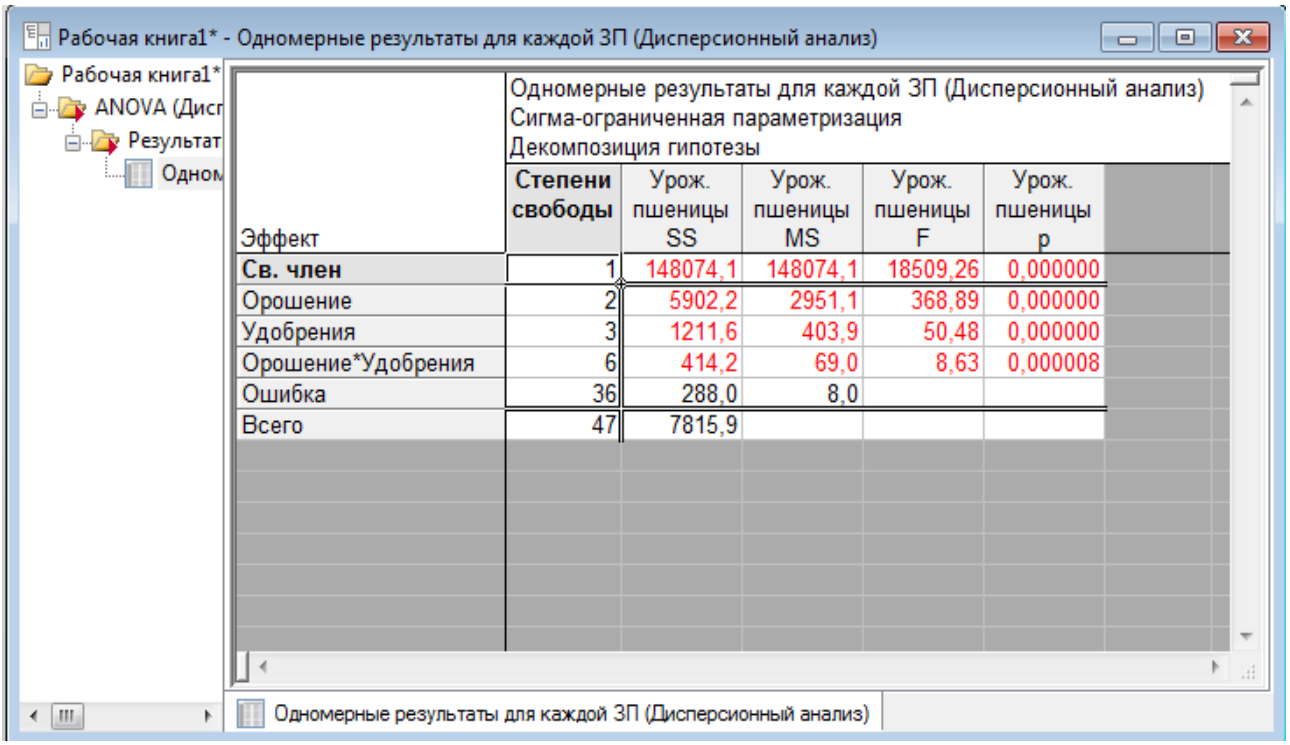


Рис. 5.22. **Таблица дисперсионного анализа**

Для получения значений описательной статистики по данным нашего опыта в расширенном диалоговом окне вывода результатов дисперсионного анализа нажмем на кнопку **Статистики ячеек** (рис. 5.21). После нажатия на клавишу **Ок** получим в рабочей книге полную таблицу статистических показателей по результатам двухфакторного опыта: средние значения, показатели вариации, ошибки средних и доверительные интервалы по фактору А – Орошение, фактору В – Удобрение и их сочетания – АВ (рис. 5.23). В принципе по 95% доверительным интервалам для генеральных средних (последние две колонки: нижняя и верхняя границы интервалов) можно провести оценку существенности главных эффектов, а также частных средних. Если 95% доверительные интервалы для генеральных средних двух сравниваемых вариантов опыта перекрываются, значит, между этими вариантами нет существенных различий, если же доверительные интервалы не перекрываются, различия существенны на 05% уровне значимости.

Эффект	Описательные статистики (Все примеры)							
	Уровень Фактор	**Уровень Фактор**	N	Урож. пшеницы Среднее	Урож. пшеницы Ст.откл.	Урож. пшеницы Стд.ош.	Урож. пшеницы -95,00%	Урож. пшеницы +95,00%
Всего			48	55,54	12,90	1,86	51,80	59,29
Орошение	О		16	40,00	3,92	0,98	37,91	42,09
Орошение	Умеренное		16	61,50	7,07	1,77	57,73	65,27
Орошение	Обильное		16	65,13	7,89	1,97	60,92	69,33
Удобрения	0		12	47,00	8,57	2,47	41,55	52,45
Удобрения	60		12	59,00	12,45	3,59	51,09	66,91
Удобрения	90		12	56,83	13,52	3,90	48,25	65,42
Удобрения	120		12	59,33	13,76	3,97	50,59	68,08
Орошение*Удобрения	О	0	4	36,50	5,45	2,72	27,83	45,17
Орошение*Удобрения	О	60	4	43,25	2,22	1,11	39,72	46,78
Орошение*Удобрения	О	90	4	38,75	1,71	0,85	36,03	41,47
Орошение*Удобрения	О	120	4	41,50	1,91	0,96	38,45	44,55
Орошение*Удобрения	Умеренное	0	4	51,00	1,15	0,58	49,16	52,84
Орошение*Удобрения	Умеренное	60	4	62,25	2,63	1,31	58,07	66,43
Орошение*Удобрения	Умеренное	90	4	67,75	0,96	0,48	66,23	69,27
Орошение*Удобрения	Умеренное	120	4	65,00	4,97	2,48	57,10	72,90
Орошение*Удобрения	Обильное	0	4	53,50	3,70	1,85	47,62	59,38
Орошение*Удобрения	Обильное	60	4	71,50	1,73	0,87	68,74	74,26
Орошение*Удобрения	Обильное	90	4	64,00	1,63	0,82	61,40	66,60
Орошение*Удобрения	Обильное	120	4	71,50	1,29	0,65	69,45	73,55

Рис. 5.23. **Статистические показатели данных двухфакторного опыта 3х4**

Для оценки частных различий возвращаемся к диалоговому окну вывода результатов дисперсионного анализа (рис. 5.24) и выделим вкладку **Апостериорные**, в окошке **Эффект** укажем Орошение*Удобрения, отметим **Значимые разности** и **Межгрупповую ошибку**, выберем классический тест *НСР* с этой целью нажмем на кнопку **Фишера НЗР**.

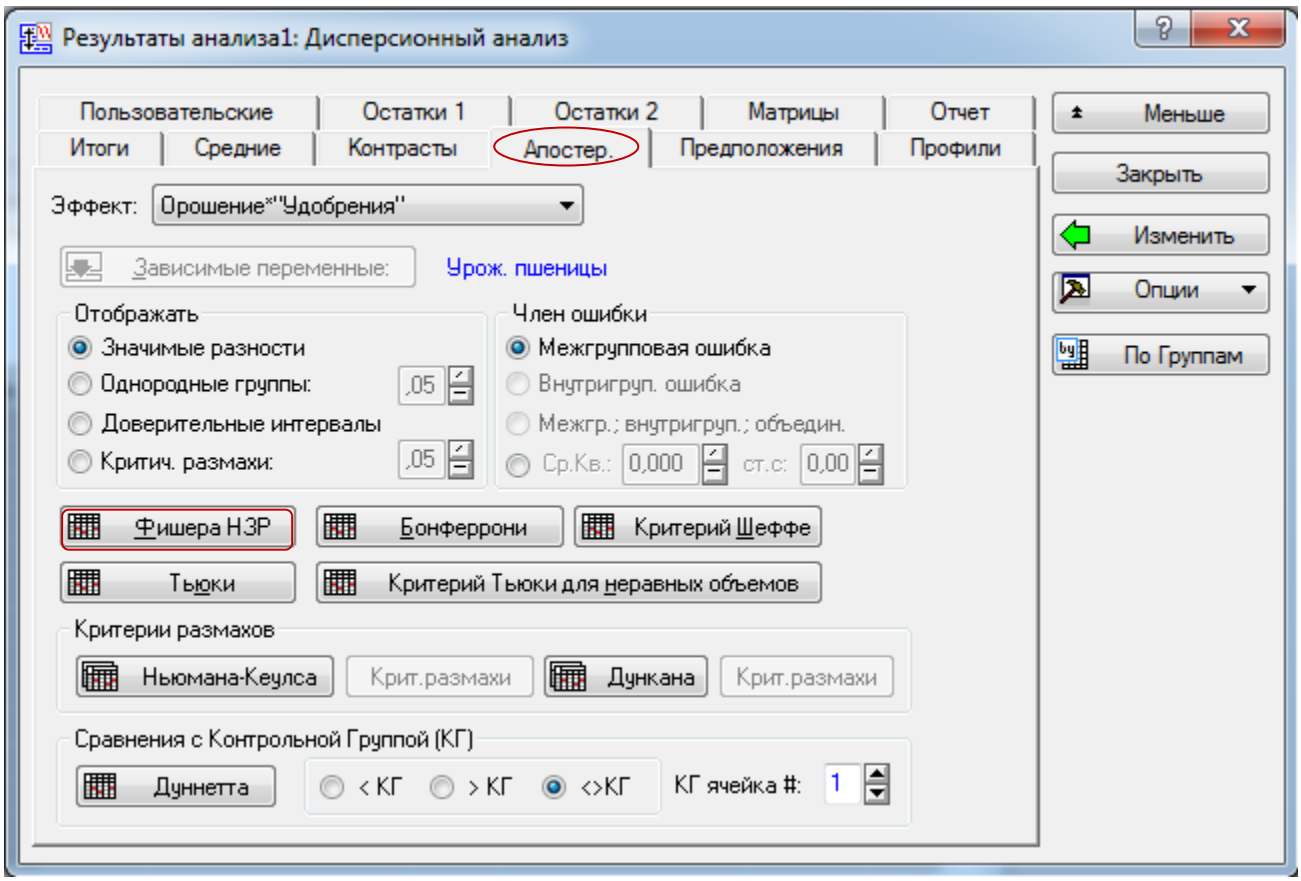


Рис. 5.24. **Диалоговое окно вывода результатов дисперсионного анализа**

В итоге получаем полную таблицу сравнения всех частных средних между собой, в которой красным цветом выделены существенные разности между средними и черным цветом – несущественные разности.

НЗР крит.; перем. Урож. пшеницы (Все примеры)
Вероятности для апостер. критериев
Ошибка: Межгр. MS = 8,0000, сс = 36,000

N ячейки	Орошение	Удобрения	{1} 36,500	{2} 43,250	{3} 38,750	{4} 41,500	{5} 51,000	{6} 62,250	{7} 67,750	{8} 65,000	{9} 53,500	{10} 71,500	{11} 64,000	{12} 71,500
1	О	0		0,00	0,27	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	О	60	0,00		0,03	0,39	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3	О	90	0,27	0,03		0,18	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4	О	120	0,02	0,39	0,18		0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
5	Умеренное	0	0,00	0,00	0,00	0,00		0,00	0,00	0,00	0,22	0,00	0,00	0,00
6	Умеренное	60	0,00	0,00	0,00	0,00	0,00		0,01	0,18	0,00	0,00	0,39	0,00
7	Умеренное	90	0,00	0,00	0,00	0,00	0,00	0,01		0,18	0,00	0,07	0,07	0,07
8	Умеренное	120	0,00	0,00	0,00	0,00	0,00	0,18	0,18		0,00	0,00	0,62	0,00
9	Обильное	0	0,00	0,00	0,00	0,00	0,22	0,00	0,00	0,00		0,00	0,00	0,00
10	Обильное	60	0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,00	0,00		0,00	1,00
11	Обильное	90	0,00	0,00	0,00	0,00	0,00	0,39	0,07	0,62	0,00	0,00		0,00
12	Обильное	120	0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,00	0,00	1,00	0,00	

Рис. 5.25. **Вероятности значимости существенных и несущественных разностей между частными средними**

Для оценки главных эффектов изучаемых факторов (А и В) в диалоговом окне вывода результатов дисперсионного анализа (рис. 5.26)

активируем вкладку **Апостер** в поле ввода **Эффект** выберем последовательно «*Орошение*» и «*Удобрения*», нажмем на кнопку **Фишера НЗР.**

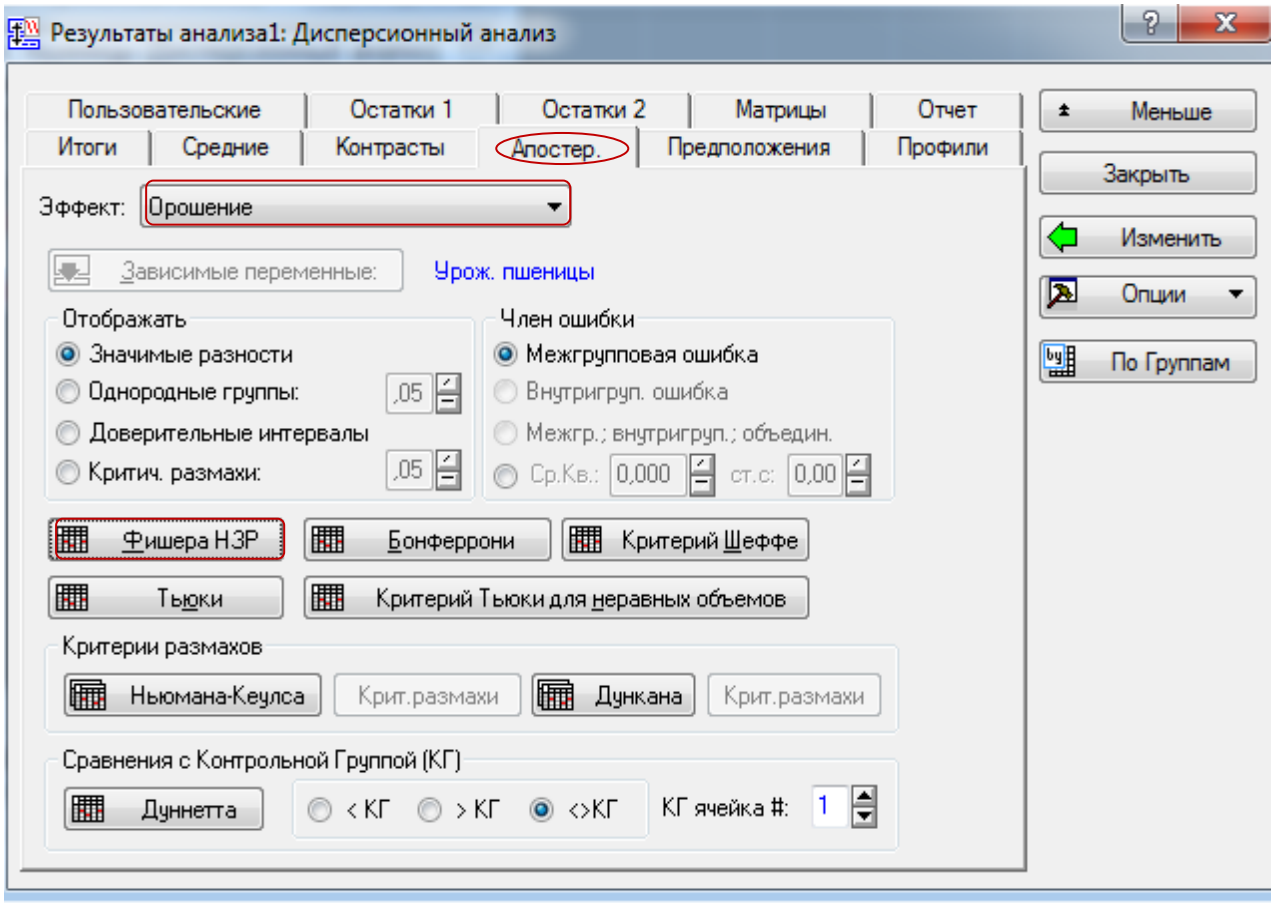


Рис. 5.26. **Диалоговое окно вывода результатов дисперсионного анализа**

В рабочей книге получим 2 таблицы с результатами оценки главных эффектов фактора *А* – *Орошение* и фактора *В* – *Удобрение* (рис. 5.27). В первой таблице представлены вероятности значимости результатов сравнения средних по фактору *Орошения* независимо от удобрений (главный эффект фактора А), а во второй – вероятности значимости результатов сравнения средних по фактору *Удобрения* независимо от орошения (главный эффект фактора В). Красным цветом выделены значимые разности.

N ячейки	НЗР крит.; перем. Урож. пшеницы (Дисперсионный анализ) Вероятности для апостер. критериев Ошибка: Межгр. MS = 8,0000, cc = 36,000 **Орошение**	{1} 40,000	{2} 61,500	{3} 65,125
1	О		0,000000	0,000000
2	Умеренное	0,00		0,000886
3	Обильное	0,00	0,000886	

N ячейки	НЗР крит.; перем. Урож. пшеницы (Дисперсионный анализ) Вероятности для апостер. критериев Ошибка: Межгр. MS = 8,0000, cc = 36,000 **Удобрения**	{1} 47,000	{2} 59,000	{3} 56,833	{4} 59,333
1	0		0,000000	0,000000	0,000000
2	60	0,000000		0,068727	0,774486
3	90	0,000000	0,068727		0,037089
4	120	0,000000	0,774486	0,037089	

Рис.5.27. **Результаты оценки главных эффектов**

Для графического представления средних значений в диалоговом окне выбора формы результатов дисперсионного анализа (рис. 5.28) активируем вкладку **Средние,** в окошке **Показать средние эффекта** укажем сочетание «Орошение*Удобрения» и нажмем на клавишу **График** напротив **Наблюдаемые, взвеш.** В появившейся панели (рис. 5.28) выберем факторы для представления на X-оси и Y-оси – шаблон линии. В качестве X-оси лучше выбрать переменную (фактор), имеющую большее число градаций, в нашем случае – *Удобрения*, а шаблон линии – *Орошение*.

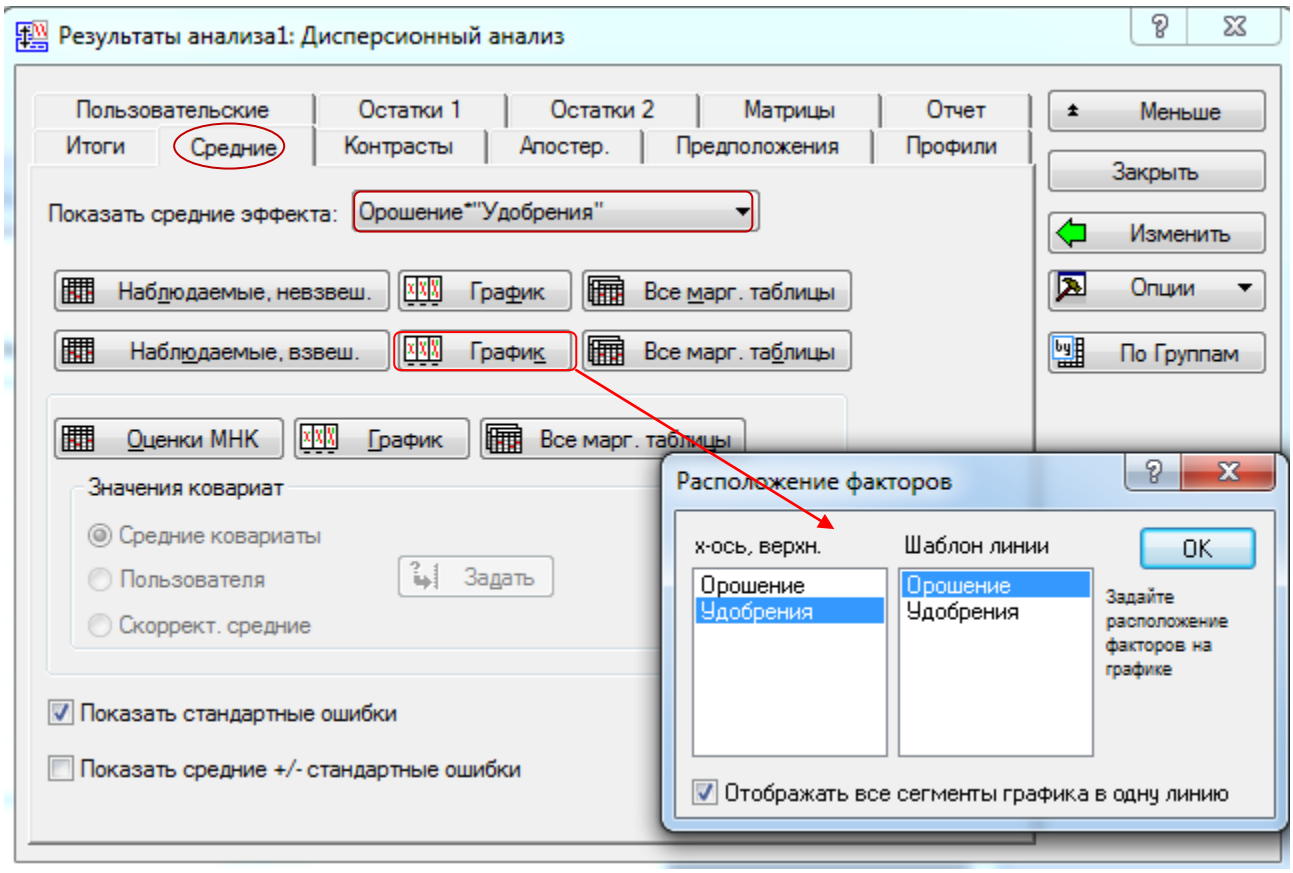


Рис. 5.28. **Диалоговое окно для выбора графика и расположения факторов**

После нажатия на кнопку **Ок** в рабочей книге получаем график (рис. 5.29).

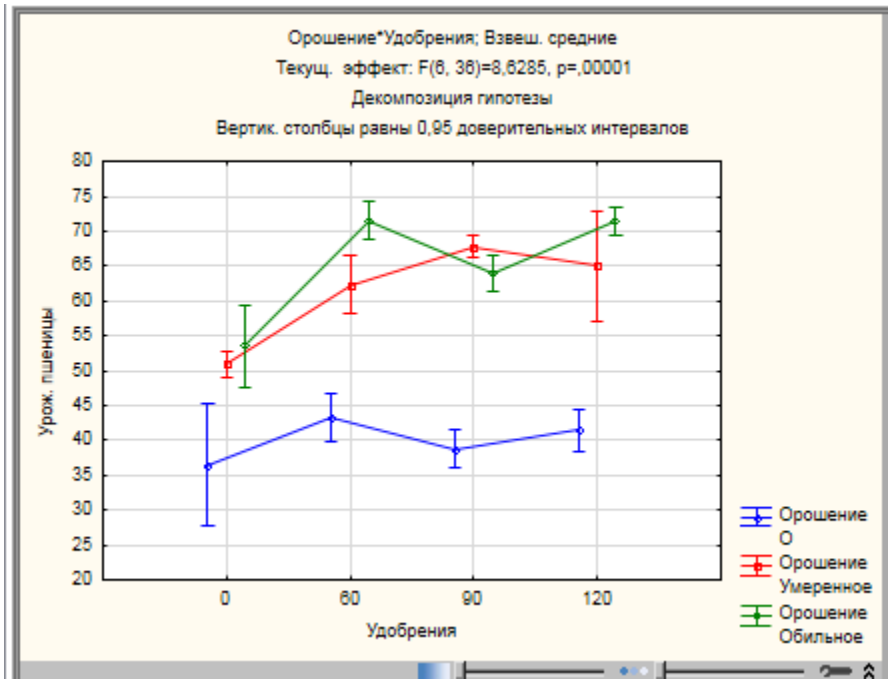


Рис. 5.29. **График средних значений урожайности озимой пшеницы**

На графике точками обозначены средние значения урожайности озимой пшеницы при разных дозах азотных удобрений и разных нормах орошения, а вертикальными отрезками по каждой средней показаны 95% доверительные интервалы для генеральных средних, по которым можно провести оценку существенности разности частных средних.

Подобные графики можно построить для оценки главных эффектов факторов *Орошение и Удобрения*.

5.4 Дисперсионный анализ данных с неоднородными выборками

Если исходные данные соответствуют предпосылкам дисперсионного анализа (однородные дисперсии, нормальное распределение ошибок) и проходят тест Бартлетта, то проводится обычный дисперсионный анализ без трансформации исходных данных с применением параметрических критериев (*F, t*), как указано в вышеприведенных примерах.

Если же исходные данные не соответствуют указанным предпосылкам, рекомендуется провести непараметрический дисперсионный анализ или преобразовать (трансформировать) исходные данные, и провести параметрический дисперсионный анализ с преобразованными данными. В этом случае после проведения дисперсионного анализа проводят проверку нулевой гипотезы по критерию Фишера и критерию Стьюдента (*НСР*) для преобразованных средних и результаты оценки автоматически переносят на исходные средние.

Пример 4. В полевом опыте изучали влияние гербицидов на засоренность посевов озимой пшеницы. После применения гербицидов провели количественный учет сорняков, шт/м^2. Результаты опыта представлены в таблице.

Гербициды	Повторность			
	I	II	III	IV
1. Без гербицида	398	414	501	399
2. Аворекс	306	357	380	372
3. Деметра	157	214	135	123
4. Ланцелот	42	36	29	41
5. Деметра + Ланцелот	33	21	11	9

Так как между вариантами опыта наблюдается значительная вариация, что может указывать на неоднордность дисперсий, рекомендуется проверить данные опыта на соответствие предпосылкам дисперсионного анализа по критерию Бартлетта.

Для проведения дисперсионного анализа создадим в программе Statistica новый файл с данными из двух переменных: первая переменная *Гербициды,* в строках которой укажем варианты (наименование гербицидов) и вторая переменная *Число сорняков,* в строках которой впишем числовые значения по сорнякам, подобно тому, как на рис. 5.1.

Запустим модуль **Однофакторный ДА One-way Anova** из меню **Дисперсионный анализ (Anova),** рис. (5.30).

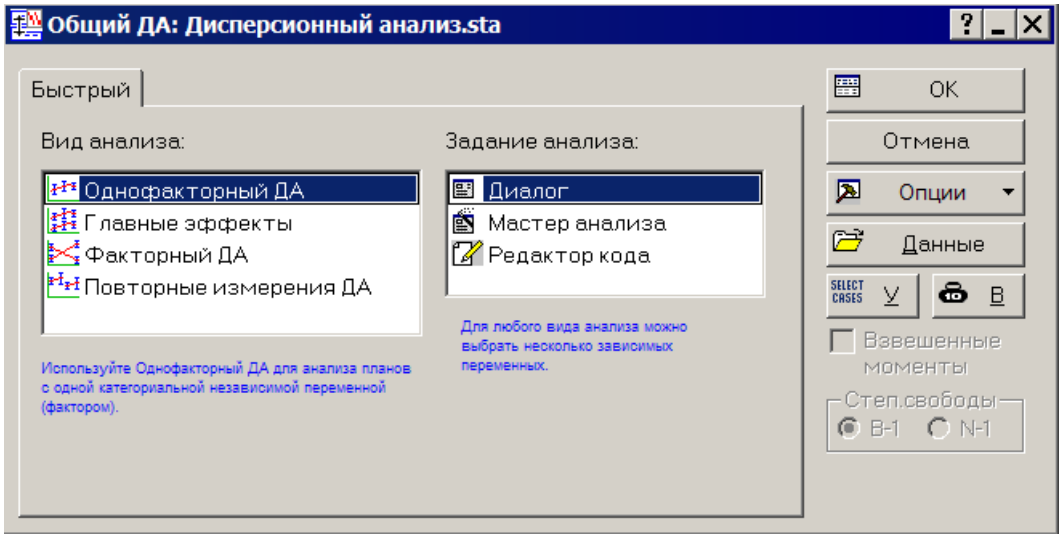


Рис. 5.30. **Диалоговое окно модуля Однофакторный ДА**

В окне выбора переменных выберем в качестве зависимой переменной *Число сорняков* и категориального предиктора – *Гербициды* (рис. 5.31).

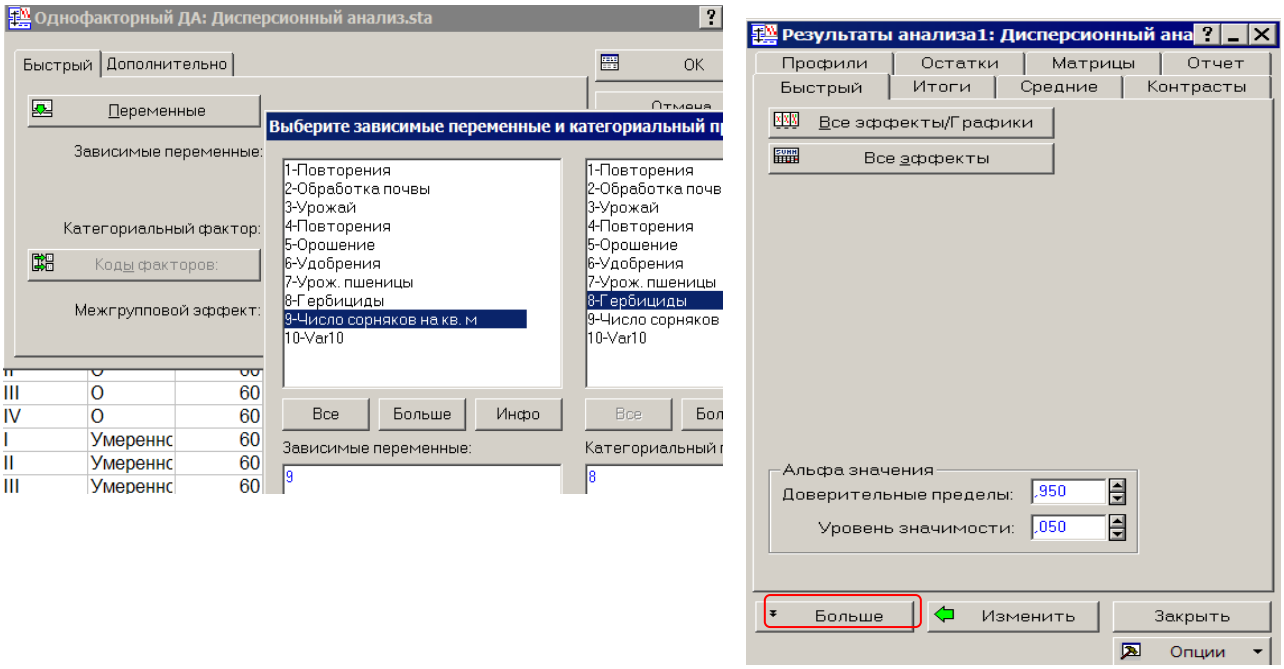


Рис. 5.31. **Диалоговые окна выбора переменных и дополнительных результатов**

В диалоговом сжатом окне выбора результатов нажмем на кнопку **Больше,** появится окно (рис 5.32) для выбора дополнительных результатов дисперсионного анализа, в котором откроем вкладку **Предположения (Assumptions).** Для проверки нулевой гипотезы на однородность дисперсий нажмем на кнопку **Кохрена С, Хартли, Бартлетта**.

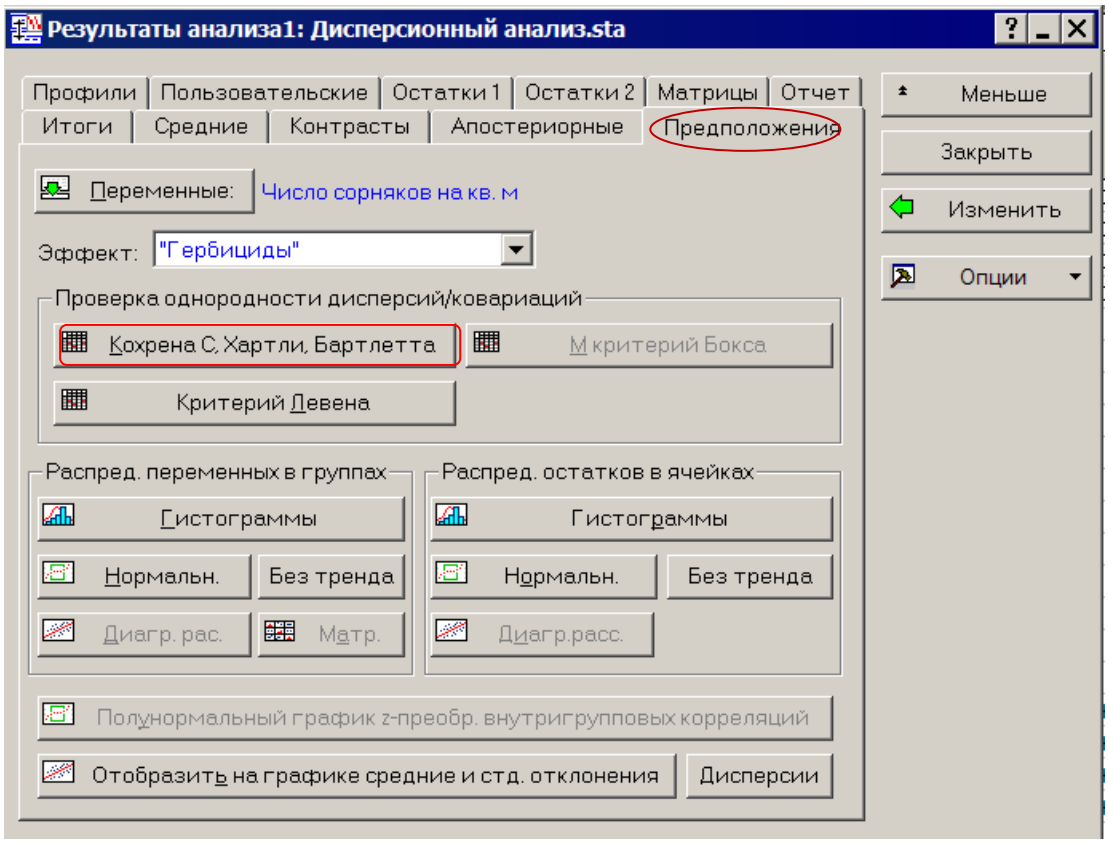


Рис. 5.32. **Диалоговое окно выбора критериев для проверки на однородность исходных данных**

В рабочей книге появится таблица, в которой показаны фактические значения выбранных нами критериев для проверки нулевой гипотезы на однородность дисперсий. Результаты теста на однородность по всем критериям (*p=0,023, p<0,05*; значения критериев выделены красным цветом) однозначно показывают, что дисперсии численности сорняков по вариантам опыта неоднородны – данные не соответствуют предпосылкам дисперсионного анализа , поэтому необходимо провести преобразование исходных данных.

	Критерии однородности дисперсий (Дисперсио Эффект: "Гербициды"				
	Хартли F-макс	Кохрена C	Бартлетт Хи-квад.	ст.св.	p
Число сорняков на кв. м	68,54717	0,453870	11,33686	4	0,023028

Рис. 5.33. **Результаты проверки H_0 на однородность дисперсий**

Проведем преобразование по следующей формуле $x = \sqrt{X}$, где x – преобразованные данные; X – исходные данные (число сорняков). Для

трансформации исходных данных нажмем на панели форматирования на значок **Преобразовать переменные** и в появившемся окне **Преобразование группы переменных** (рис. 5.34) впишем формулу в следуюшей транскрипции: *Преобраз = Sqrt(Сорняки)*. После нажатия на кнопку **Ок** в таблице исходных данных появится новая переменная **Преобраз (Var12)** с автоматически преобразованными (корень квадратный из числа сорняков) данными.

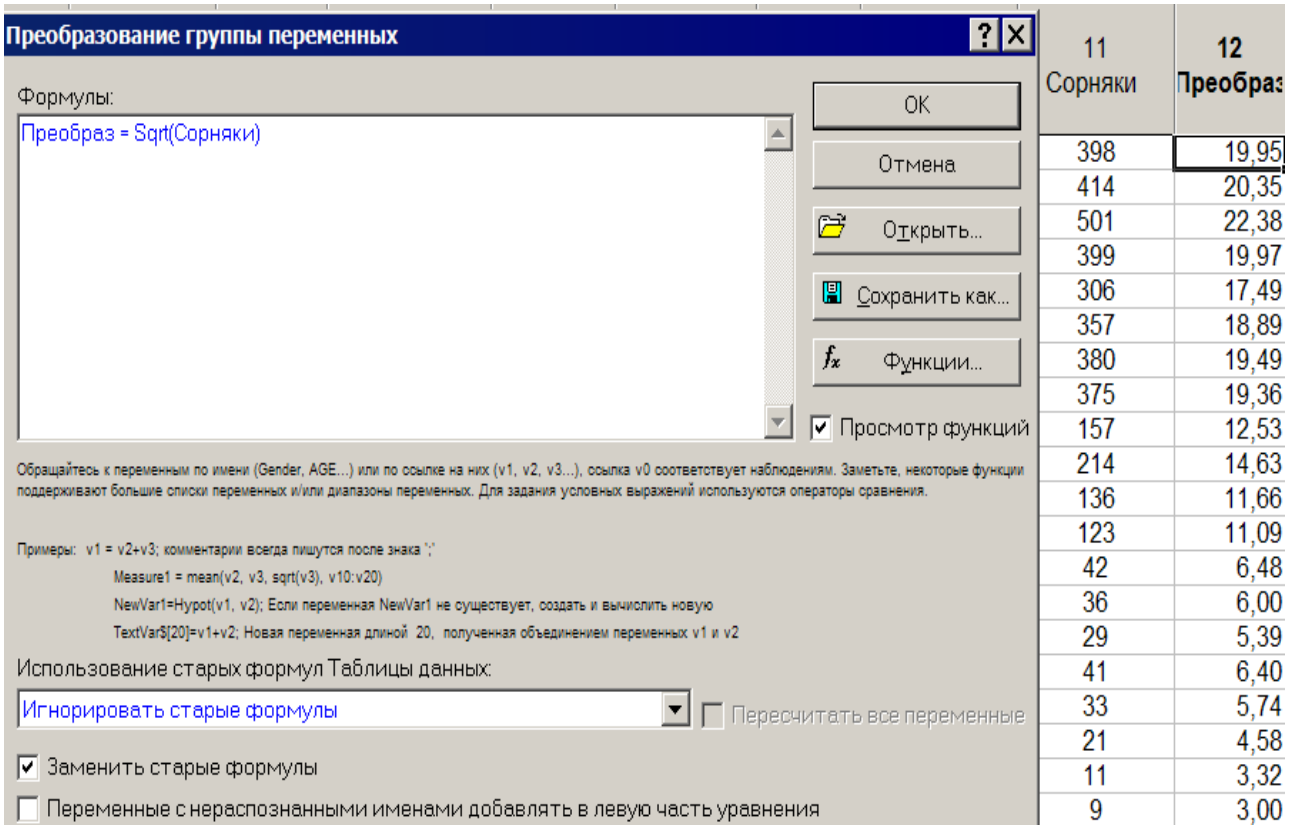


Рис. 5.34. **Рабочее окно для преобразования исходных данных**

После трансформации исходных данных проведем дисперсионный анализ с преобразованными данными (в качестве категориального фактора укажем Гербициды, а зависимой переменной – Var12), (рис. 5.35).

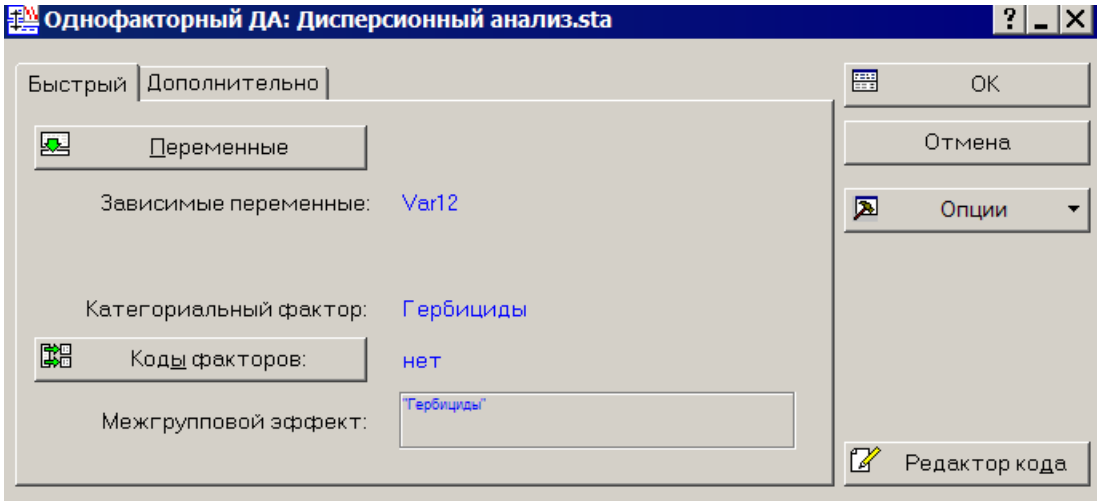


Рис. 5.35. **Диалоговое окно выбора переменных ДА**

В окне с дополнительными результатами дисперсионного анализа (рис. 5.36) откроем вкладку **Предположения (Assumptions)** Для проверки нулевой гипотезы на однородность дисперсий нажмем на кнопку **Кохрена С, Хартли, Бартлетта**

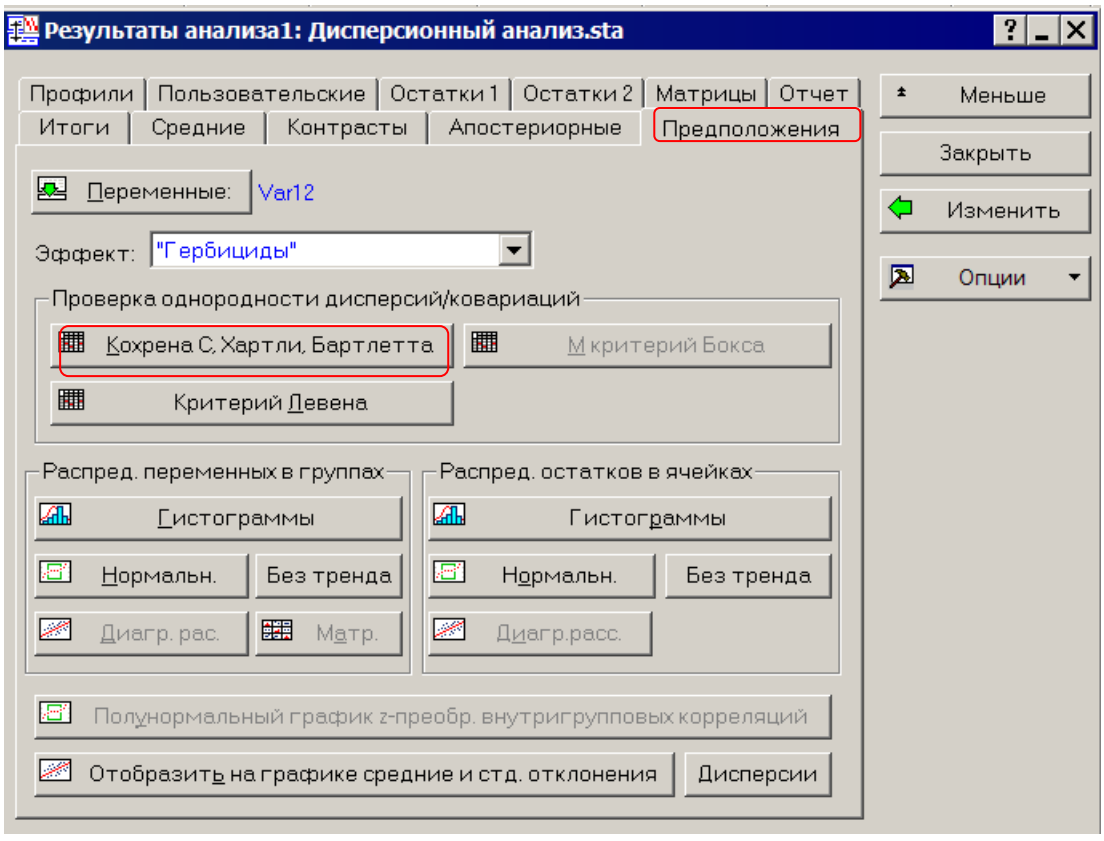


Рис. 5.36. **Диалоговое окно выбора критериев для проверки на однородность преобразованных данных**

В рабочей книге появится новая таблица, (рис. 5.37) в которой показаны фактические значения выбранных нами критериев для проверки нулевой гипотезы на однородность дисперсий.

	Критерии однородности дисперсий (Дисперсион Эффект: "Гербициды"				
	Хартли F-макс	Кохрена C	Бартлетт Хи-квад.	ст.св.	p
Var12	9,580594	0,374517	3,114709	4	0,538816

Рис. 5.37. **Результаты проверки H_0 на однородность дисперсии**

Так как вероятность $p=0,539 > 0,05$ и критерии не окрашены красным цветом, нулевая гипотеза об однородности дисперсий для преобразованных данных принимается, преобразованные данные соответствуют предпосылкам дисперсионного анализа. Теперь можно оценить различия для преобразованных значений в целом по опыту с помощью параметрического критерия Фишера и сравнить попарно варианты между собой по величине *НСР*.

В окне с дополнительными результатами дисперсионного анализа откроем вкладку **Апостериорные** и нажмем кнопку **Фишера НЗР** (аналогично рис. 5.8**).** Появится рабочая книга сравнения средних преобразованных между собой (рис. 5.38), из которой видно, что разности между вторым и первым (*p=0,035*), пятым и четвертым вариантами (*p=0,031*) несущественны на 1% уровне значимости, в то время как разности между другими вариантами существенны. Если проводить сравнение на 5%, то все разности значимы.

	НЗР крит.; перем. Var12 (Дисперсионный анализ) Вероятности для апостер. критериев Ошибка: Межгр. MS = 1,2851, cc = 15,000					
N ячейки	**Гербициды**	{1} 20,664	{2} 18,811	{3} 12,478	{4} 6,0673	{5} 4,1609
1	Без гербицидов		0,035479	0,000000	0,000000	0,000000
2	Аворекс	0,035479		0,000001	0,000000	0,000000
3	Деметра	0,000000	0,000001		0,000001	0,000000
4	Ланцелот	0,000000	0,000000	0,000001		0,031125
5	Деметра+ланцелот	0,000000	0,000000	0,000000	0,031125	

Рис. 5.38. **Таблица сравнения средних преобразованных значений**

После статистической оценки средних преобразованных результаты оценки переносим на средние исходные (количество сорняков, м^2). Для этого

составим таблицу средних преобразованных и средних исходных данных с отклонениями (см. ниже). Одной звездочкой обозначаем существенные разности на 5% уровне значимости, двумя звездочками существенные разности на 1% уровне значимости для преобразованных средних и эти звездочки автоматически переносим на исходные данные.

Агрономический вывод: с вероятностью 95% можно утверждать, что обработка посевов озимой пшеницы гербицидом Аворекс приводит к существенному снижению засоренности ($p<0,05$), а при обработке гербицидами Деметра, Ланцелот и смесью Деметра + Ланцелот наблюдается еще более значимое уменьшение численности сорняков ($p<0,01$).

Таблица 2

Влияние гербцидов на засоренность посевов озимой пшеницы, шт/м^2

Гербициды	Средние преобразованные	Разность, d	Средние исходные	Разность, d
Без гербицидов	20,66	–	428,0	–
Аворекс	18,81	-1,85*	354,5	-73,5*
Деметра	12,48	-8,18**	157,5	-270,5**
Ланцелот	6,07	-14,59**	37,0	-391,0**
Деметра+ланцелот	4,16	-16,5**	18,5	-409,5**

*) существенные различия на 5% уровне значимости

**) существенные различия на 1% уровне значимости

Контрольные вопросы:

1. В чем сущность дисперсионного анализа?
2. Схемы, модели дисперсионного анализа данных агрономических исследований.
3. Критерий Фишера. Проверка нулевой гипотезы с помощью критерия Фишера?
4. Предпосылки дисперсионного анализа.
5. Статистическая обработка данных наблюдений и анализов с неоднородными выборками.
6. Оценка существенности разности средних при дисперсионном анализе.
7. Какие критерии применяются для множественного сравнения средних?
8. Особенности применения критерия Дункана?
9. Апостериорные сравнения в программе Statistica.
10. Как графически оценить различия между средними по вариантам.
11. Как провести дисперсионный анализ данных полевого опыта с рандомизированными методами размещения вариантов в программе Statistica.
12. Дисперсионный анализ данных многофакторного опыта в программе Statistica.
13. Как рассчитать частные и главные эффекты?
14. Дисперсионный анализ данных наблюдений и анализов с неоднородными выборками в программе Statistica.
15. Как проводится сравнение исходных данных с неоднородными выборками?
16. Дисперсионный анализ данных с многосборовыми культурами и данных многолетних экспериментов.

Глава 6. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ ДАННЫХ АГРОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ

Корреляционно-регрессионный анализ – статистический метод анализа, данных, который дает возможность:

– дать качественную оценку наличия силы или тесноты, направления и формы зависимости изучаемых признаков (корреляционный анализ);

– провести количественную оценку изучаемой зависимости: нахождение уравнения регрессии и построения теоретической линии регрессии (регрессионный анализ).

Все изучаемые признаки делятся на две группы:

– *независимые или факторные признаки* (аргумент), которые обозначаются буквой ***X,***

– *зависимые или результативные* (функция), которые изменяются под влиянием независимых признаков. Эти признаки обозначаются буквой ***Y.***

Корреляции подразделяют по направлению, форме и числу связей. По направлению корреляция может быть прямой и обратной. При ***прямой корреляции*** с увеличением значения признака ***X*** увеличивается значение признака *Y*. *Например, чем выше продуктивная кустистость, тем выше урожайность, чем больше питательных элементов в почве, тем выше урожайность, чем больше длина листа, тем больше его площадь: чем лучше освещенность растений, тем интенсивнее фотосинтез и т.п.*

При ***обратной корреляции*** с увеличением значения признака *X* значение признака *Y* уменьшается. *Например, при увеличении засоренности полей или пораженности культурных растений уменьшается урожай, при постоянном увеличении массы корнеплодов свеклы уменьшается их сахаристость и т.п.*

По форме корреляция бывает прямолинейной и криволинейной. При ***прямолинейной (линейной)*** связи между признаками ***X*** и ***Y*** зависимость носит линейный характер и выражается уравнением прямой линии $\boldsymbol{Y = a + bX.}$ При

линейной зависимости одинаковые приращения аргумента ***X*** приводят к одинаковым приращениям функции ***Y***.

Когда при одинаковых приращениях аргумента ***X*** функция имеет неодинаковые изменения ***Y***, зависимость называется ***нелинейной*** или ***криволинейной.***

В зависимости от числа изучаемых признаков корреляция может быть ***простой***, если имеется связь между двумя признаками и ***множественной***, когда изучается зависимость между тремя и более признаками.

Количественно связь между признаками описывается уравнением регрессии: при простой регрессии связь кратко обозначается ***Y = f(X),*** а при множественной ***Y = (X, Z, T …).***

6.1 Прямолинейная корреляция и регрессия

Оценка прямолинейной корреляционной зависимости проводится по **коэффициенту корреляции**

$$r = \frac{\sum (X - \bar{x}) \cdot (Y - \bar{y})}{\sqrt{\sum (X - \bar{x})^2 \cdot \sum (Y - \bar{y})^2}}$$

По величине коэффициента корреляции можно судить о направлении и силе или тесноте связи. Коэффициент корреляции величина безразмерная, он изменяется в интервале от ***–1 ÷ +1***. Знак коэффициента корреляции указывает на направление связи, если «минус», то связь обратная, если «плюс», то положительная.

Тесноту или силу связи между изучаемыми признаками можно определить по величине коэффициента корреляции. При полных зависимостях, когда корреляционная связь превращается в функциональную, значение коэффициента корреляции равно для положительных связей +1, для обратных связей –1. Если же ***r*** принимает значение около *0*, то это дает основание говорить об отсутствии связи между ***Y*** и ***X***.

Значимость коэффициента корреляции (нулевая гипотеза об отсутствии связи между факторным и результативным признаками H_0: $r = 0$) проверяется на основе t-критерия Стьюдента или вероятности значимости (p). Для проверки H_0 следует рассчитать t-статистику ($t_ф$) и сравнить ее с табличным значением (t_m) по заданному уровню значимости (α) и числу степеней свободы ($d.f.$) или рассчитать (p). Если $t_ф > t_m$, $p<0,05$ то гипотеза H_0 отвергается с вероятностью 95%. Это свидетельствует о значимости линейного коэффициента корреляции и статистической существенности зависимости между факторным и результативным признаками.

Квадрат коэффициента корреляции называют **коэффициентом детерминации (r^2)**. Коэффициент детерминации показывает долю общей дисперсии результативного признака (Y), которая объясняется вариацией факторного признака (X).

Этапы корреляционного анализа:

- выявление наличия взаимосвязи между признаками;
- определение формы связи;
- определение силы (тесноты связи);
- определение значимости корреляционной зависимости.

В процессе проведения регрессионного анализа рассчитывают коэффициент регрессии (b_{yx}). Коэффициент регрессии b_{yx} показывает, в каком направлении и насколько изменится результативный признак Y при изменении независимого (факторного) признака X на единицу измерения и выражается в единицах Y. Коэффициент регрессии имеет знак коэффициента корреляции и может принимать любые значения, он привязан к единицам измерения обоих признаков, знак коэффициента регрессии указывает на направление связи.

Полный регрессионный анализ включает следующие этапы:

Спецификация:

- определение вида функции, описывающей связь между результативным признаком и факторными признаками;

- выбор модели регрессии: линейные, нелинейные; однофакторные модели (парная модель регрессии) и многофакторные модели (модель множественной регрессии).

Идентификация:

- определение коэффициентов регрессии;
- определение параметров, входящих в модель регрессии;
- расчет теоретических значений результативного признака для отдельных наборов значений факторов;
- исследование отклонений расчетных значений от эмпирических данных.

Верификация:

- оценка качества полученной модели и проверка соответствующих статистических гипотез о регрессии;
- анализ остатков.

Прямолинейная зависимость описывается уравнением парной линейной регрессии: $\mathbf{Y = bX + a}$, где Y – результирующий признак, X – факторный признак, b – коэффициент регрессии, a – свободный член.

Оценка качества модели проводится на основе гипотез о значимости модели в целом и каждого ее параметра и анализе остатков. Оценка *значимости уравнения в целом* проводится по ***F- критерию***, которому предшествует дисперсионный анализ (ANOVA) регрессии. Для каждого значения F можно вычислить соответствующую вероятность (p). Если $F_{фак} \geq F$ табл., значение вероятности (p) меньше принятого *уровня значимости p* или вероятности ошибки регрессия признается значимой, уравнение в целом значимо и корректно описывает проверяемую зависимость.

Оценка *значимости параметров регрессии* проводится по *t -критерию Стьюдента*.

Анализ остатков позволяет получить представление, насколько хорошо подобрана сама модель и насколько правильно выбран метод оценки коэффициентов. Анализ остатков проводится по графику вероятности.

Пример 1. В опыте изучали зависимость между массой зерна (***X***) и содержанием жира (***Y***)

Масса зерна, г, X	11	19,9	15,9	16,3	10,2	21,4	15,8	21,6	12,3	17,3
Содержание жира,%, Y	1,2	5,1	2,3	3,1	0,9	4,1	2,1	4,2	1,1	3,4

Для проведения корреляционно-регрессионного анализа создадим в программе Statistica файл **Корреляция** и занесем наши данные в переменных *Масса зерна* и *Содержание жира.* Если исследователя интересует только корреляция без дальнейшего регрессионного анализа, можно воспользоваться опцией **Парные и частные корреляции**. Для проведения простой парной корреляционной зависимости в меню **Анализ (Statistics)** выберем модуль **Основные статистики и таблицы** и в появившейся стартовой панели **Основные статистики и таблицы** выберем опцию **Парные и частные корреляции (Paired and Partial Correlations**) (рис. 6.1).

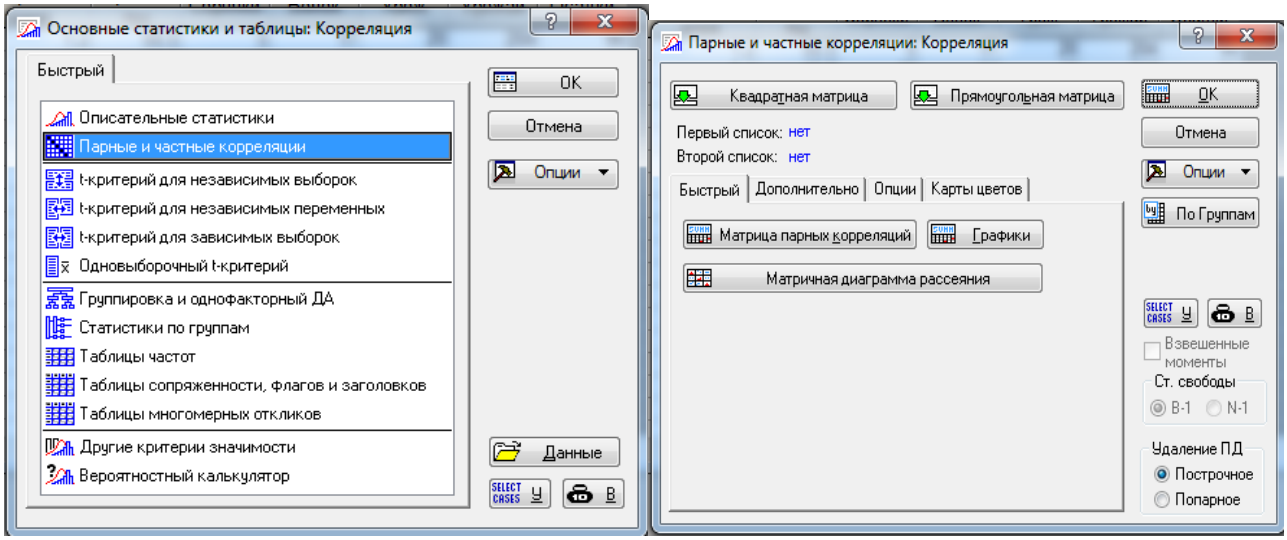


Рис. 6.1. **Стартовая панель и диалоговое окно *Парные и частные корреляции***

В появившемся диалоговом окне в поле списков вводим анализируемые переменные: *Масса зерна и Содержание жира*, причем для нахождения коэффициента корреляции не имеет значение, какая из переменных является зависимым признаком, а какая факторным.

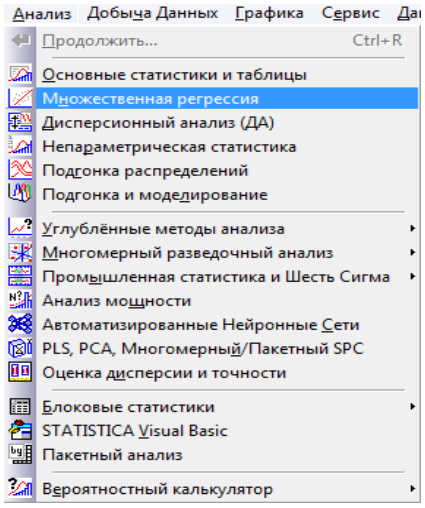
После нажатия на клавишу **Ок** в рабочей книге получим таблицу с результатами корреляционного анализа, где для нас наибольший интерес

представляет коэффициент корреляции между содержанием жира и массой зерна ***r = 0,93911***.

Переменная	Корреляции (Корреляция) Отмеченные корреляции значимы на уровне p <,05000 N=10 (Построчное удаление ПД) **Средние**	Ст.откл.	Масса зерна	Содержание жира
Масса зерна	16,17000	4,077050	1,000000	0,939111
Содержание жира	2,75000	1,459262	0,939111	1,000000

На основании проведенного анализа мы можем сделать вывод о том, что между содержанием жира и массой зерна ячменя установлена прямая и очень тесная корреляционная зависимость, причем корреляция значима с вероятностью 95%, так как коэффициент корреляции отмечен красным цветом и ***p<0,05***.

Итак, проведя анализ в этой опции, мы получили информацию о качественной оценке связи между содержанием жира и массой зерна. Однако, чтобы установить какова количественная зависимость между данными признаками, представленных результатов явно недостаточно. Для проведения полноценного корреляционно-регрессионного анализа необходимы другие опции. Такие опции представлены в меню **Анализ (Statistics)**, например: **Множественная регрессия (Multiple Regression).**



После выбора указанного модуля в появившемся диалоговом окне (рис. 6.2) нажмем на вкладку **Дополнительно (Advanced)** и поставим галочку напротив *Показать описат. статист*. Щелкнем по кнопке **Переменные**

(Variablets) и в качестве зависимой переменной укажем – *Содержание жира* (номер переменной – 7) и независимой или факторной переменной – *Масса зерна* (номер переменной – 6).

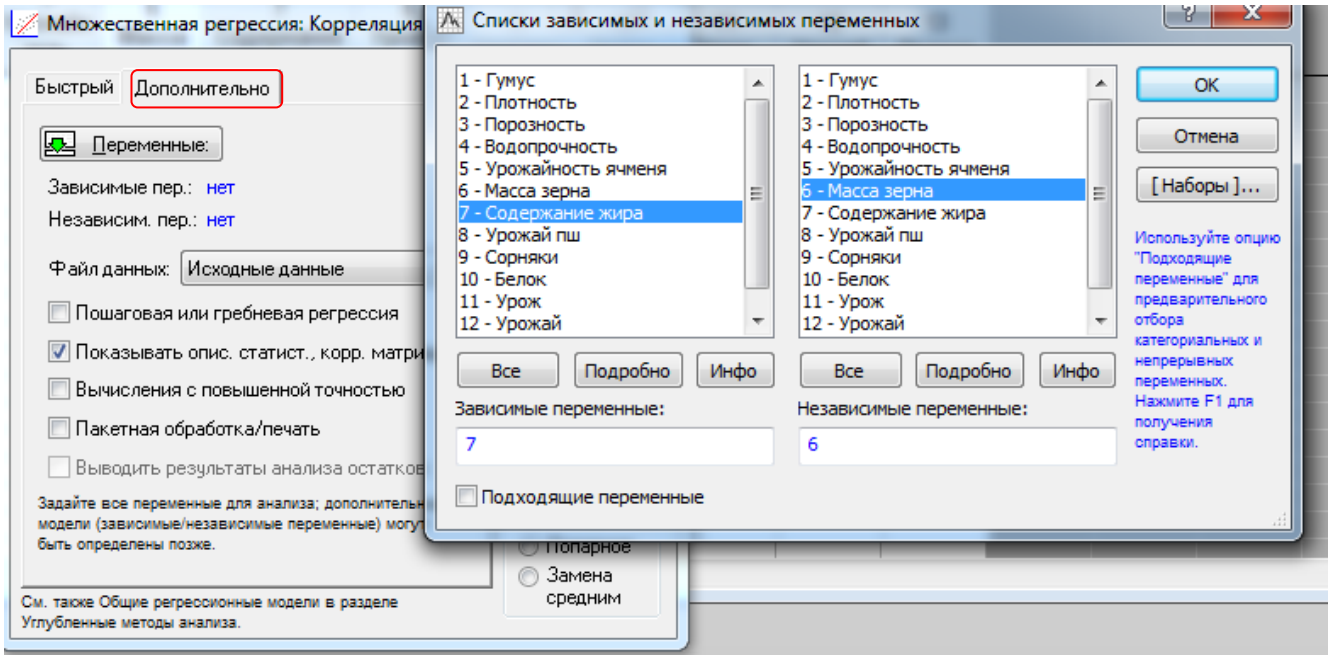
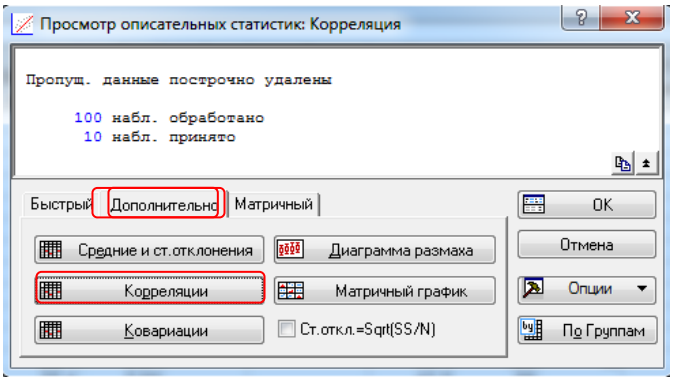


Рис. 6.2. **Диалоговое окно выбора зависимой и независимых переменных**

Для вывода результатов корреляционного анализа активируем вкладку **Дополнительно (Advanced)**, нажмем на кнопку **Корреляции (Correlation)** и в рабочей книге получим значение коэффициента корреляции (***r=0,939111***).



Переменная	Корреляции (Корреляция) Масса зерна	Содержание жира
Масса зерна	1,000000	0,939111
Содержание жира	0,939111	1,000000

Рис. 6.3. **Диалоговое окно вывода показателей корреляции и регрессии**

Для вывода результатов регрессии нажмем на вкладку **Быстрый (Quik)** и кнопку **Ок** (рис. 6.3). Программа произведет вычисления и на экране

появится расширенное диалоговое окно с результатами регрессии (рис. 6.4). Верхняя часть окна является информационной, здесь приводятся наиболее важные параметры корреляционно-регрессионного анализа: множественный коэффициент корреляции (***R=0,93911***), который для нашего примера с двумя переменными и есть коэффициент парной корреляции (***r***). Коэффициент детерминации $R^2 = 0,8819$ показывает, что в изменении содержания жира в зернах ячменя 88, 2% обусловлено влиянием массы зерна. О значимости коэффициента корреляции свидетельствует ***критерий Стьюдента*** $t_ф=3,714$, который больше $t_{05}=2,31$ *(df=8)*. На значимость коэффициента корреляции указывает и очень малое значение вероятности ***p =0,0059, p<0,05.***

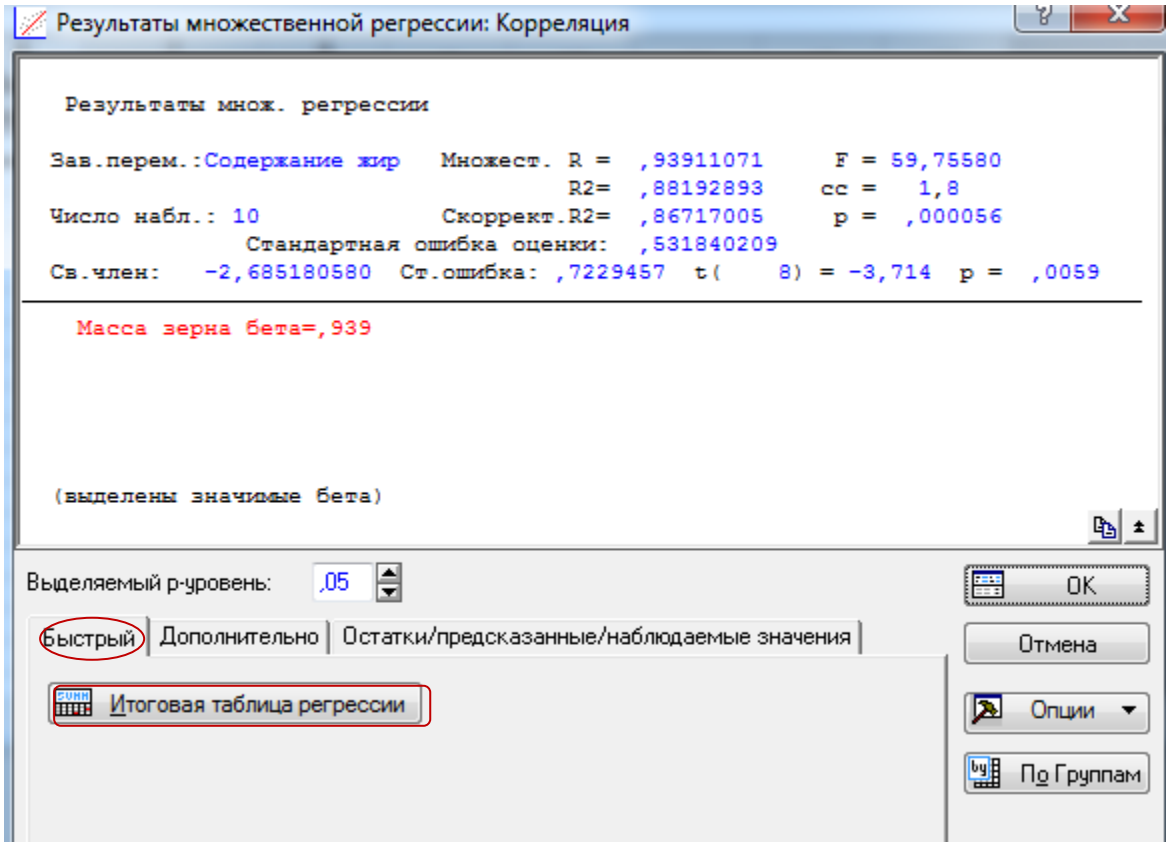


Рис. 6.4. **Расширенное диалоговое окно результатов корреляции и регрессии**

Нижняя часть содержит функциональные кнопки, позволяющие просмотреть результаты регрессионного анализа более детально. Для вывода результатов регрессии нажмем на кнопку **Итоговая таблица регрессии (Summary: regression results)** и получим таблицу итогов регрессии для зависимой переменной.

N=10	Итоги регрессии для зависимой переменной: Содержание жира R= ,93911071 R2= ,88192893 Скоррект. R2= ,86717005 F(1,8)=59,756 p<,00006 Станд. ошибка оценки: ,53184 БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(8)	p-знач.
Св.член			-2,68518	0,722946	-3,71422	0,005920
Масса зерна	0,939111	0,121486	0,33613	0,043482	7,73019	0,000056

В итоговой таблице представлены: коэффициент Бета (для парной корреляции он равен коэффициенту корреляции – ***r***), параметры для уравнения прямолинейной регрессии: коэффициент регрессии b_{yx}***= 0,33613*** и величина свободного члена ***a= –2,68518***. На основании проведенного регрессионного анализа получено следующее уравнение прямолинейной зависимости ***Y=0,34 – 2,68X***. О значимости рассчитанных параметров уравнения регрессии свидетельствуют низкие значения вероятности (***p***), представленные в последнем столбце таблицы. В верхней части таблицы указано фактическое значение критерия Фишера, по которому можно судить об адекватности регрессионной модели в целом. Согласно критерию Фишера ($F_{фак}$ ***=59,756,*** F_{05}***(df=1, 8) =5,32 и p<0,0006)*** полученное уравнение регрессии высоко значимо. Анализ значимости параметров уравнения регрессии с помощью *t*-критерия показал, что параметры уравнения ***a= –2,68518*** (свободный член) и b_{yx}***= 0,33613*** (коэффициент регрессии) с вероятностью 99% достоверны, поэтому мы имеем право использовать представленное уравнение с рассчитанными коэффициентами при *0,01%* уровне значимости.

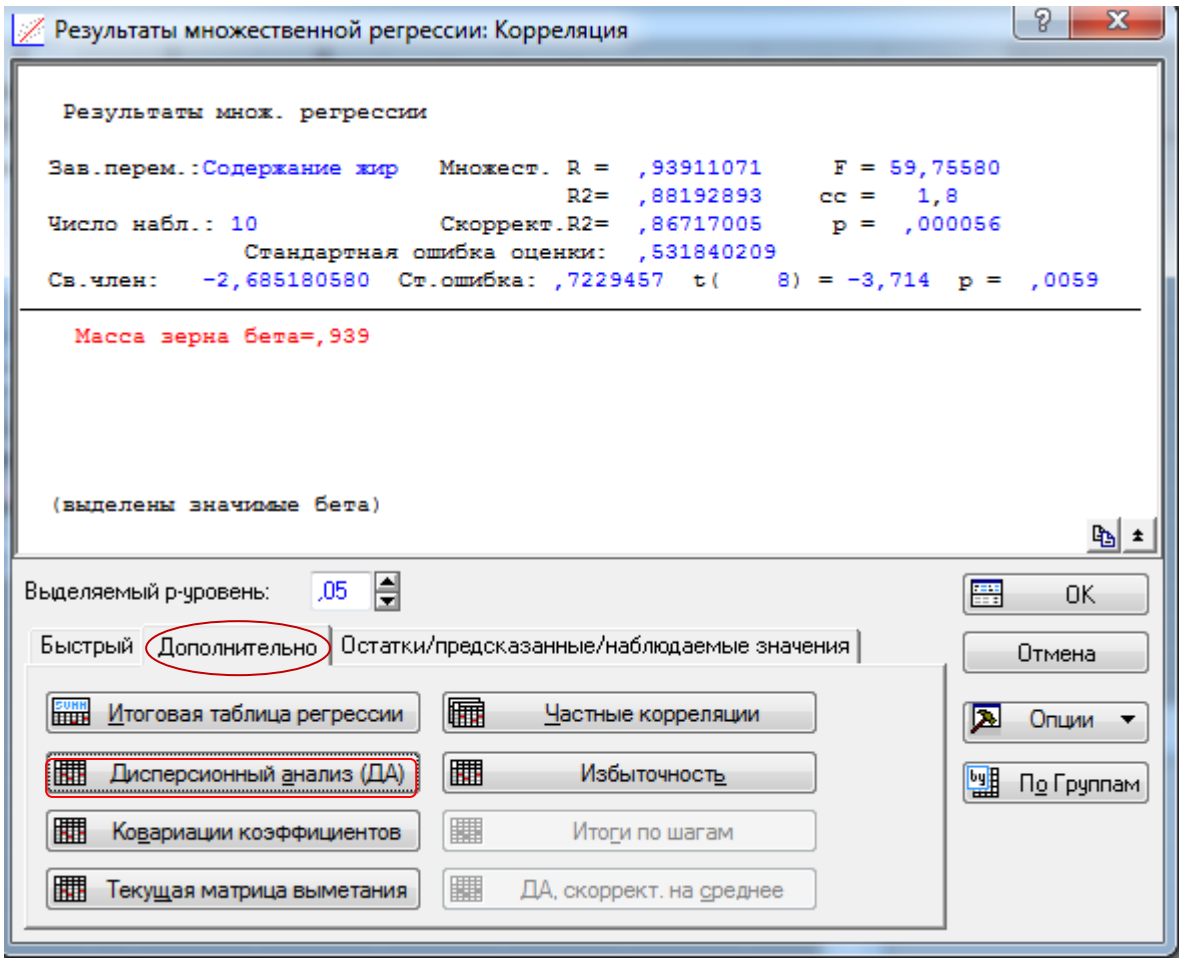


Рис. 6.5. **Расширенное диалоговое окно результатов корреляции и регрессии**

Кнопка **Дисперсионный анализ (Analysis of variance)** регрессии при активной вкладке **Дополнительно** (рис. 6.5) позволяет более детально ознакомиться с результатами дисперсионного анализа уравнения регрессии. В строках таблицы дисперсионного анализа регрессии из общей суммы вариации **Итого (Total),** равной **19,165** на регрессию **Регрессия (Regress)** приходится 16,903 (88,3%) и 2,263 на остаточную или случайную вариацию **Остатки (Residual).** ***F*** - критерий полученного уравнения регрессии значим на 5% уровне. Вероятность нулевой гипотезы (*p-level*) ***p= 0,000056*** значительно меньше *0,05*, что говорит об общей значимости предлагаемой модели регрессии.

Эффект	Дисперсионный анализ; ЗП: Содержание жира (Корреляция)				
	Сумма квадр.	сс	Средн. квадр.	F	р-знач.
Регресс.	16,90217	1	16,90217	59,75580	0,000056
Остатки	2,26283	8	0,28285		
Итого	19,16500				

Для построения точечного графика и теоретической линии регрессии активируем вкладку **Остатки/предсказанные/наблюдаемые значения (Residuals/assumptions/prediction)** и нажмем на кнопку **Анализ остатков (Perform residual analysis)** (Рис.6.6).

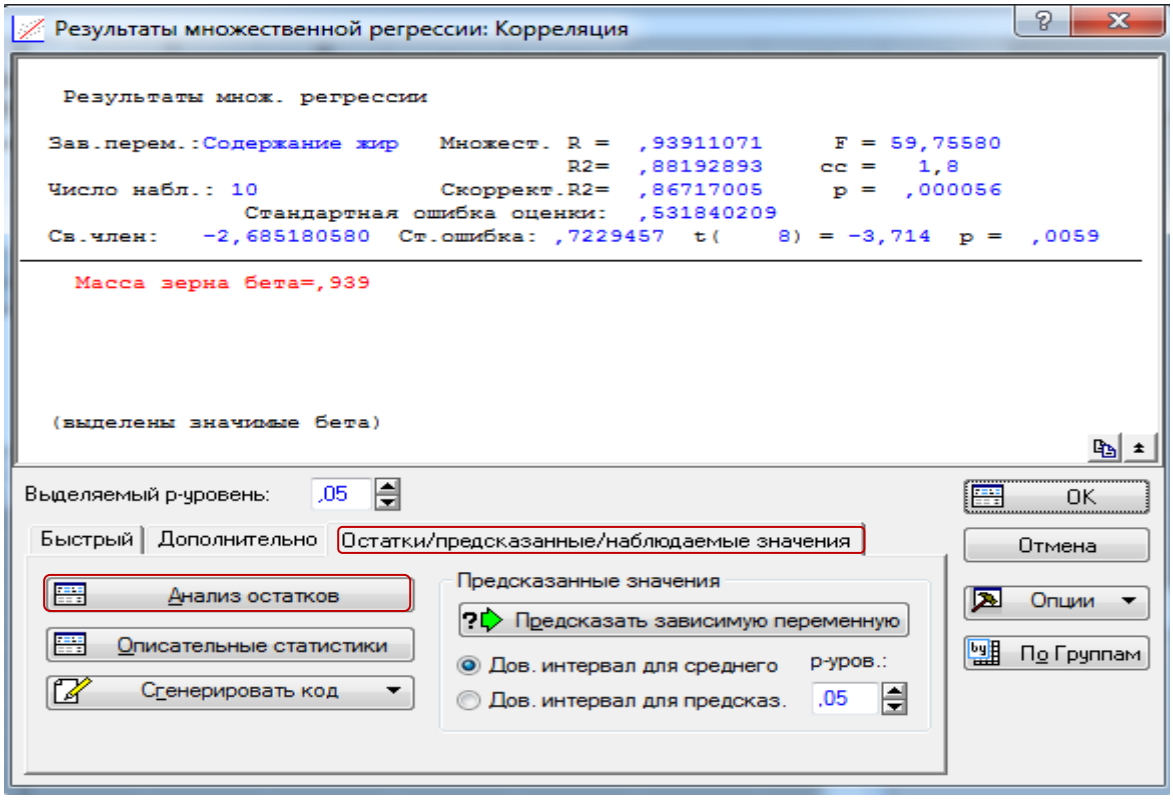


Рис. 6.6. **Расширенное диалоговое окно результатов корреляции и регрессии**

После нажатия на кнопку **Ок** появится диалоговое окно (рис. 6.7), в котором активируем вкладку **Диаграммы рассеяния** и нажмем на кнопку **Две переменные.** В появившемся окне выберем две переменные для построения диаграммы рассеяния: на оси ***X*** размещаем независимую переменную – *Масса зерна*, а на оси ***Y*** – *Содержание жира*.

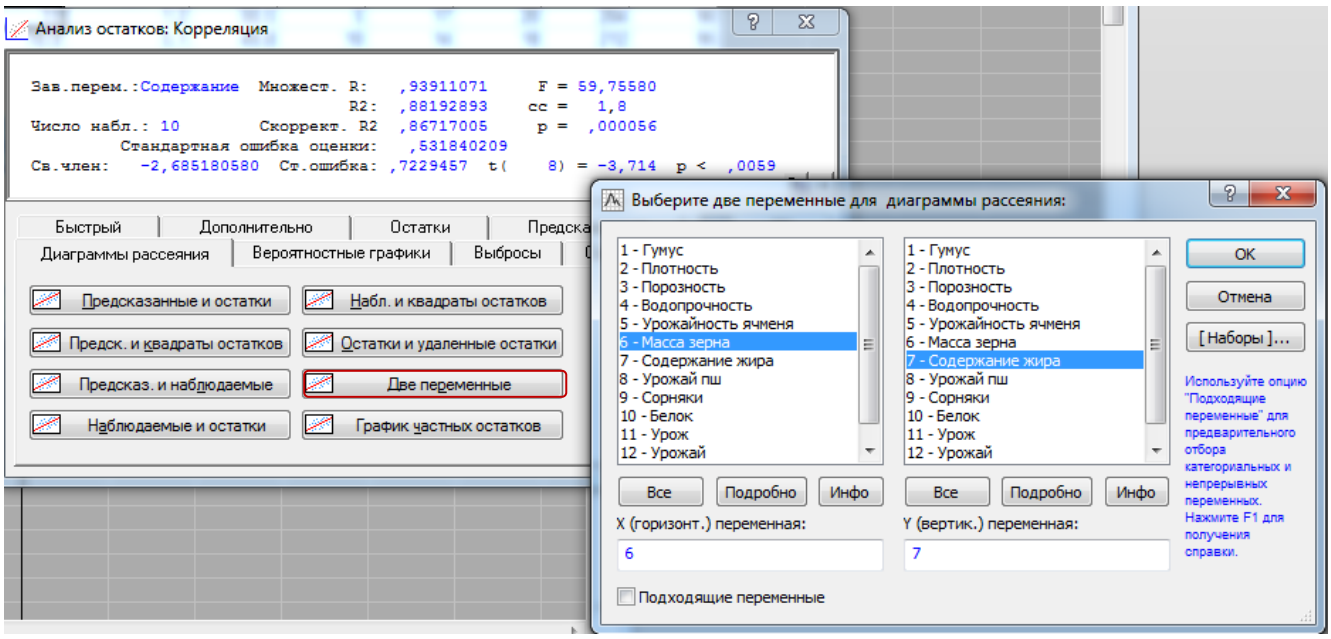


Рис. 6.7. **Диалоговое окно выбора переменных для построения графика**

После нажатия на кнопку **Ок** в рабочей книге получаем следующий график (рис. 6.8)

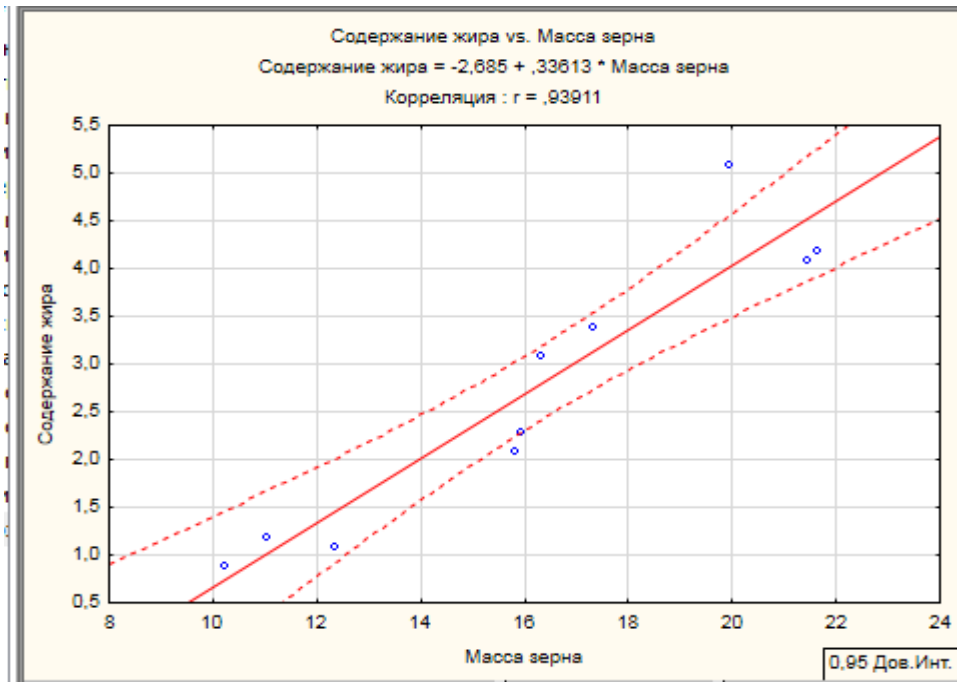


Рис. 6.8. **Корреляционное поле и теоретическая линия линейной регрессии**

На графике представлена теоретическая линия прямолинейной регрессии, пунктирными линиями показан диапазон 95% доверительного интервала линии регрессии. В верхней части графика уравнение регрессии ***Y= –2,685 + 0,33613X*** и коэффициент корреляции ***r=0,94***

Линия регрессии выражает наилучшее предсказание зависимой переменной (***Y***) по независимым переменным (***X***). Однако отмечается разброс наблюдаемых точек относительно теоретически рассчитанной прямой линии. Отклонение отдельной точки от линии регрессии (от предсказанного значения) называется *остатком*. Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем, очевидно, лучше прогноз. Анализом остатков завершается корреляционно-регрессионный анализ. Анализ остатков проводится по графику вероятности.

Для построения нормального вероятностного графика остатков выберем вкладку **Вероятностные графики (Probability plots)** и нажмем кнопку **Нормальный график остатков (Normal plot of residuals)** (рис. 6.9).

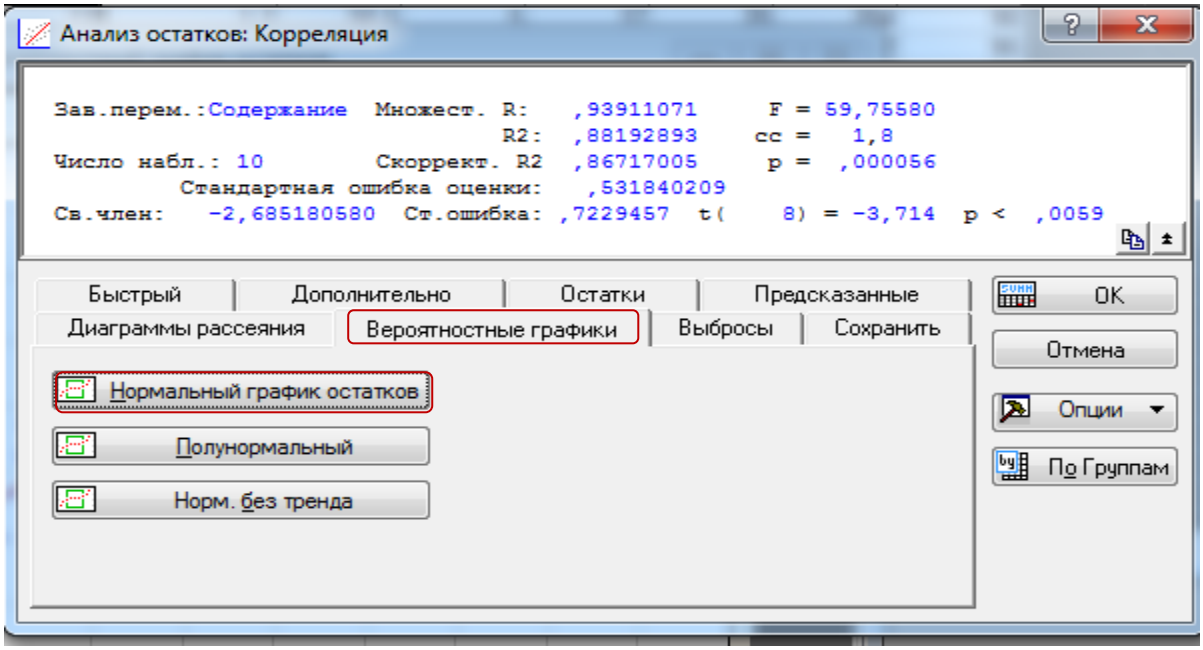


Рис. 6.9. **Диалоговое окно для выбора вероятностных графиков**

После нажатия на кнопку **Ок** в рабочей книге получаем нормальный вероятностный график остатков.

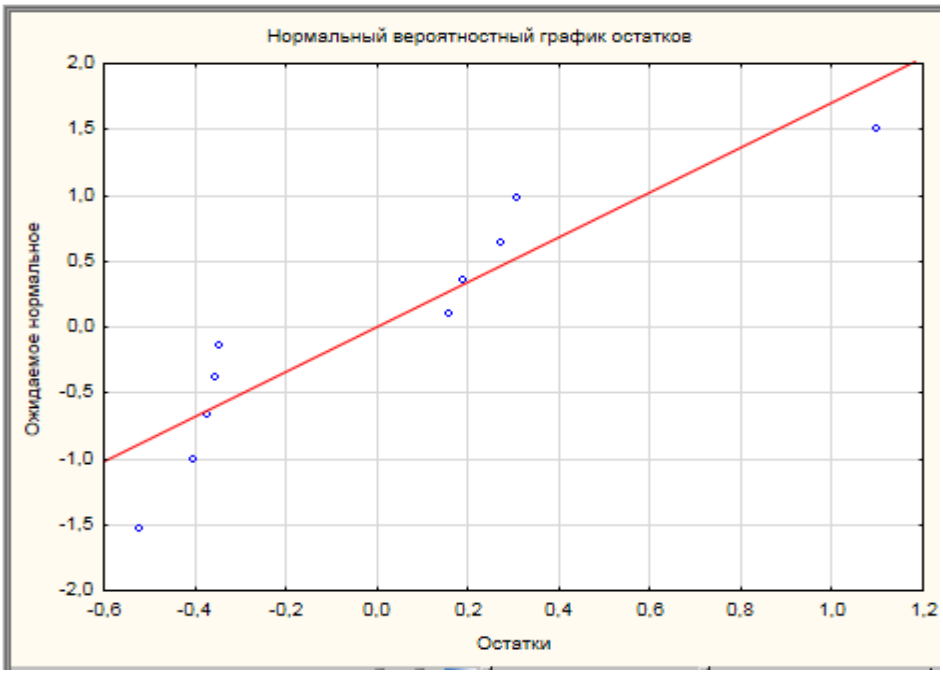
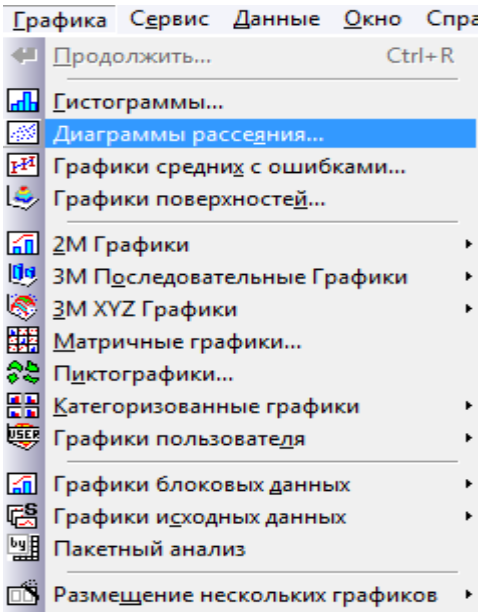


Рис. 6.9.1. **Нормальный вероятностный график**

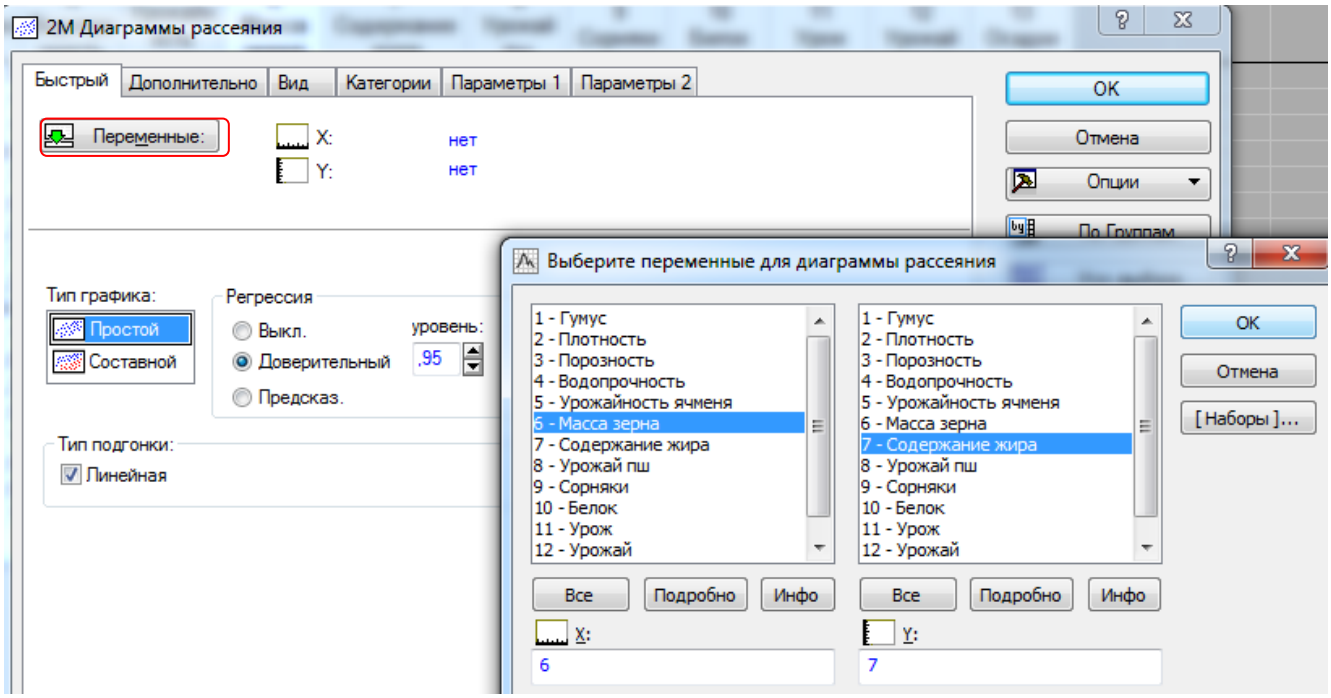
Ввиду того, что разброс точек вокруг прямой незначителен, представленный вероятностный график остатков визуально показывает, что методы оценки коэффициентов корреляции и регрессии и рассчитанное уравнение в целом, верно, отражают зависимость между содержанием жира и массой зерна ячменя.

Проведение регрессионного анализа в модуле Графика

Ускоренно, без детального анализа регрессионный анализ с построением графика можно провести в меню **Графика (Grafics).** Для чего в меню **Графика** выберем модуль **Диаграмма рассеяния (Scatterplots)**



В открывшемся диалоговом окне (рис. 6.10) отмечаем тип графика – **Простой**, тип подгонки – **Линейный (Linear),** Регрессия – **Доверительный уровень 0,95.** Нажмем на кнопку **Переменные** и в окне выбора переменных укажем по оси **X** – *Масса зерна*, а по оси **Y** – *Содержание жира*.



6.10. **Диалоговое окно выбора типа графика и переменных для графика рассеяния**

После нажатия на кнопку **Ок** попадаем в диалоговое окно (рис. 6.11) для выбора параметров представления результатов анализа. По умолчанию оставляем Тип графика – **Простой**, Подгонка – **Линейная**, галочками выберем нужные статистики: *R-квадрат*, корреляция и *p-уровень*, уравнение регрессии и нажмем на клавишу **Ок.**

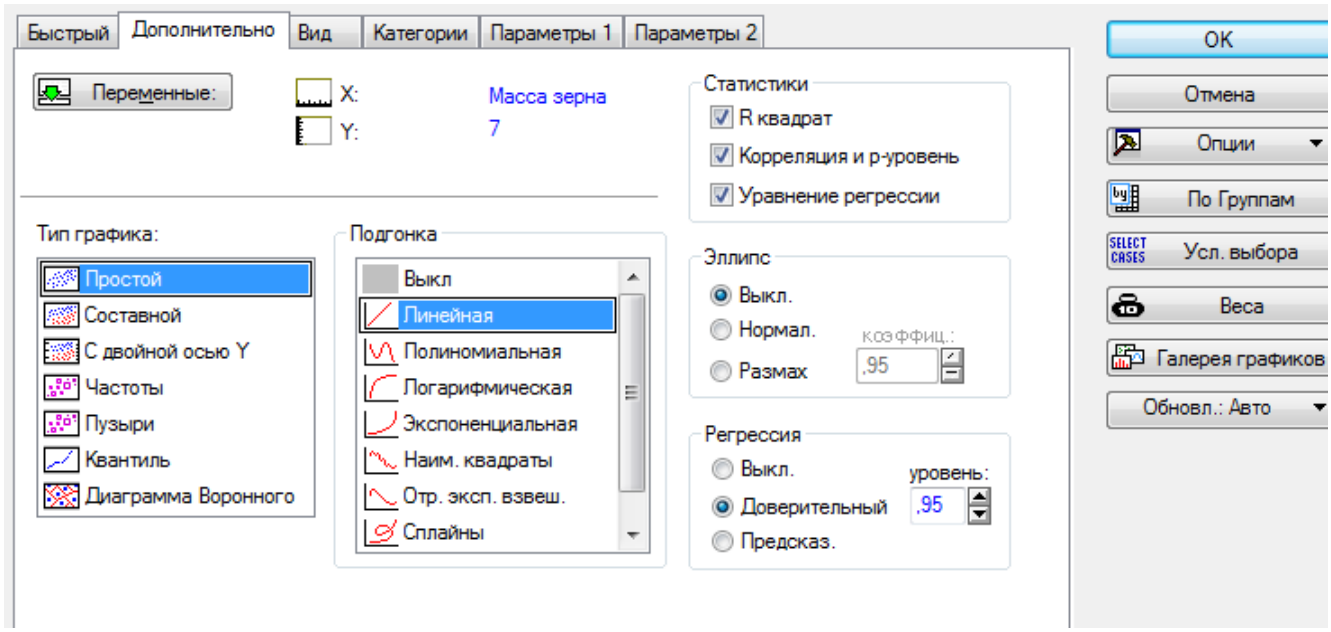


Рис. 6.11. **Диалоговое окно выбора параметров регрессии**

В рабочей книге появляется график регрессионной зависимости между содержанием жира и массой зерна ячменя, в верхней части выведено уравнение регрессии ***Y=-2,68 +0,34X,*** где ***Y*** – содержание жира, ***X***- масса зерна. Пунктирными линиями выделены границы 95% доверительной зоны линии регрессии (рис. 6. 12).

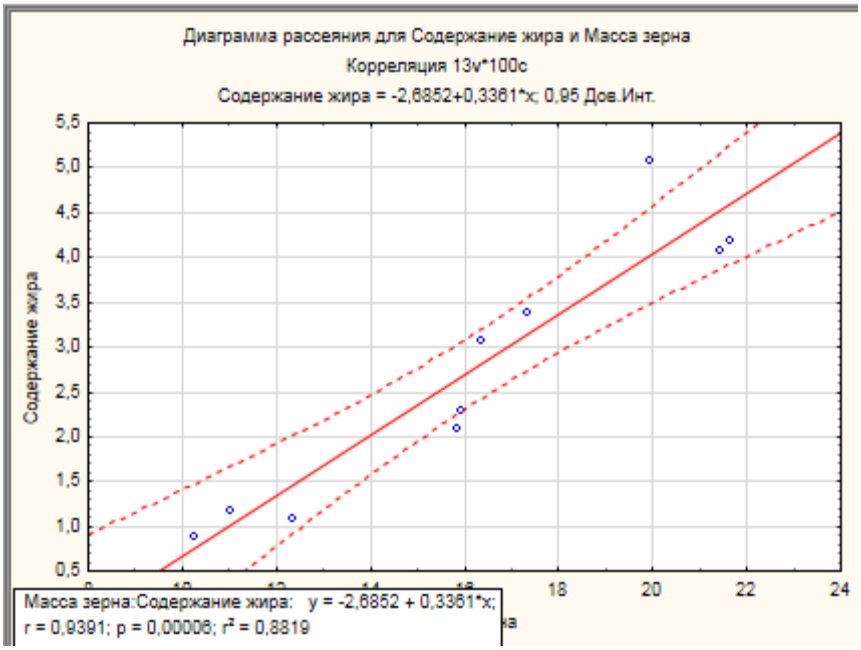


Рис. 6.12. **Точечный график и теоретическая линия регрессия**

В нижней части графика представлены коэффициент корреляции (***r=0,9391***), коэффициент детерминации (***d_{yx}=r^2=0,8819≈88,2%***) и вероятность значимости (***p=0,00006***). На основании указанных коэффициентов можно сделать вывод о том, что между содержанием жира и массой зерна установлена тесная, значимая на *0,00006* уровне вероятности корреляционная зависимость, 88,2% изменений в содержании жира в зерне ячменя обусловлено изменением массы зерна.

6.2 Нелинейная (криволинейная) корреляция и регрессия

Использование коэффициента прямолинейной зависимости предполагает выполнение следующих условий:

- значение обоих анализируемых признаков распределено нормально;
- связь между признаками является линейной.

Если же не выполняется одно из этих условий, применение коэффициента Пирсона (*r*) приводит к ложным выводам о корреляционной зависимости.

Пример 2. Изучали зависимость между урожайностью (*X*) и содержанием белка в зерне озимой пшеницы (*Y,%*).

X	20	18	17	17	20	18	24	18	23	18	24	23
У	17	14	13	12	17	15	13	13	15	14	12	14

Сформируем файл исходных данных с двумя переменными *Урож и Белок* с данными по урожайности и содержания белка. Проведем корреляционный анализ данного примера в программе Statistica с использованием опции **Парные корреляции,** как было показано на рис. 6.1.

Корреляции (Корреляция)
Отмеченные корреляции значимы на уровне p <,05000
N=12 (Построчное удаление ПД)

Переменная	Средние	Ст.откл.	Белок	Урож
Белок	14,08333	1,676486	1,000000	-0,000000
Урож	20,00000	2,763397	-0,000000	1,000000

В итоге в рабочей книге получаем нулевой коэффициент прямолинейной корреляции, что свидетельствует об отсутствии какой-либо связи между содержанием белка и урожайностью озимой пшеницы. Но, если мы построим точечный график (рис.6.13), то с большой долей вероятности можем предположить, что между содержанием белка и урожайностью озимой пшеницы наблюдается определенная корреляционная зависимость, и она по форме криволинейна. Графическое изображение изучаемой зависимости является очень удобным визуальным способом оценки адекватности регрессионной модели.

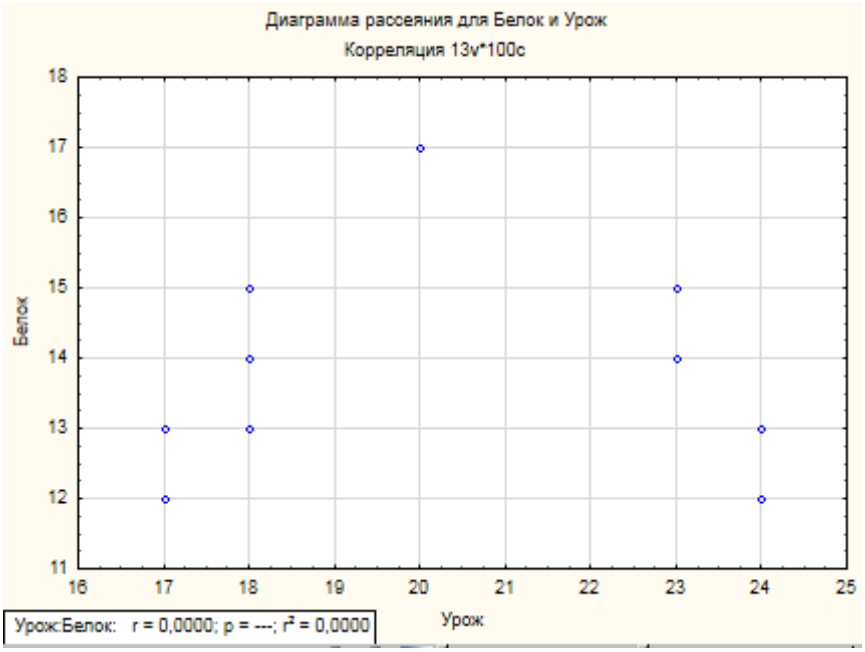


Рис. 6.13. **Точечный график зависимости содержания белка от урожайности**

Построение графиков является быстрым и наглядным методом получения информации о характере и типе зависимости между изучаемыми признаками. В принципе во всех случаях, когда перед исследователем стоит задача проведения корреляционно-регрессионного анализа, прежде, чем приступать к определению параметров корреляции и регрессии будет полезным сначала построить точечный график (диаграмма рассеяния) и по корреляционному полю сделать предварительные выводы о характере зависимости между изучаемыми признаками и на основании визуальной оценки выбрать те или иные модели регрессионного анализа, предлагаемые программой Statstica.

Так как связь между содержанием белка и урожайностью носит нелинейный характер, что видно из представленной диаграммы, мы не можем использовать линейную регрессию, поэтому для аппроксимации экспериментальных данных необходимо опираться на параметры криволинейной регрессии. Основная задача заключается в том, чтобы подобрать вид нелинейной функции и определить численные значения неизвестных параметров выбранной функции.

Поиск наилучшей регрессионной модели представляет собой довольно громоздкий процесс. Для оценки нелинейной зависимости программа Statistica

предлагает несколько методов. Рассмотрим два из них: модуль **Общие регрессионные модели (General regression models)** и модуль **Множественная нелинейная регрессия (Multy-nonlinear).**

6.2.1 Регрессионный анализ в модуле Общие регрессионные модели

Для проведения регрессионного анализа с использованием первого модуля в меню **Анализ** выберем модуль **Углубленные методы анализа**, в открывающемся меню второго порядка выберем **Общие регрессионные модели** (рис. 6.14).

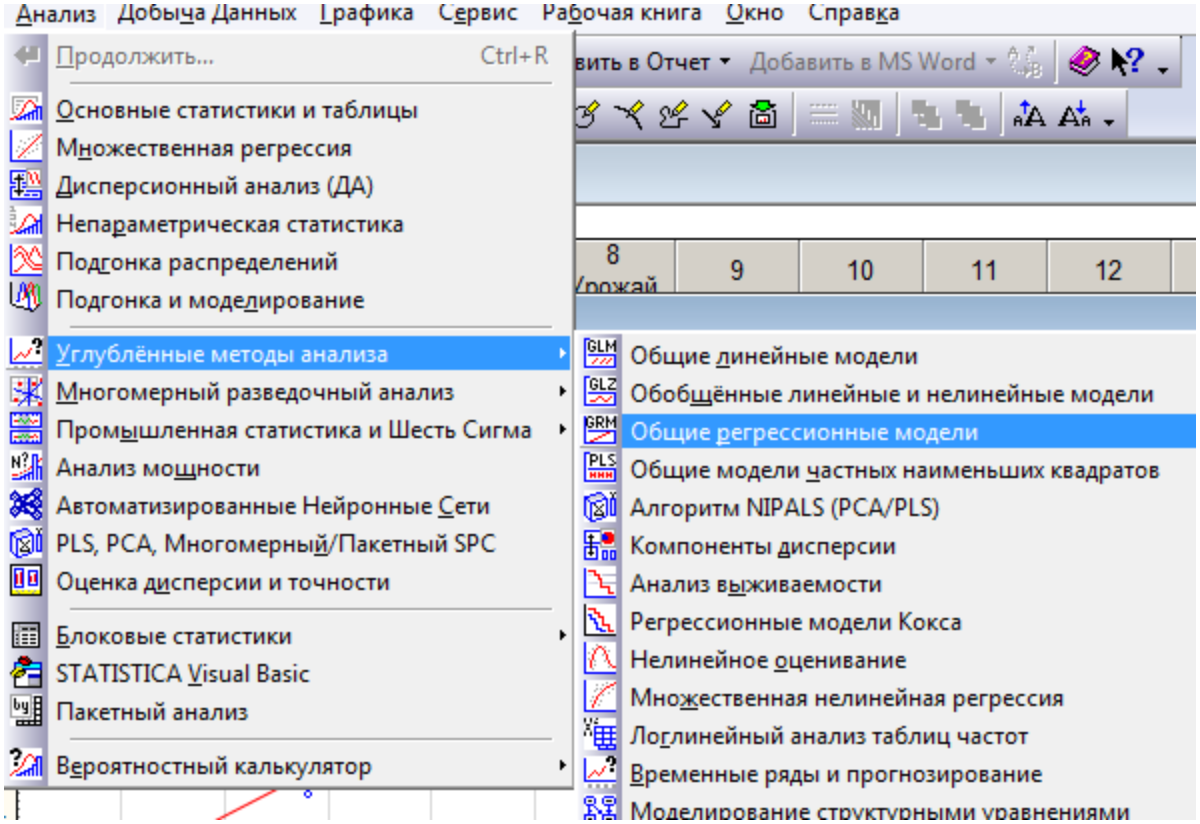
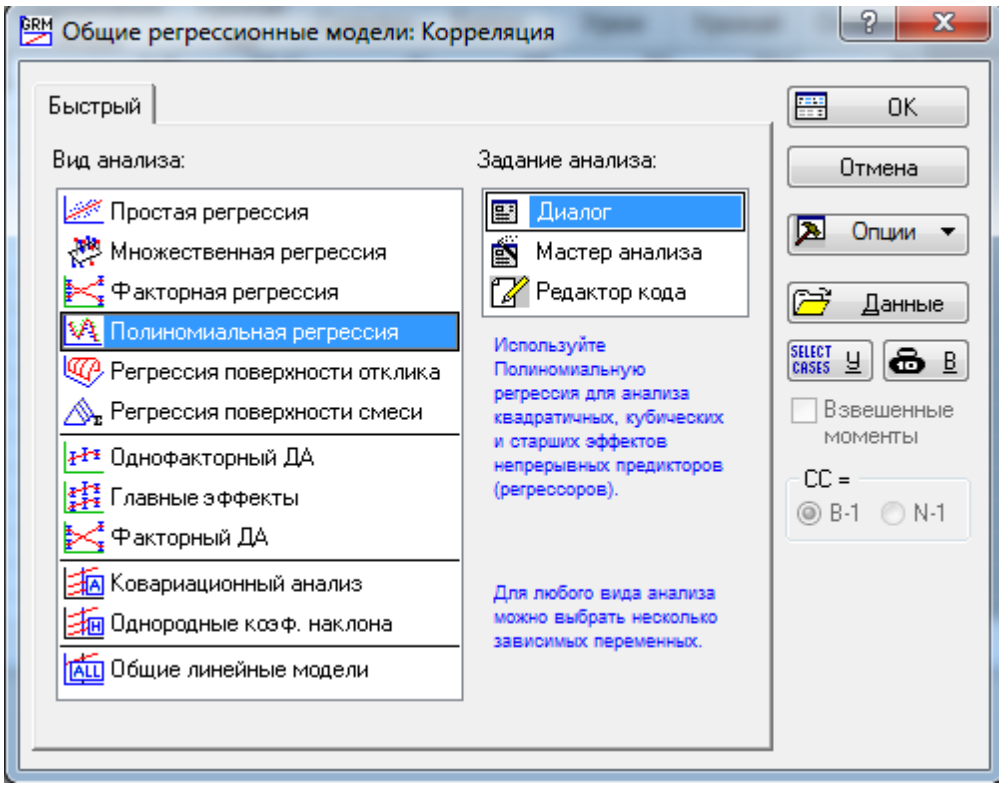
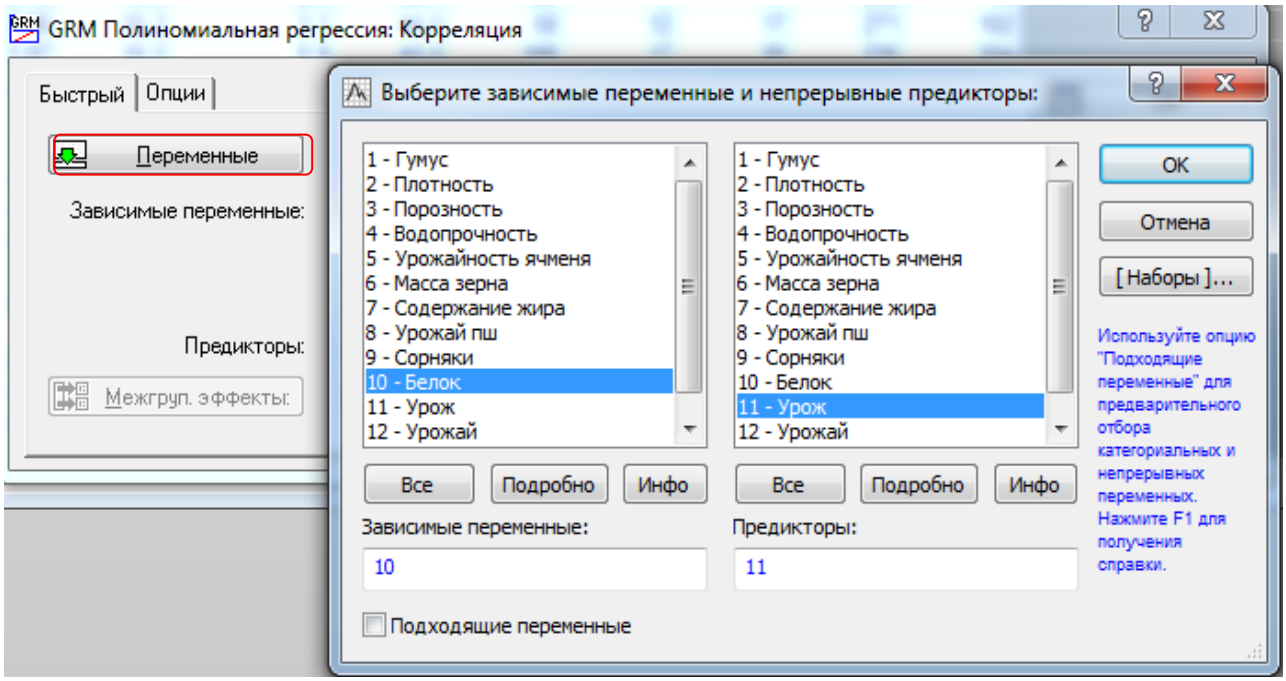


Рис. 6.14. **Выбор модуля для нелинейной регрессии**

Открывается диалоговое окно **Общие регрессионные модели.** Так как на основании визуальной оценки мы предполагаем, что анализируемая зависимость вероятнее всего описывается полиномом, выберем опцию **Полиноминальная регрессия (Polynomial regression),** (рис. 6.15).



6.15. **Диалоговое окно выбора полиноминальной регрессии**



6.16. **Диалоговое окно выбора переменных для регрессионного анализа**

В окне выбора переменных (рис. 6.16) в качестве Зависимой переменной (Y) введем *Белок,* предикторы (X) – *Урожай*, после нажатия на клавишу Ок,

получим панель для выбора результатов корреляционно-регрессионного анализа.

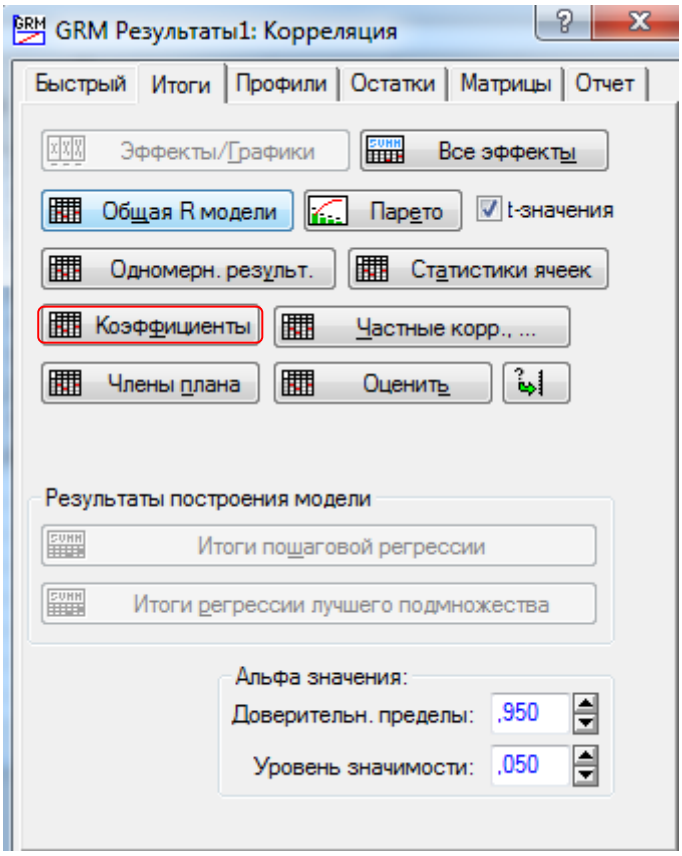


Рис. 6.17. **Панель результатов корреляционно-регрессионного анализа**

Следующий шаг – это нахождение параметров уравнения регрессии нелинейной зависимости, в нашем случае – уравнение полинома. В панели результатов (рис. 6.17) выберем кнопку **Коэффициенты** и в итоге получим таблицу с параметрами уравнения регрессии, критерии Стьюдента для этих параметров и вероятности значимости этих параметров.

Эффект	Оценки параметров (Корреляция) Сигма-ограниченная параметризация					
	Белок Парам.	Белок Ст. Ош.	Белок t	Белок p	-95,00% Дов.инт	+95,00% Дов.инт
Св.член	-135,989	20,90046	-6,50653	0,000111	-183,270	-88,7093
Урож	14,847	2,06342	7,19556	0,000051	10,180	19,5152
Урож^2	-0,361	0,05012	-7,20068	0,000051	-0,474	-0,2475

На основании проведенных расчетов зависимость между содержанием белка и урожайностью озимой пшеницы описывается уравнением полинома (парабола) $\boldsymbol{Y=14,85X - 0,36X^2 - 135}$

Все параметры криволинейной регрессионной зависимости значимы, так как значения ***p*** максимально приближены к нулю**,** и они выделены красным цветом ***(p<0,05)***, поэтому предлагаемая модель нелинейной связи наиболее точно описывает зависимость между урожайностью и содержанием белка в зерне.

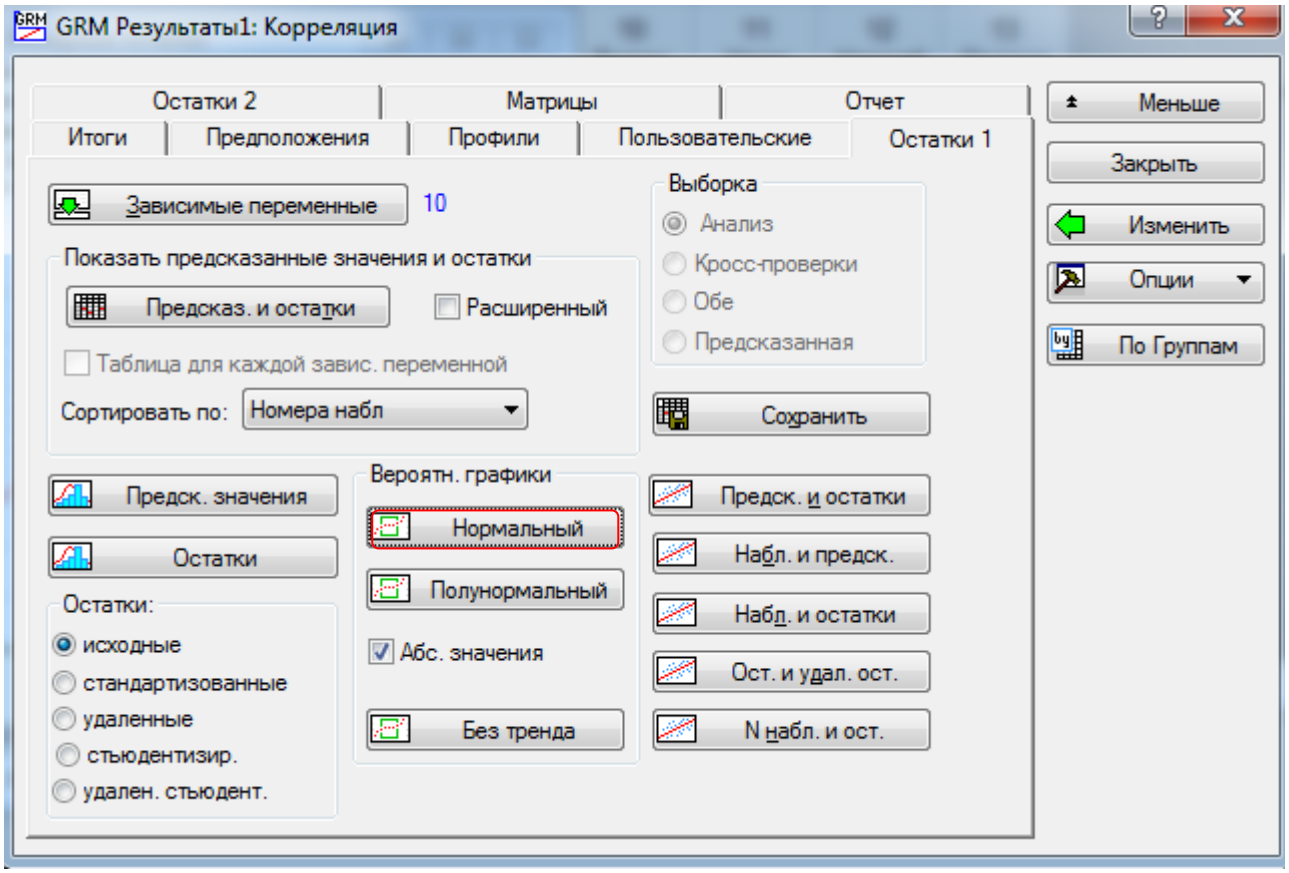


Рис. 6.18. **Диалоговое окно выбора нормального вероятностного графика**

Для оценки приведенного уравнения регрессии проведем анализ остатков. В панели вывода результатов активируем вкладку **Остатки (Residual),** получаем новую панель результатов, в которой нажмем на кнопку **Нормальный (Normal)** (рис. 6.18) и в итоге получаем нормальный вероятностный график (рис. 6.19). Очень близкое расположение точек вокруг прямой линии свидетельствует об адекватности рассчитанной модели по полиному.

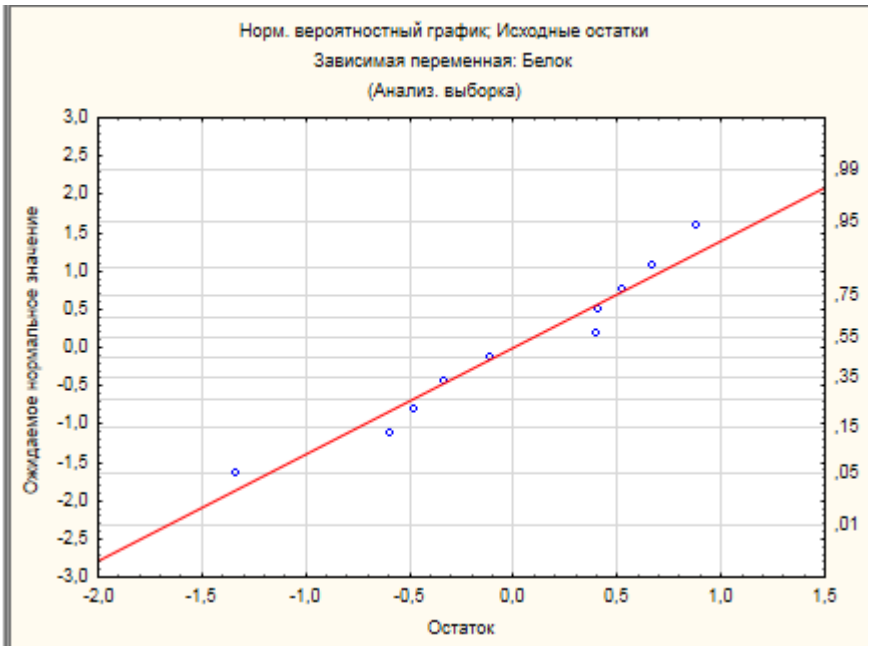


Рис. 6.19. **Нормальный вероятностный график**

6.2.2. Подбор кривых и уравнения регрессии для нелинейных связей

Для построения полиноминальной линии заходим в меню **График** и выберем модуль **2M Диаграмма рассеяния**, в котором укажем переменные для построения графика как показано на рис. 6.20.

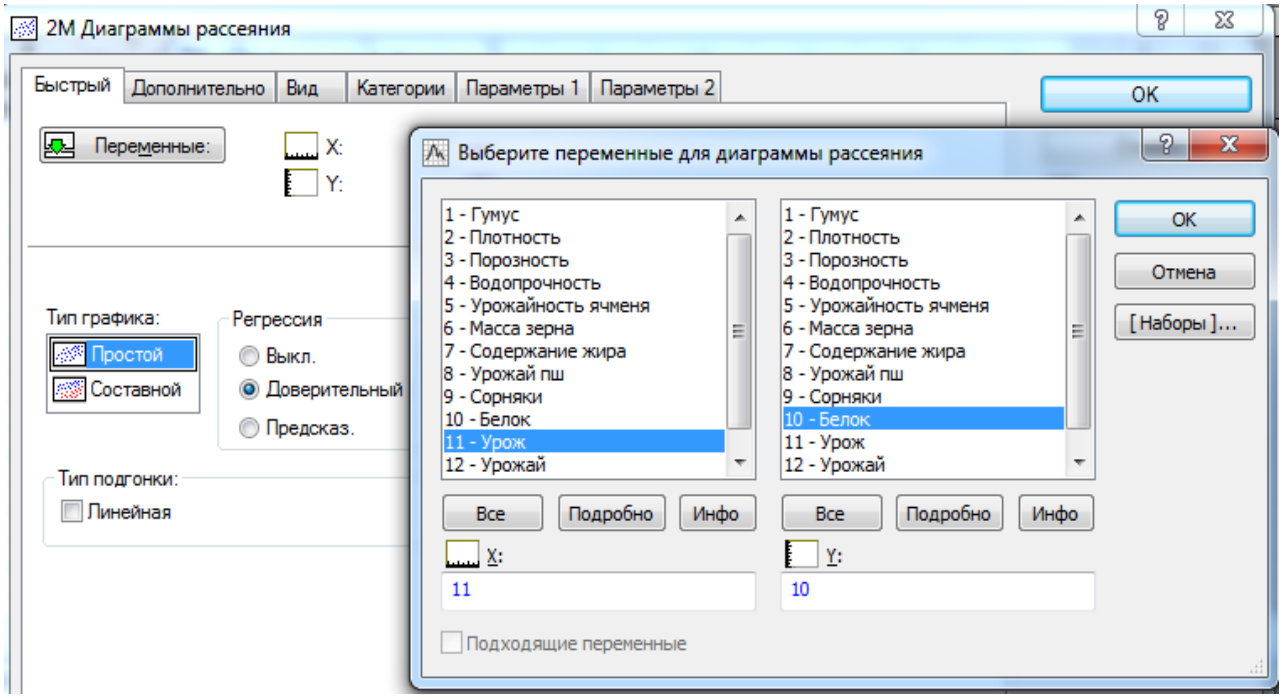


Рис. 6.20. **Диалоговое окно выбора переменных для диаграммы рассеяния**

В появившемся окне (рис. 6.21) выберем **Простой (Simple)** тип графика, в графе **Подгонка (Fit)**, отметим **Полиноминальная (Polynomial)** для вывода на

графике показателей **Статистики** галочками укажем **корреляцию, р-уровень, уравнение регрессии** и нажмем на кнопку **Ок.**

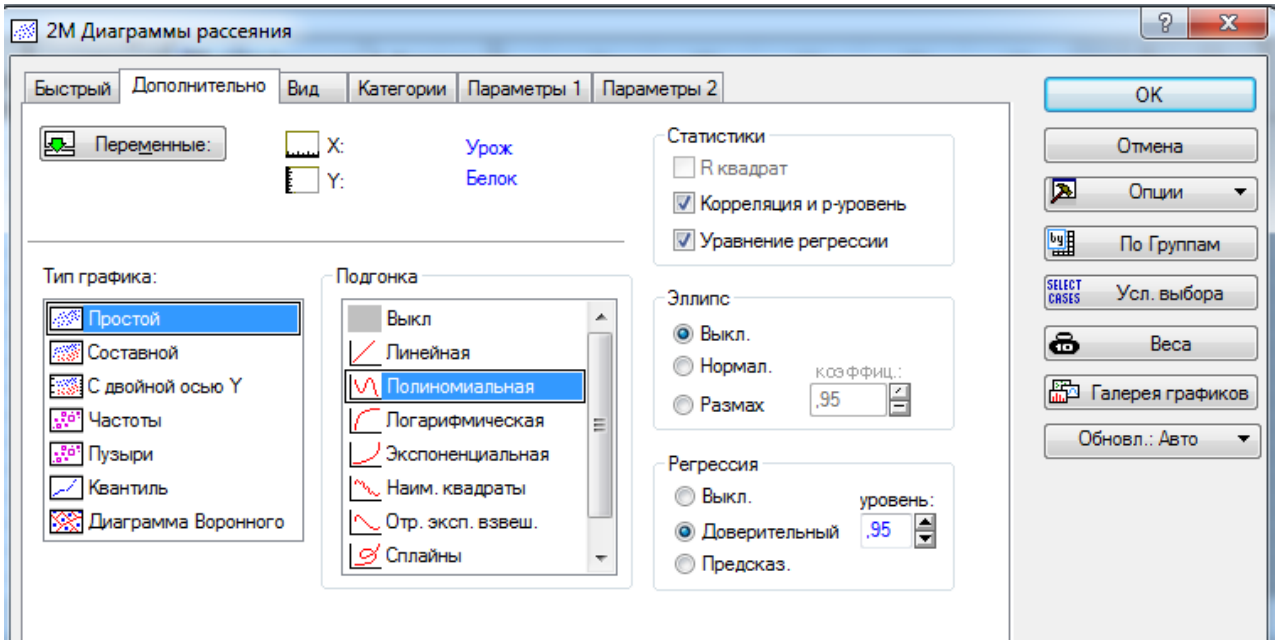


Рис. 6.21. **Диалоговое окно выбора нелинейных кривых**

В рабочей книге (рис. 6.22) выводится график полиноминальной зависимости содержания белка и урожайности зерна озимой пшеницы с фактическими точками, теоретической линией регрессии с 95% доверительной зоной. На графике указано уравнение полинома $\boldsymbol{Y=14,85X - 0,36X^2 - 136}$

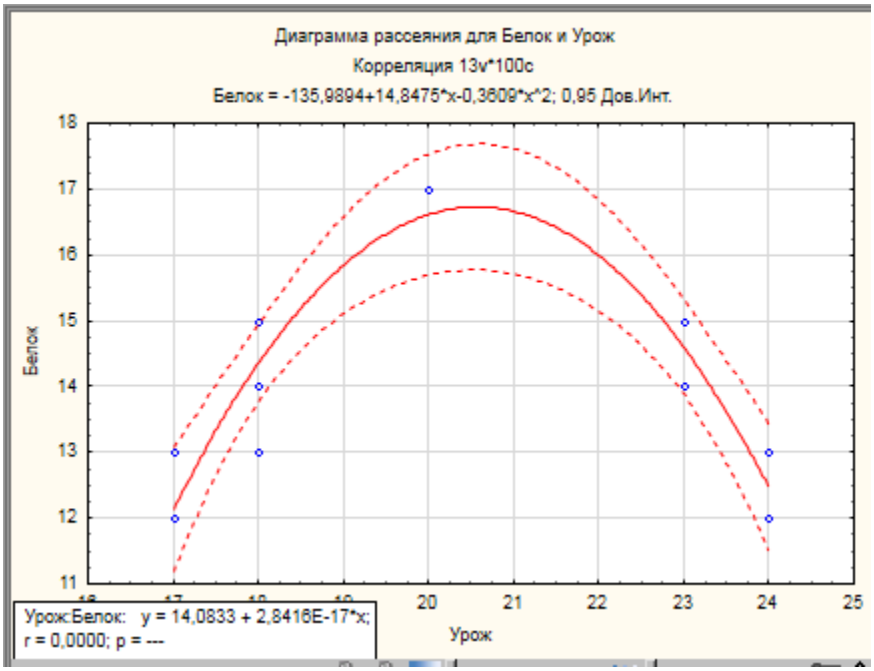


Рис. 6.22. **График полиноминальной зависимости**

6.2.3 Регрессионный анализ в модуле Множественная нелинейная регрессия

Пример 3. В полевом опыте изучали зависимость урожайности озимой пшеницы *Y* (ц/га) от засоренности посевов *X* (количество сорняков, шт/м^2).

X	5	10	163	50	100	258	20	286	38	94
Y	48,3	48,2	41,5	45,8	42,4	41,5	48,9	39,6	45,5	43,8
X	75	120	68	75	200	30	145	217	87	235
Y	44,8	41,2	43	45,1	40,6	45,1	41,7	40,1	44	41,4

Для проведения нелинейной регрессии создадим в программе Statstica файл исходных данных с двумя переменными *Сорняки и Урожай* по 20 наблюдений в каждой и внесем цифровые данные по количеству сорняков и урожайности.

Для визуальной оценки связи между изучаемыми признаками построим точечный график, из которого видно, что связь между количеством сорняков и урожайностью, вероятнее всего, носит криволинейный характер (рис. 6.23). Поэтому корреляционно-регрессионный анализ проведем с применением нелинейной статистики.

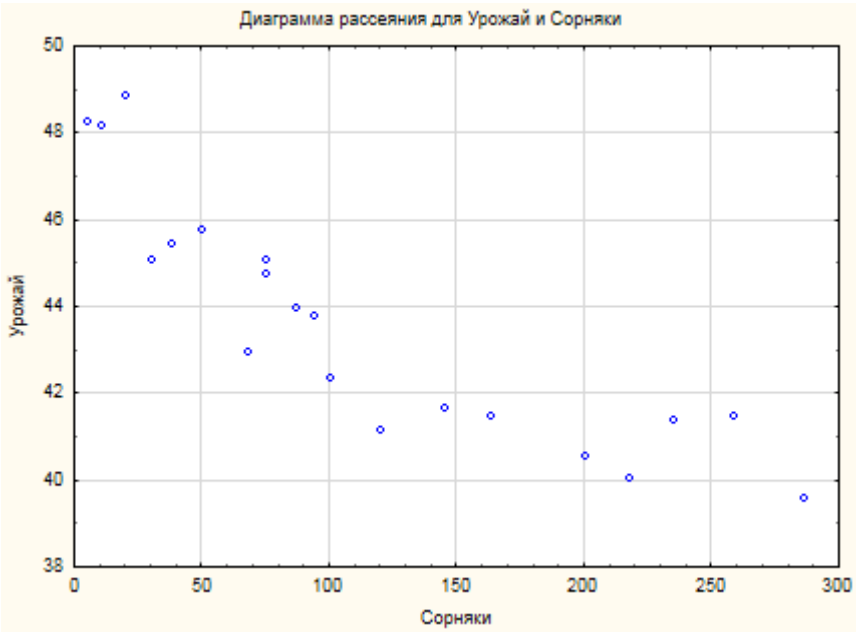


Рис. 6.23. **Точечный график зависимости между урожайностью и количеством сорняков**

В меню **Анализ** выберем модуль **Углубленные методы анализа**, в открывающемся меню второго порядка выберем **Множественная нелинейная регрессия** (рис. 6.23).

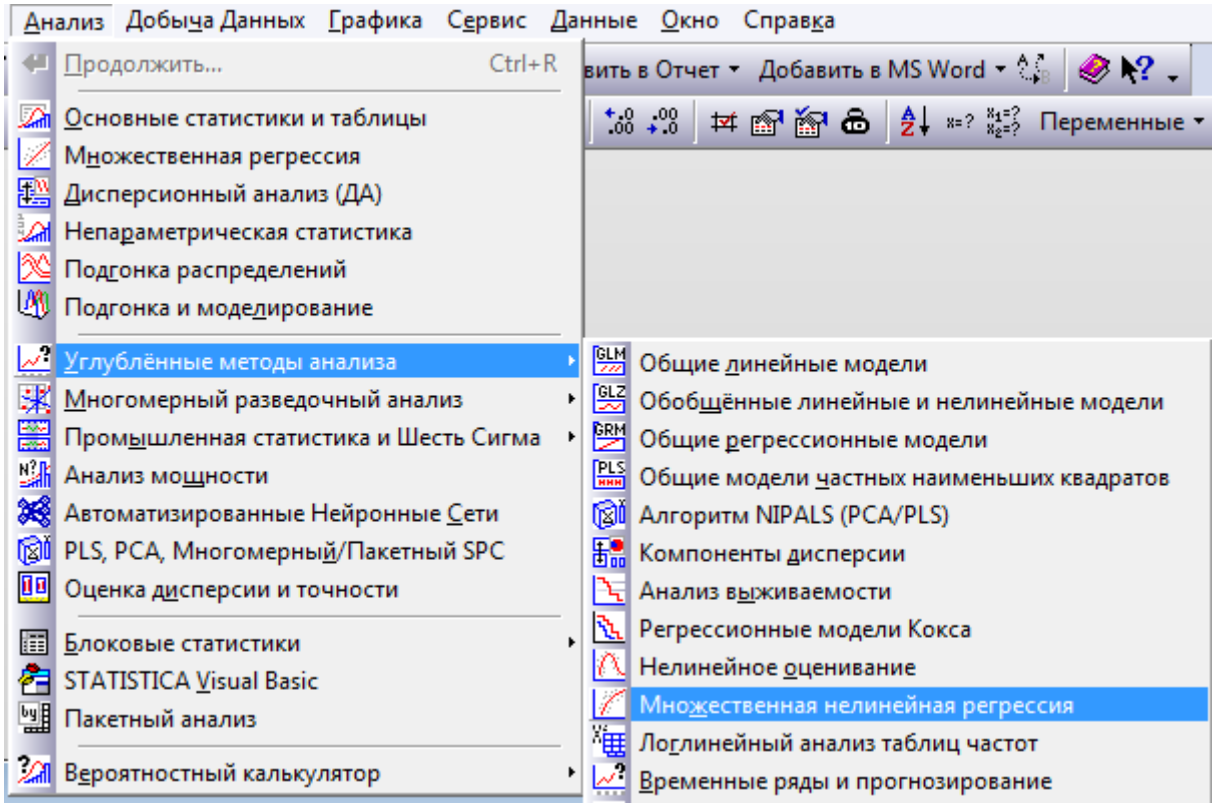


Рис. 6.23. **Диалоговое окно выбора модуля нелинейной регрессии**

В окне выбора переменных (рис. 6.24) отмечаем переменные **Сорняки и Урожай** и нажмем на кнопку **ОК**

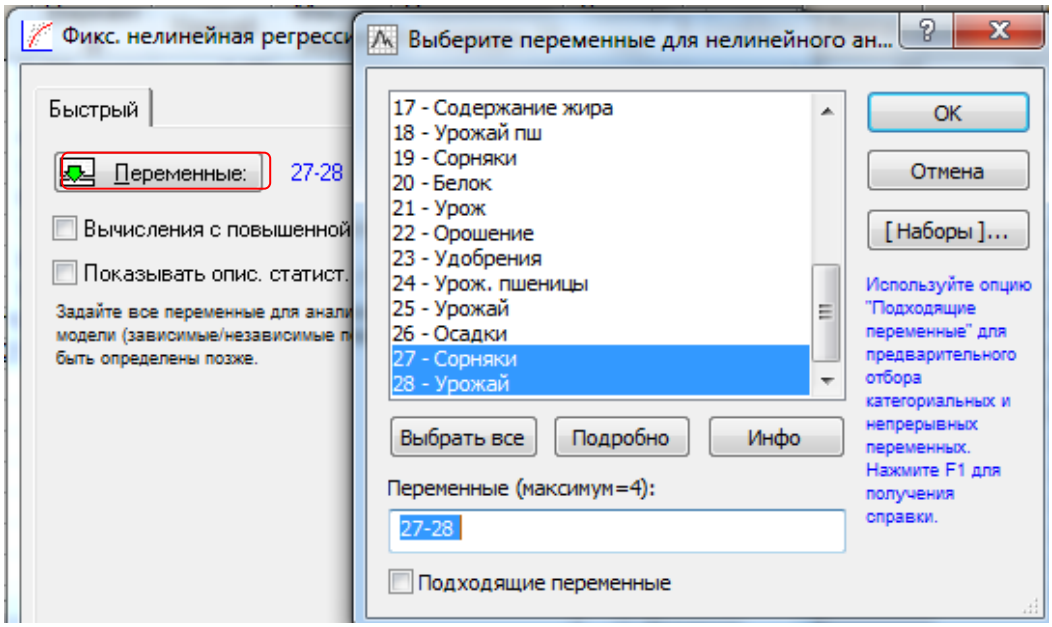


Рис. 6.24. **Окно выбора переменных**

В модуле **Множественная нелинейная регрессия** для нахождения коэффициентов уравнения нелинейного вида проводится преобразование переменных, которые затем могут быть приведены к линейной модели. После выбора переменных и нажатия на кнопку **Ок** появляется окно **Регрессия с нелинейными компонентами (Non-linear Components Regression),** в котором предлагаются различные типы преобразований переменных (рис. 6.25). Исходя из визуальной оценки модели регрессионной зависимости, галочками выберем следующие нелинейные функции преобразований, показанные ниже. В принципе можно выбрать все функции. Однако, для успешного проведения выбранного преобразования, данные должны попадать в допустимый диапазон значений, заданный для данного преобразования; недопустимые наблюдения будут исключены из анализа.

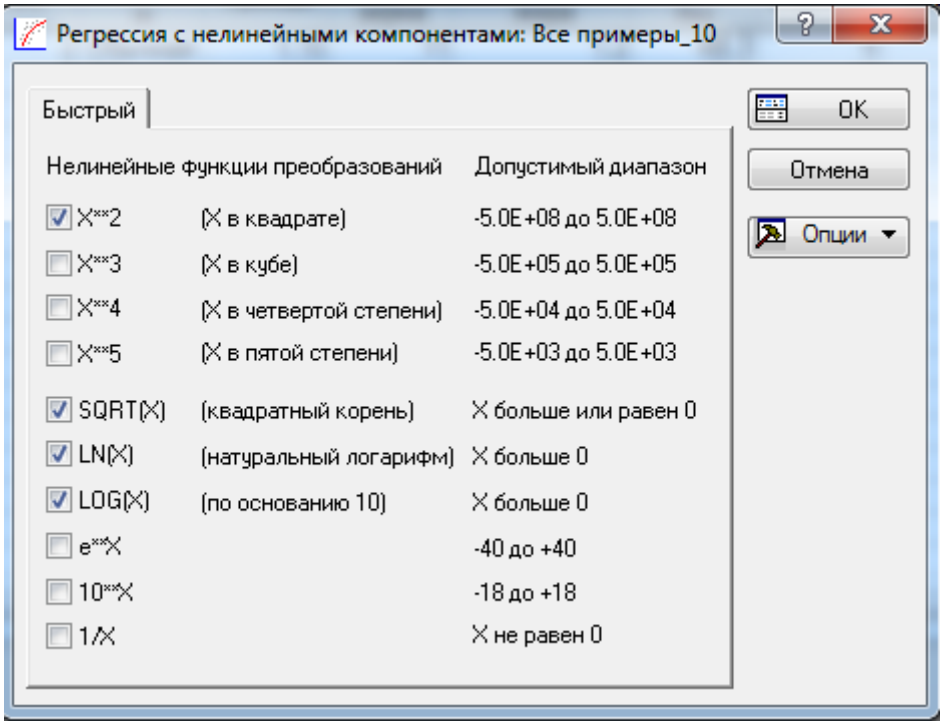


Рис. 6.25. **Диалоговое окно выбора типов преобразования переменных**

После того, как тип (типы) преобразования определен, в оперативной памяти будут созданы списки с дополнительными переменными для каждой переменной и преобразования (рис. 6.26).

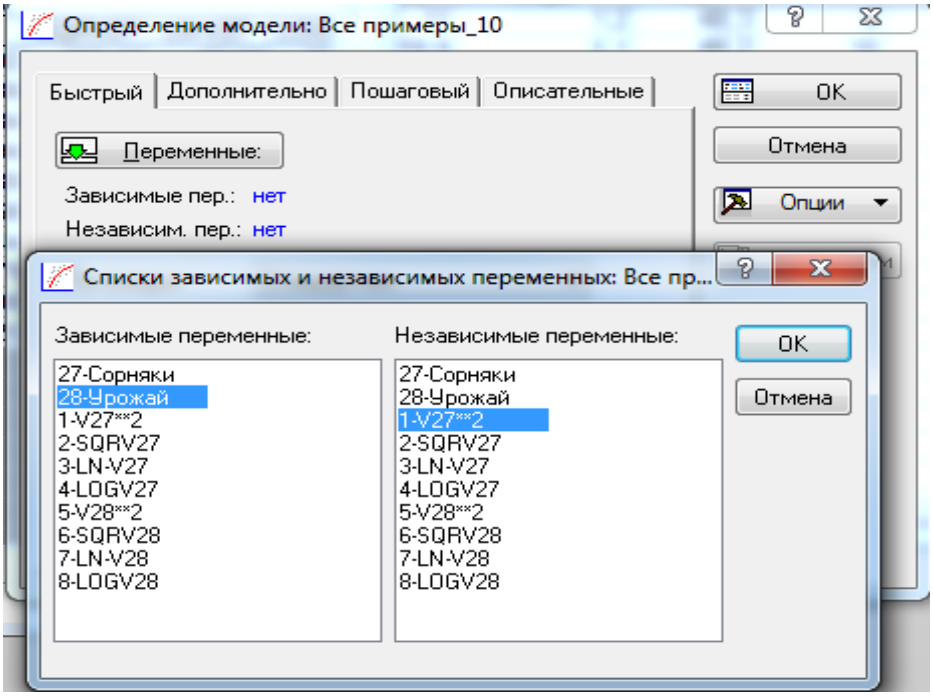


Рис. 6.26. **Списки с преобразованными переменными**

Далее проводим четыре регрессионных анализа, для чего, последовательно выберем в качестве зависимой переменную Урожай, а в качестве независимой переменной преобразованные переменные (*1.V27^2; 2. SQRT V27; 3. LN-V27; 4. LOGV27*). В итоге получаем 4 окна (рис. 6.27) результатов множественной регрессии (для нашего примера с 2-мя переменными – простой регрессии), в верхней части каждого окна указаны коэффициенты корреляции, детерминации, критерии Фишера и уровень значимости (*p*) на основании которых можно судить об адекватности регрессионной модели.

Для более точной оценки результатов регрессии и вывода параметров уравнения нелинейной зависимости в каждом окне результатов множественной регрессии при активной вкладке **Быстрый** нажимаем на кнопку **Итоговая таблица регрессии (Regression summary),** (рис. 6.27).

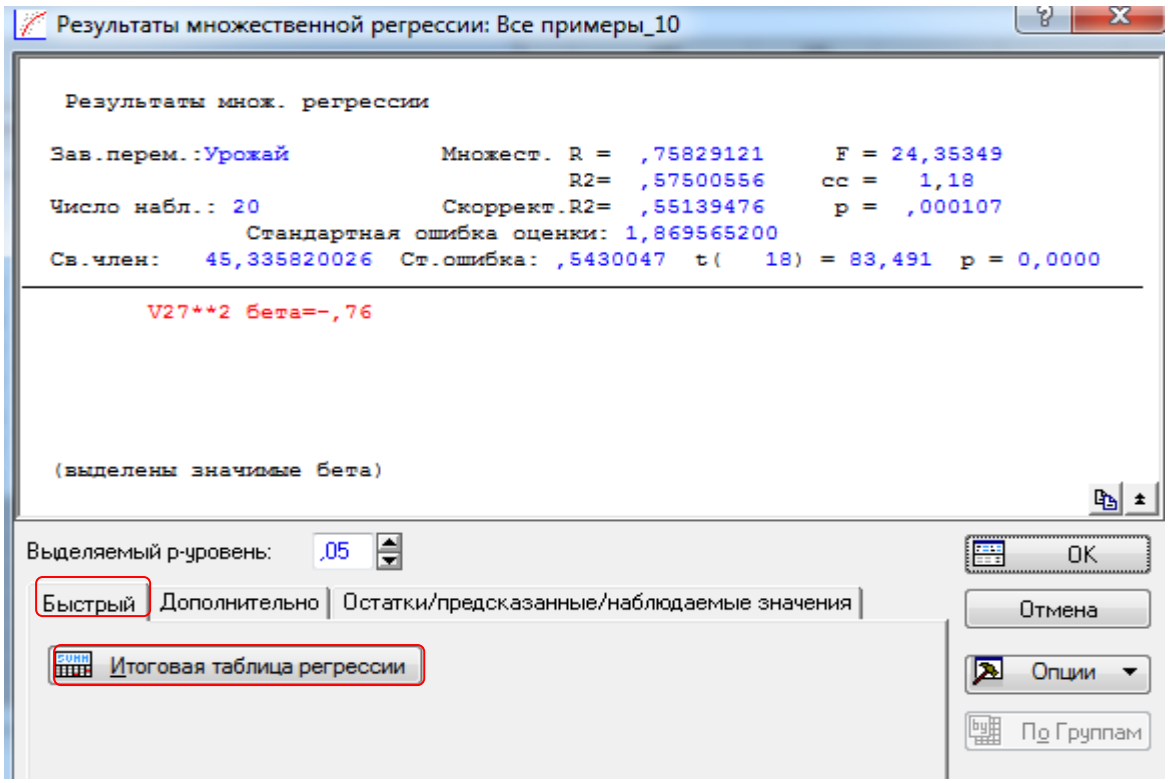


Рис. 6.27. **Диалоговое окно результатов регрессионного анализа**

В рабочей книге получаем четыре таблицы с итогами корреляционно-регрессионного анализа (рис. 6.28), в которых показаны значения коэффициентов корреляции, детерминации, регрессии, величина свободного члена, фактическое значение критериев *F* и *t* и уровень значимости параметров уравнения. Для простых (парных) зависимостей коэффициент БЕТА равен коэффициенту корреляции. В представленных таблицах в первой колонке указан тип преобразования независимой переменной.

Итоги регрессии для зависимой переменной: Урожай (Все г
R= ,75829121 R2= ,57500556 Скоррект. R2= ,55139476
F(1,18)=24,353 p<,00011 Станд. ошибка оценки: 1,8696

N=20	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(18)	p-знач.
Св.член			45,33582	0,543005	83,49066	0,000000
V27**2	-0,758291	0,153658	-0,00009	0,000017	-4,93493	0,000107

Итоги регрессии для зависимой переменной: Урожай (Все п
R= ,93456545 R2= ,87341258 Скоррект. R2= ,86637995
F(1,18)=124,19 p<,00000 Станд. ошибка оценки: 1,0203

N=20	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(18)	p-знач.
Св.член			49,56020	0,579338	85,5463	0,000000
SQRV27	-0,934565	0,083861	-0,60522	0,054308	-11,1442	0,000000

Итоги регрессии для зависимой переменной: Урожай (Все г
R= ,92859538 R2= ,86228937 Скоррект. R2= ,85463878
F(1,18)=112,71 p<,00000 Станд. ошибка оценки: 1,0642

N=20	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(18)	p-знач.
Св.член			53,87626	0,994432	54,1779	0,000000
LN-V27	-0,928595	0,087468	-2,36920	0,223163	-10,6164	0,000000

Итоги регрессии для зависимой переменной: Урожай (Все г
R= ,92859538 R2= ,86228937 Скоррект. R2= ,85463878
F(1,18)=112,71 p<,00000 Станд. ошибка оценки: 1,0642

N=20	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(18)	p-знач.
Св.член			53,87626	0,994432	54,1779	0,000000
LOGV27	-0,928595	0,087468	-5,45528	0,513852	-10,6164	0,000000

Рис. 6.28. **Итоговые таблицы регрессии с преобразованными значениями**

Для выбора параметров уравнения регрессии и лучшей модели нелинейной зависимости проанализируем указанные таблицы итогов регрессии с преобразованными значениями. Во всех таблицах все параметры уравнения регрессии (БЕТА-коэффициент, коэффициент регрессии и свободный член) окрашены красным цветом, так как $p< 0,0001$, что показывает на их высокую значимость (рис. 6.28)

Для окончательного выбора наилучшей модели продолжим сравнение итогов 4-х таблиц. Наиболее точная аппроксимация достигается при более высоких значениях коэффициента детерминации (R^2), коэффициентов Фишера (F), Стьтюдента (t) и наименьших значениях уровня регрессии (p). Этим критериям аппроксимации удовлетворяют логарифмические модели и преобразование численности сорняков через корень квадратный.

Таким образом, нелинейную зависимость между урожайностью (Y) и количеством сорняков (X) можно описать следующими уравнениями:

Y= 49,56 – 0,60√X; Y= 53,88 – 2,37·lnX; Y= 53,88 – 5,45·logX

Для построения графика нелинейной зависимости между урожайностью и количеством сорняков выберем в меню **График** модуль **Диаграммы рассеяния** (рис. 6.29).

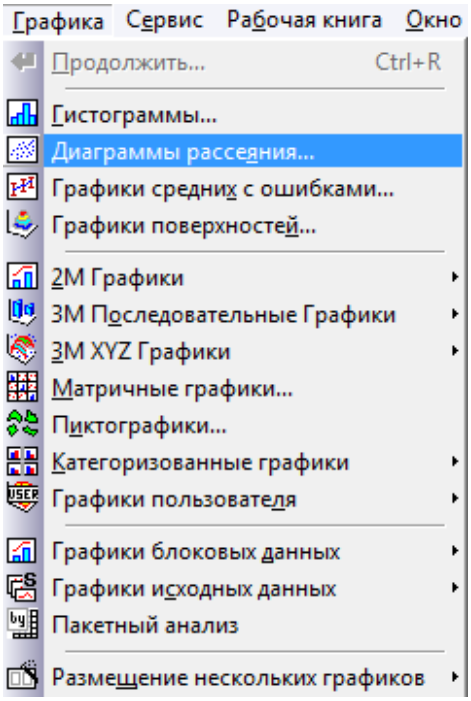


Рис. 6.29. **Панель выбора графиков**

В появившемся диалоговом окне (рис. 6.30) укажем переменные X – *Сорняки* и Y – *Урожай,* выберем **Простой тип графика** и в списке **Подгонка – Логарифмическая (Logarithmic)**, отметим **Уравнение регрессии и 95% доверительный интервал.**

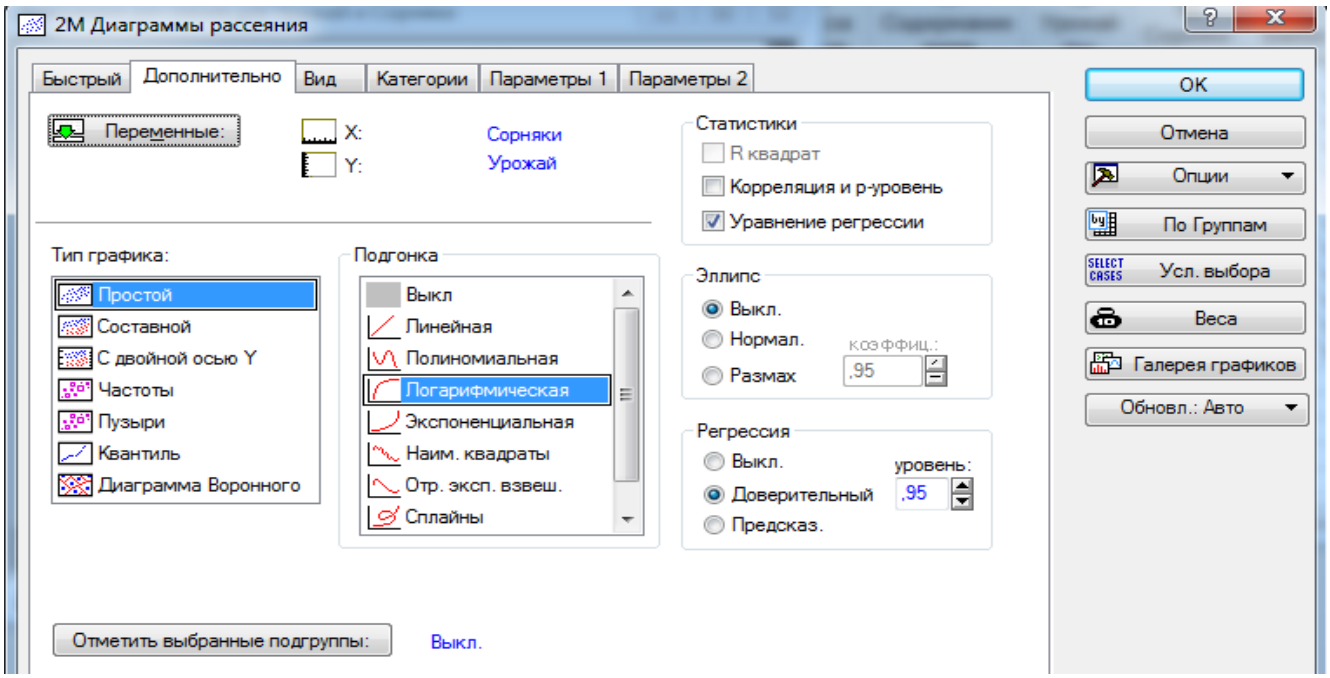


Рис. 6.30. **Диалоговое окно выбора типа криволинейной регрессии**

После нажатия на кнопку **Ок** в рабочей книге получим график (рис. 6.31.), на котором показаны распределение фактических точек на графике, что отражает связь между количеством сорняков и урожайностью озимой пшеницы, и теоретическая линия регрессии (логарифмическая), описывающая эту зависимость. В верхней части графика представлено уравнение криволинейной связи $\boldsymbol{Y = 53{,}88 - 5{,}45 \cdot Log^{10}(X)}$

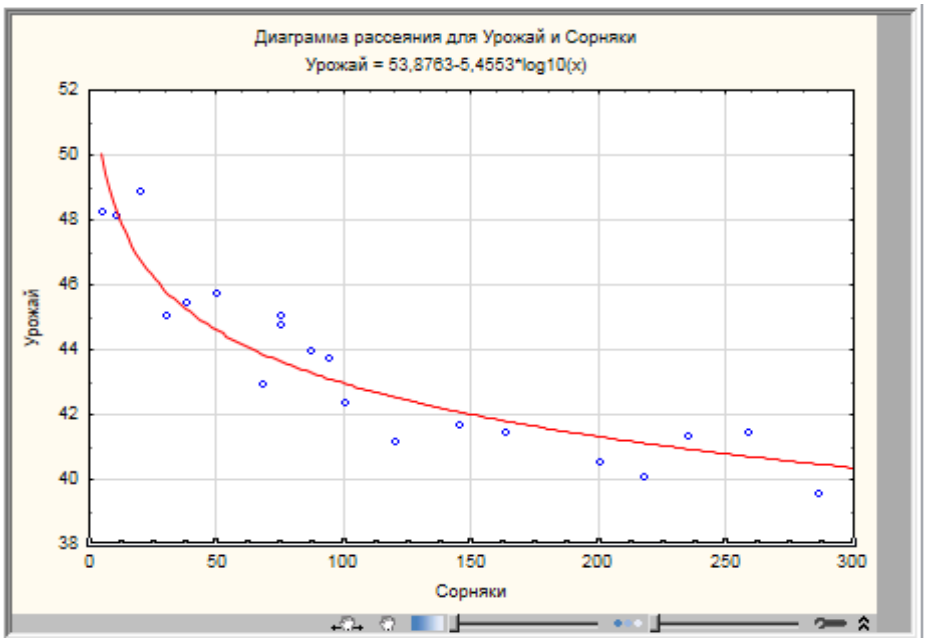


Рис. 6.31. **График зависимости между урожайностью и количеством сорняков**

6.3. Множественная корреляция и регрессия

При изучении зависимости между признаками, когда их число больше двух, мы имеем дело с множественной корреляцией и регрессией.

Множественная корреляция характеризует тесноту и направленность связи между результативным (зависимым) и несколькими независимыми (факторными) признаками. Для оценки множественной корреляции рассчитывается коэффициент множественной корреляции

$$R_{y \cdot xz} = \sqrt{\frac{r_{yx}^2 + r_{yz}^2 - 2r_{yx} \cdot r_{yz} \cdot r_{xz}}{1 - r_{xz}^2}}$$

Коэффициент множественной корреляции R – величина безразмерная и изменяется в пределах от 0 до 1. С помощью множественного коэффициента корреляции мы можем судить только о тесноте взаимосвязи между изучаемыми признаками в целом. Значимость множественного коэффициента корреляции проверяется на основе t-критерия Стьюдента.

Квадрат множественного коэффициента корреляции называют **множественным коэффициентом детерминации (R^2)**. R^2 оценивает долю вариации результативного фактора за счет представленных в модели факторов в общей вариации результата.

Коэффициент детерминации R^2 показывает, *насколько изменения зависимого признака (в долях или в процентах) объясняются изменениями совокупности независимых признаков*.

Для измерения связей между отдельными признаками служит матрица **парных коэффициентов корреляции**. По ней можно в первом приближении судить о тесноте связи независимых признаков между собой и с результативным признаком, а также осуществлять предварительный отбор признаков для включения их в уравнение регрессии.

Более точную характеристику тесноты зависимости дают **частные коэффициенты корреляции**. Частный коэффициент корреляции служит показателем линейной связи между двумя признаками, исключая влияние всех остальных представленных в модели факторов.

Статистическая надежность регрессионного уравнения в целом оценивается на основе F-критерия Фишера: Для проверки $\boldsymbol{H_0}$ следует рассчитать значение F- критерия ($\boldsymbol{F_ф}$) и сравнить его с табличным значением ($\boldsymbol{F_{05}}$ ***или*** $\boldsymbol{F_{01}}$).

Множественная регрессия описывает форму связи в виде уравнения множественной регрессии: $\boldsymbol{Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k}$

где: $\boldsymbol{X_1, X_2, \ldots, X_k}$ независимые переменные (факторы);

$\boldsymbol{b_1, b_2, \ldots, b_k}$ соответствующие им коэффициенты регрессии

Параметры уравнения множественной регрессии $\boldsymbol{b_1, b_2, \ldots, b_k}$ называют **коэффициентами множественной регрессии.** Полученные коэффициенты множественной регрессии являются именованными числами и показывают, как в среднем изменится значение *результативного признака*, если соответствующий *факторный (независимый) признак* увеличится на единицу *при фиксированных значениях всех остальных факторов.*

Коэффициенты регрессии можно преобразовать в сравнимые относительные показатели – **стандартизованные коэффициенты регрессии** или *β-коэффициенты* (БЕТА-коэффициенты). *β-коэффициент* позволяет оценить меру влияния вариации факторного признака

Пример. 4. В опыте изучается зависимость урожайности ячменя (Y) от содержания гумуса (X_1) и агрофизических показателей почвы: плотности (X_2), порозности (X_3) и водопрочности структуры (X_4). Необходимо оценить степень и значимость влияния изучаемых признаков на урожайность ячменя и построить теоретическую модель регрессии.

В программе Statistica создадим файл исходных данных «Корреляция» с пятью переменными с указанными названиями и введем значения, как представлено ниже (табл. 3).

Таблица 3

Зависимость урожайности ячменя (*Y*) от содержания гумуса (X_1) плотности (X_2), порозности (X_3) и водопрочности структуры (X_4) почвы.

Гумус, %, X_1	Плотность, г/см^3 X_2	Порозность, %, X_3	Водопрочность %, X_4	Урожайность ячменя, т/га, Y
3,83	1,28	55,80	27,30	2,82
4,21	1,14	63,30	28,30	3,08
3,97	1,29	55,80	27,00	2,72
3,69	1,27	57,80	26,40	2,66
4,01	1,28	56,30	25,50	2,87
3,92	1,31	53,90	28,00	2,82
4,21	1,14	64,10	29,40	3,11
3,54	1,32	54,00	27,40	2,65
3,69	1,30	55,30	27,10	2,73
3,84	1,31	54,20	25,80	2,74
3,92	1,28	56,70	28,10	2,86
4,01	1,25	60,80	26,10	3,01
3,40	1,28	57,20	27,40	2,62
4,00	1,20	61,50	29,10	2,90
3,85	1,27	57,10	28,70	2,84
3,80	1,28	57,20	27,30	2,77
4,01	1,24	58,40	29,80	2,96
3,70	1,30	55,40	25,60	2,73
4,11	1,24	61,30	30,50	3,18
4,07	1,27	59,40	28,40	2,90

В меню **Анализ (Statistics)** выберем модуль **Множественная регрессия (Multiple Regression),** (рис. 6.32)

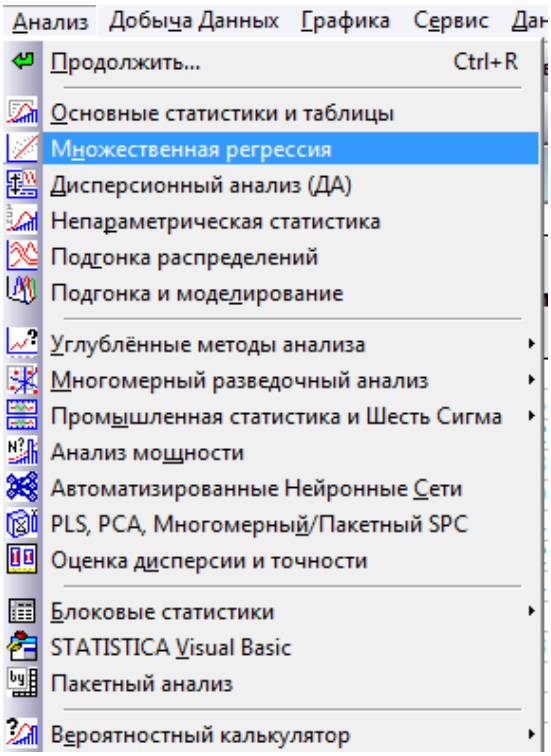


Рис. 6.32. **Выбор модуля Множественная регрессия**

В появившемся диалоговом окне (рис. 6.33) активируем вкладку **Дополнительно (Advanced)** и галочкой отметим опцию **Показать описательную статистику, корр. матрицы.**

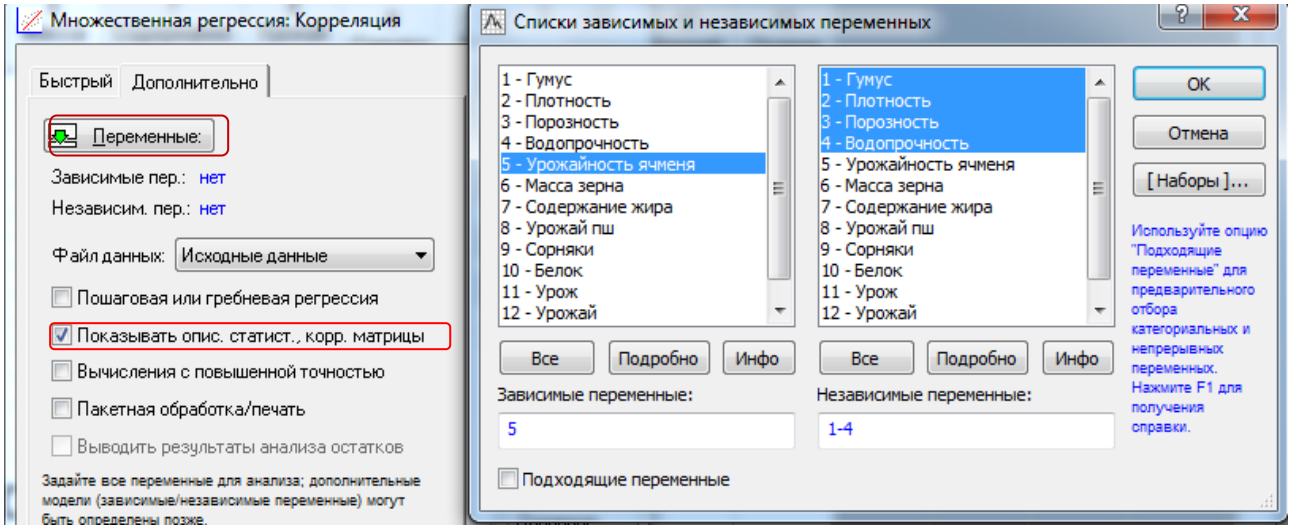


Рис. 6.33. **Диалоговое окно настройки опций и ввода переменных множественной регрессии**

Щелкнем по кнопке **Переменные (Variables)** и в появившемся диалоговом окне (рис. 6.33) в качестве **Зависимой переменной (Dependent)** укажем Урожайность ячменя и **независимых переменных (Independent) –**

переменные: *Гумус, Плотность, Порозность и Водопрочность*. В окне просмотра описательных статистик активируем вкладку **Быстрый** и нажмем на кнопку **Корреляции** (рис. 6.34).

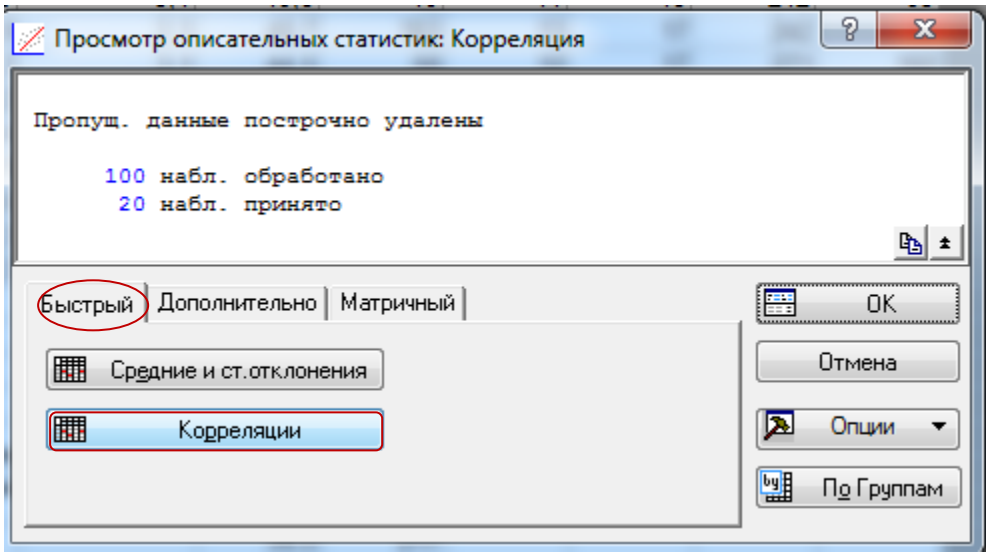


Рис. 6.34. **Окно просмотра описательных статистик**

После нажатия на кнопку **Ок** в рабочей книге получим таблицу, в которой представлена полная матрица парных коэффициентов корреляции между всеми изучаемыми признаками (рис. 6.35).

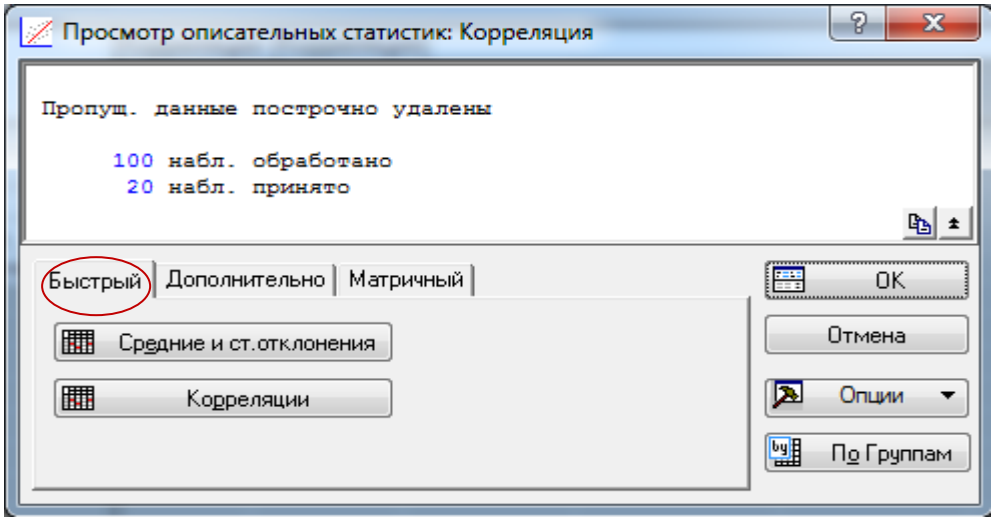
Переменная	Корреляции (Корреляция)				
	Гумус	Плотность	Порозность	Водопрочность	Урожайность ячменя
Гумус	1,000000	-0,679341	0,670172	0,450513	0,875774
Плотность	-0,679341	1,000000	-0,934768	-0,535512	-0,748347
Порозность	0,670172	-0,934768	1,000000	0,550481	0,795848
Водопрочность	0,450513	-0,535512	0,550481	1,000000	0,612129
Урожайность ячменя	0,875774	-0,748347	0,795848	0,612129	1,000000

Рис. 6.35. **Матрица парных корреляций**

Из данных этой таблицы видно, что между урожайностью ячменя и содержанием гумуса, а также порозностью отмечается тесная положительная связь, между плотностью почвы и урожайностью связь обратная и тесная, влияние водопрочной структуры на урожайность носит средний характер (рис. 6.35). Помимо влияния независимых признаков на урожайность в таблице представлены парные корреляции между независимыми признаками. Определенный интерес представляет очень высокий коэффициент между

плотностью и порозностью, который в дальнейшем будет учитываться для устранения *мультиколлинеарности.*

Для возврата в предыдущее диалоговое окно нажмем в нижнем левом углу на свернутую панель **Просмотр описател…**



Активируем вкладку **Быстрый (Quick),** для вывода результатов анализа нажмем на кнопку **Ок.** Программа произведет вычисления и на экране появится расширенное диалоговое окно (рис. 6.36) с результатами множественной регрессии: коэффициент множественной корреляции (*R= 0,936*) указывает на тесную связь урожайности ячменя со всем комплексом независимых признаков, коэффициент детерминации равен R^2 *= 0, 876 = 87,6%.* Коэффициент детерминации является одной из основных статистик в данном окне, он показывает долю в вариации урожайности, обусловленную действием независимых признаков (гумус, водопрочная структура, плотность и порозность). Чем ближе коэффициент детерминации к единице, тем более тесная связь. Для оценки регрессии служат критерий *F* и уровень значимости. Значения критерия Фишера- *F = 26,6,* уровень значимости- *p = 0,00001* свидетельствуют об адекватности указанной регрессионной модели.

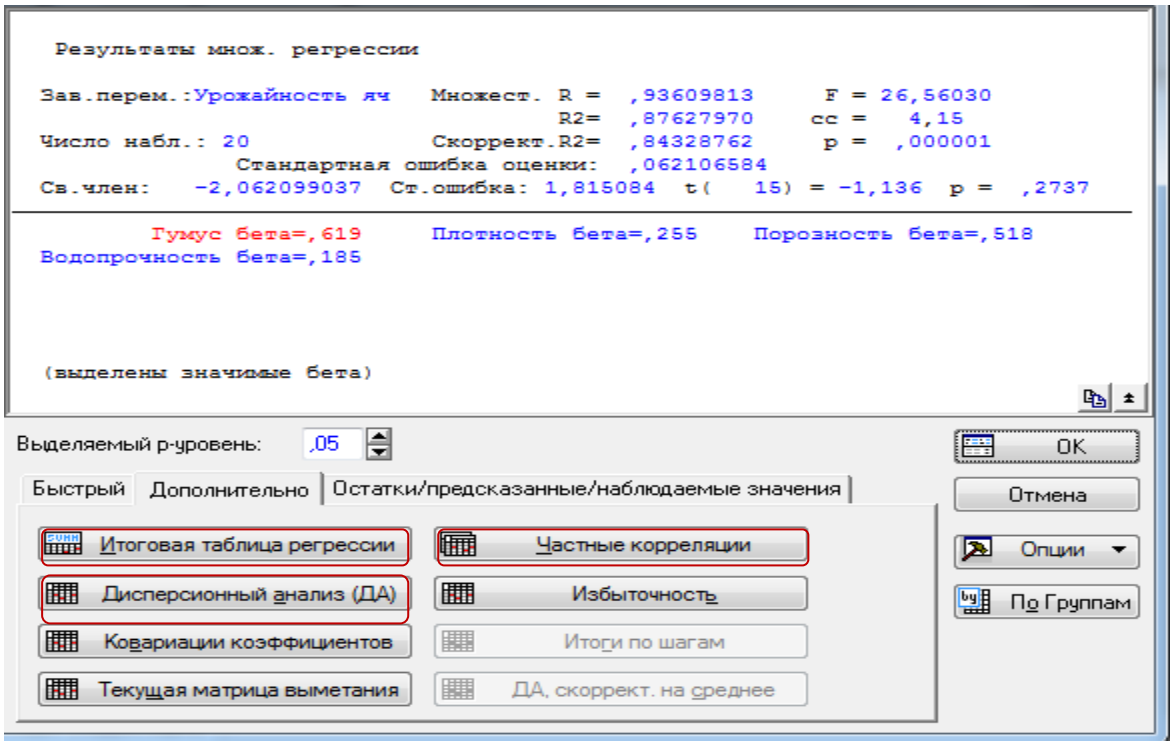


Рис. 6.36. **Расширенное диалоговое окно для вывода результатов множественной регрессии**

Нижняя часть расширенного окна (рис. 6.36) содержит функциональные кнопки, нажимая на которые мы можем более подробно рассмотреть результаты вычислений показателей множественной корреляции или регрессии (итоговая регрессия, дисперсионный анализ регрессии, частные коэффициенты, остатки и др.).

После нажатия на кнопку **Итоговая таблица регрессии (Summary regression results)** получаем таблицу с ***БЕТА коэффициентами,*** коэффициентами регрессии – $в_{yx}$, *t*- критериями и *p-значениями (рис. 6.37).* Значение $t= 4,917$ и уровень вероятности $p = 0,000186$ показывают, что из всех независимых признаков значимым на 95% уровне вероятности является только содержание гумуса.

N=20	Итоги регрессии для зависимой переменной: Урожайность я R= ,93609813 R2= ,87627970 Скоррект. R2= ,84328762 F(4,15)=26,560 p<,00000 Станд. ошибка оценки: ,06211 БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(15)	p-знач.
Св.член			-2,06210	1,815084	-1,13609	0,273745
Гумус	0,618684	0,125814	0,46065	0,093676	4,91743	0,000186
Плотность	0,254961	0,261214	0,79438	0,813857	0,97606	0,344513
Порозность	0,517756	0,260960	0,02676	0,013487	1,98404	0,065860
Водопрочность	0,184924	0,109854	0,02062	0,012247	1,68336	0,112997

Рис. 6.37. **Итоги регрессии**

Бета коэффициенты – это стандартизированные коэффициенты, величины безразмерные. Введение этих показателей обусловлено тем, что при множественной регрессии изучаемые независимые признаки имеют разные единицы измерения, поэтому и коэффициенты регрессии в уравнении связи имеют разные единицы измерения, что делает их несопоставимыми, если возникает вопрос о об определении силы влияния факторов на результативный признак. Поэтому для сравнительной оценки воздействия каждого независимого признака (фактора) на результативный признак необходимо стандартизировать коэффициенты регрессии. С этой целью все переменные уравнения регрессии выражают в долях среднеквадратического отклонения. Такие стандартизированные коэффициенты регрессии называют ***БЕТА – β коэффициентами***.

Бета-коэффициент β показывает, что если величина фактора увеличится на одно среднеквадратическое отклонение, то соответствующая зависимая переменная увеличится или уменьшится на долю своего среднеквадратического отклонения. Таким образом, по значениям бета-коэффициентов можно судить о вкладе независимых признаков на урожайность ячменя. Из данных таблицы видно, что из всех признаков наибольший и значимый вклад в изменение урожайности вносит содержание гумуса. Об этом свидетельствуют выделение этой строки красным цветом и самая низкая вероятность отклонения нулевой гипотезы (***p<0,01***).

В верхней части таблицы итогов регрессии (рис. 6,37) и подробной таблице дисперсионного анализа (рис. 6.38) приведены значения критерия

Фишера *F(4,15) = 26,56* и его уровень значимости (***p=0,000001***). В скобках при критерии Фишера указаны степени свободы 4 – число переменных *m(5) –1* и 15 – число наблюдений *n(20) – m(5)*. Для проверки существенности регрессионной зависимости можно сравнить рассчитанное значение критерия Фишера $\boldsymbol{F_{ф}}$ **= *26,56*** с табличным значением критерия Фишера $\boldsymbol{F_{05}}$ **= *3,06***. Так как фактическое значение критерия Фишера значительно больше табличного это свидетельствует о статистической значимости уравнения регрессии, его параметров и показателя тесноты связи в целом с учетом действия всех 4-х переменных.

Эффект	Дисперсионный анализ; ЗП: Урожайность яч				
	Сумма квадр.	cc	Средн. квадр.	F	p-знач.
Регресс.	0,409797	4	0,102449	26,56030	0,000001
Остатки	0,057858	15	0,003857		
Итого	0,467655				

Рис. 6.38. **Дисперсионный анализ регрессии**

В программе Statistica для проверки нулевой гипотезы вместо табличного значения критерия Фишера приводятся вероятности значимости (***p***) и красным цветом выделяются значимые показатели. В нашем случае ***p*** значительно меньше заданного *p-level 0,05*, что также подтверждает значимость параметров множественной регрессии.

Более объективную характеристику тесноты связи дают **частные коэффициенты корреляции**, измеряющие влияние на результативный фактор *Y* фактора X_i при неизменном уровне других факторов. Коэффициент частной корреляции отличается от простого коэффициента линейной парной корреляции тем, что он измеряет парную корреляцию соответствующих признаков (*Y* и X_i) при условии, что влияние на них остальных факторов (X_j) устранено.

Для получения значений частных корреляция в расширенном окне вывода результатов множественной регрессии (рис. 6.36) нажмем на кнопку **Частные корреляции (Partial Correlation)** и в рабочей книге получим

таблицу с результатами частных корреляций и ***Бета-коэффициентами*** (рис. 6.39).

Переменная	Переменные, входящие в уравнение (Корреляция)						
	Бета(в)	Частная Корр.	Получаст Корр.	Толеран.	R-квадр.	t(15)	p-знач.
Гумус	0,618684	0,785597	0,446595	0,521062	0,478938	4,917434	0,000186
Плотность	0,254961	0,244377	0,088645	0,120881	0,879119	0,976063	0,344513
Порозность	0,517756	0,455934	0,180188	0,121116	0,878884	1,984040	0,065860
Водопрочность	0,184924	0,398617	0,152880	0,683466	0,316534	1,683358	0,112997

Рис. 6.39. **Частные корреляции**

Сравнивая частные коэффициенты корреляции (рис. 6.39) с соответствующими парными коэффициентами (рис. 6.35), видим, что только коэффициент корреляции между гумусом и урожайностью практически мало изменился – с 0,87 до 0,78, в то время как коэффициенты корреляций между урожайностью и агрофизическими показателями очень сильно снизились. Такая разница в величинах парных и частных корреляций связана с эффектом *мультиколлинеарности (*наличие линейной зависимости между независимыми переменными, включёнными в модель) – влиянием независимых признаков (факторов) друг на друга (r_{xz}*> 0,7).* При этом каждый независимый признак влияет на результативный как непосредственно, так и опосредованно, через связь с другими признаками. Сильная взаимная коррелированность отдельных независимых переменных в нашем уравнении затрудняет анализ влияния отдельных факторов на зависимую переменную. При наличии такого явления для корректной интерпретации результатов множественной регрессии необходимо минимизировать влияние *мультиколлинеарности*, например, из каждой группы тесно связанных независимых признаков оставлять только один.

6.3.1 Устранение эффекта мультиколлинеарности – гребневая регрессия

Анализ матрицы парных корреляций нашего примера (рис. 6.35) указывает на весьма сильную (приближающуюся к функциональной, *r =0,93*) связь между плотностью почвы и порозностью. При такой сильной зависимости между независимыми признаками, когда коэффициент корреляции между ними *r>±0,8* рекомендуется один из независимых признаков (факторов) исключить

из рассмотрения. Корреляция между другими независимыми признаками слабая или средняя (связь между содержанием гумуса с плотностью или порозностью).

Поиск наилучшей регрессионной модели представляет собой довольно громоздкий процесс. Для устранения эффекта *мультиколлинеарности* и нахождения наилучшего регрессионного уравнения в программе Statistica можно воспользоваться методом *пошагового регрессионного анализа или гребневой регрессии*, который последовательно проверяет независимые переменные на отсутствие между ними *мультиколлинеарности*.

Оценку значимости БЕТА- коэффициентов регрессии с помощью t-критерия используют для завершения отбора существенных факторов в процессе *пошагового регрессионного анализа*. Существует две схемы пошаговой регрессии: *пошаговое включение признаков и пошаговое исключение независимых признаков*.

При использовании метода *пошагового включения признаков*, первым в итоговой регрессии включается независимый признак, наиболее тесно коррелирующий с результирующим признаком, который в паре с другими отобранными дает максимальное значение коэффициента корреляции и бета-коэффициент регрессии.

Суть *пошагового исключения независимых признаков* заключается в том, что после оценки значимости всех бета коэффициентов регрессии из модели исключают ту переменную (независимый признак), бета-коэффициент при котором незначим и имеет наименьшее значение критерия t-Стьюдента. Затем уравнение регрессии строится без исключенного фактора (независимого признака), и снова проводится оценка адекватности уравнения и значимости коэффициентов регрессии. Такой процесс длится до тех пор, пока все стандартизированные коэффициенты регрессии не окажутся значимыми, что свидетельствует о наличии в регрессионной модели только существенных факторов.

Для проведения пошагового регрессионного анализа активируем вкладку **Дополнительно (Advanced),** галочками отметим **Пошаговая или гребневая регрессия (Stepwise or riddle regression) и Показать описательную статистику** (рис. 6.40).

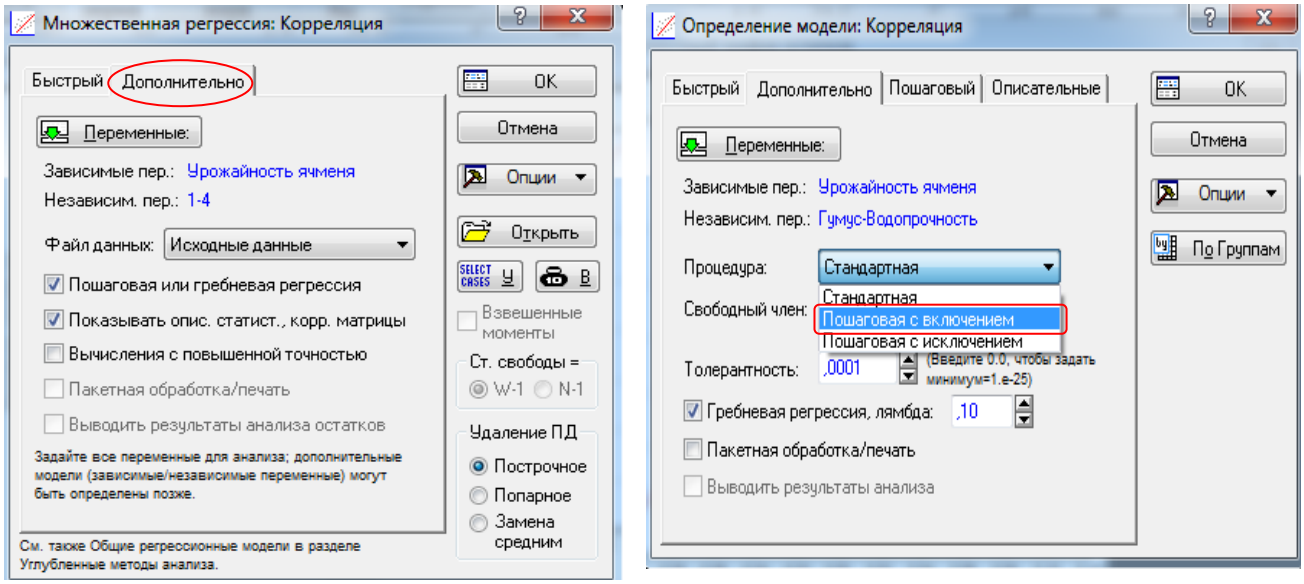


Рис. 6.40. **Диалоговое окно для проведения пошаговой регрессии**

В появившемся окне откажемся от стандартного метода регрессионного анализа и выберем метод пошагового включения переменных в регрессионную модель – процедуру **Пошаговая с включением (Forward stepwise)**. После нажатия на кнопку **Ок** появляется расширенное диалоговое окно (рис. 6.41) с результатами множественной регрессии.

В верхней части окна программа по умолчанию приводит результаты заключительного (третьего) шага гребневой регрессии: коэффициент множественной регрессии $R=0,91$, $R^2= 0,82$, $p =0,00003$, что свидетельствует о высокой значимости включенных в модель признаков. Из четырех независимых признаков в модель включены по значениям бета-коэффициентов регрессии гумус, порозность и водопрочность, при этом два признака: содержание – содержание гумуса и порозность выделены красным цветом, что указывает на значимость этих признаков в регрессионной модели. Из-за эффекта *мультиколлинеарности* плотность исключена из модели.

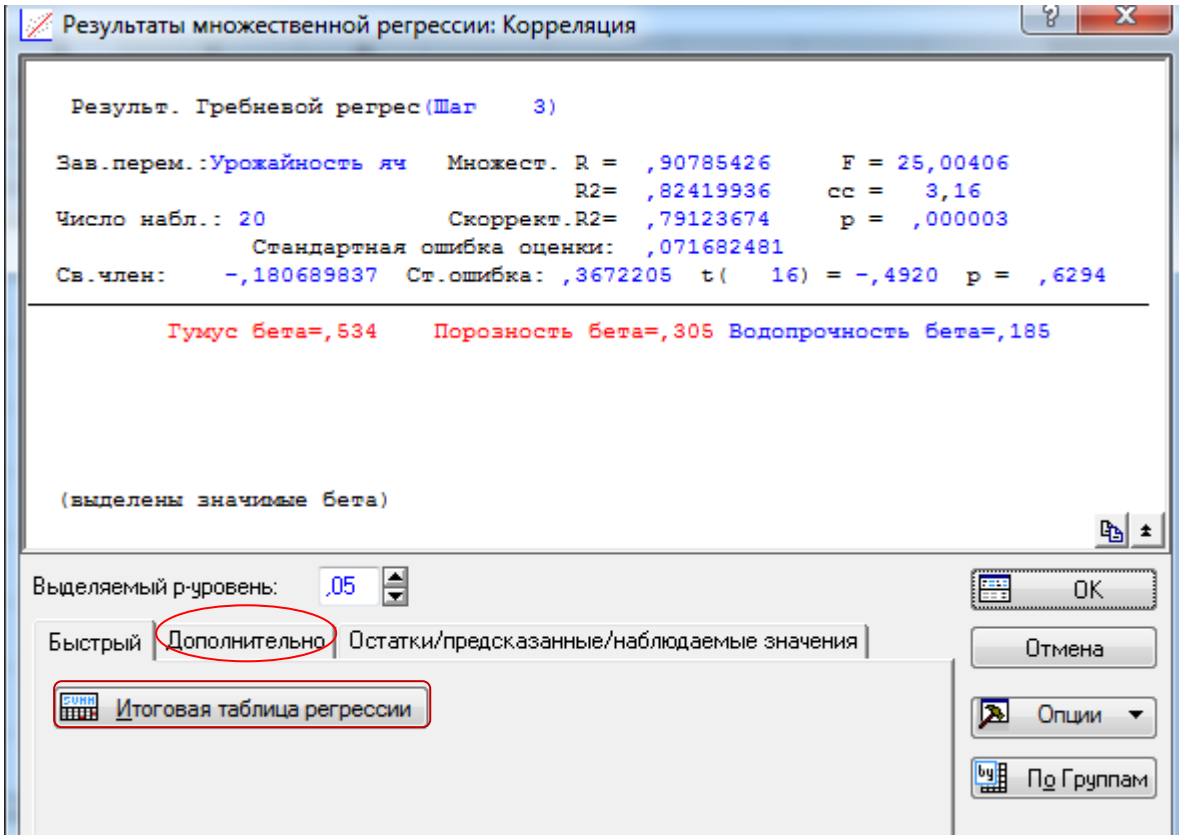


Рис. 6.41. **Расширенное диалоговое окно с результатами пошаговой (гребневой) регрессии**

При активной вкладке **Дополнительно (Advanced)** выберем кнопку **Итоговая таблица регрессии (Summary regression results).** После нажатия на кнопку **Ок** в рабочей книге получим таблицу итогов гребневой регрессии, в которой наибольший интерес представляют *БЕТА-коэффициенты*, коэффициенты регрессии и критерии ***t*** и ***p.*** Все эти показатели для переменных *Гумус* и *Порозность* выделены красным цветом, что говорит об их существенности в регрессионной модели на 05% уровне значимости.

N=20	Итоги Гребневой регрессии для зависимой переменной: Урожа I=,10000 R= ,90785426 R2= ,82419936 Скоррект. R2 ,79123674 F(3,16)=25,004 p<,00000 Станд. ошибка оценки: ,07168					
	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(16)	p-знач.
Св.член			-0,180690	0,367220	-0,492047	0,629369
Гумус	0,534386	0,127526	0,397883	0,094951	4,190418	0,000692
Порозность	0,305468	0,134377	0,015787	0,006945	2,273219	0,037148
Водопрочность	0,184751	0,116804	0,020597	0,013022	1,581725	0,133276

Рис. 6.42. **Итоги гребневой регрессии методом включения**

Проведем пошаговый регрессионный анализ вторым методом – **Пошаговая регрессия с исключением.** Для этого, находясь в окне **Определение модели** (рис. 6.43) выберем процедуру **Пошаговая с исключением**

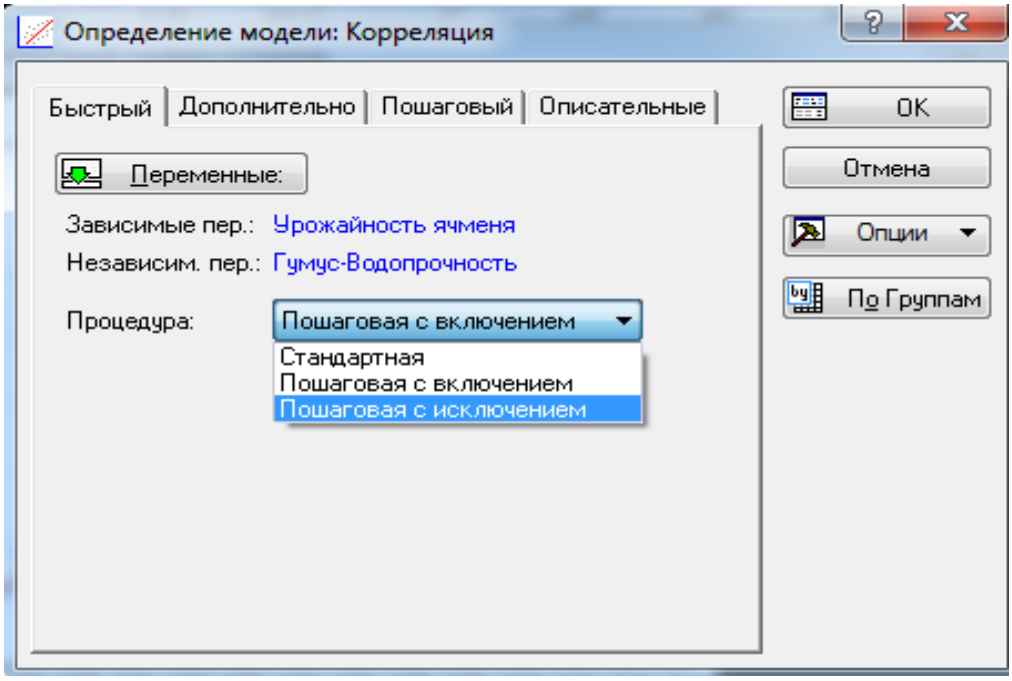


Рис. 6.43. **Диалоговое окно выбора процедуры пошаговой регрессии с исключением**

Далее нажмем на вкладку **Пошаговый** и попадаем в окно (рис. 6.44) выбора отображения результатов пошаговой регрессии, где можно выбрать два способа отображения результатов: **Только итоги (Summary only)** или **На каждом шаге (At each step).**

При выборе первой опции мы получим итоговые окончательные результаты регрессионного анализа, при выборе второй – результаты регрессии после каждого шага исключения переменных. При этом процедура продолжается до тех пор, пока не будут исключены все незначимые коэффициенты и останутся значимые коэффициенты регрессии.

Ниже показан пример процедуры выбора **Пошаговая с исключением (Backward stepwise)** и отображение результатов на каждом шаге **(At each step).**

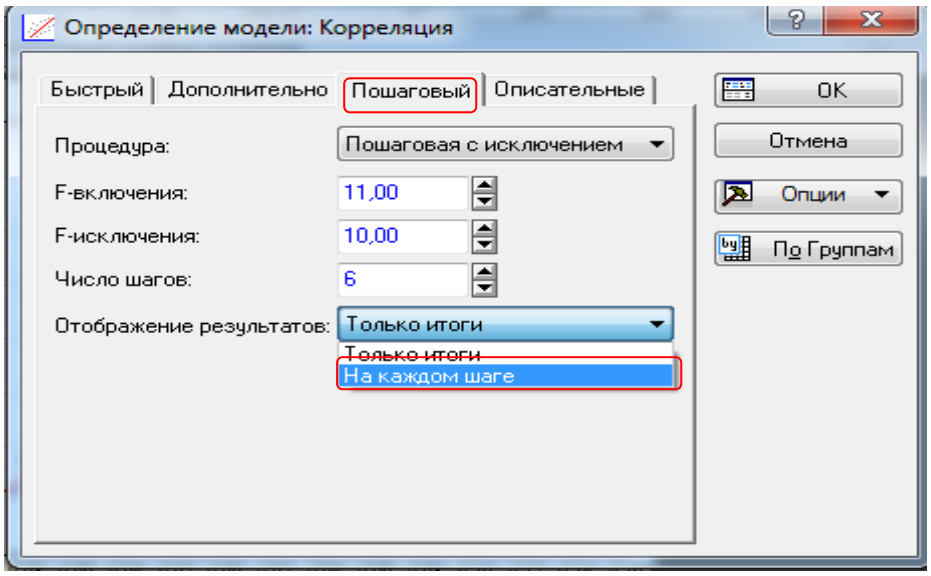


Рис. 6.44. **Диалоговое окно выбора результатов пошаговой регрессии**

После нажатия в расширенном диалоговом окне результатов множественной регрессии (рис. 6.45) на кнопку **Итоговая таблица регрессия** получим таблицу **Итоги гребневой регрессии по шагам.** Далее повторяем процедуру нажатия кнопок после вывода результатов гребневой регрессии на каждом шаге исключения, пока программа не выберет значимые независимые признаки.

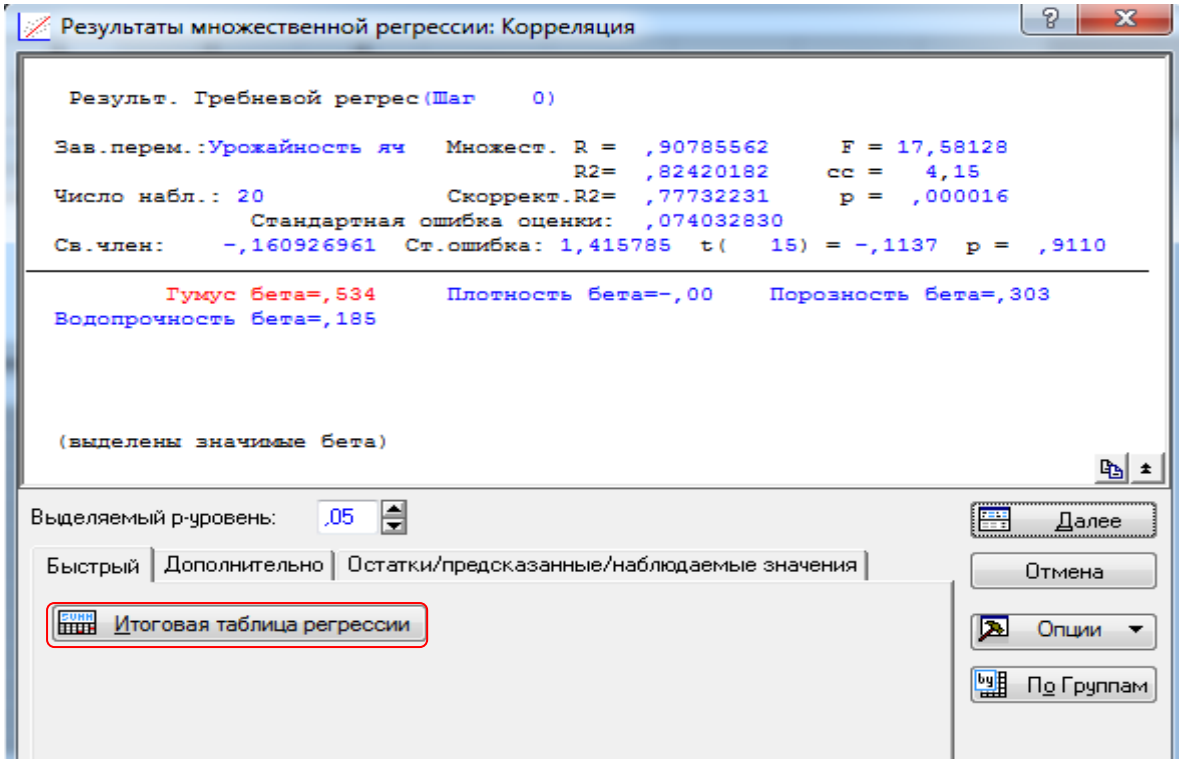


Рис. 6.45. **Расширенное диалоговое окно результатов в начале пошаговой регрессии методом исключения**

Ниже представлены результаты пошаговой (гребневой) регрессии методом исключения на каждом шаге:

Результаты гребневой регрессии (Шаг 0)

N=20	Итоги Гребневой регрессии для зависимой переменной: Урожай I=,10000 R= ,90785562 R2= ,82420182 Скоррект. R2 ,77732231 F(4,15)=17,581 p<,00002 Станд. ошибка оценки: ,07403					
	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(15)	p-знач.
Св.член			-0,160927	1,415785	-0,113666	0,911010
Гумус	0,533950	0,135107	0,397558	0,100596	3,952043	0,001278
Плотность	-0,002937	0,202727	-0,009151	0,631630	-0,014488	0,988631
Порозность	0,303328	0,202670	0,015677	0,010475	1,496663	0,155225
Водопрочность	0,184571	0,121274	0,020576	0,013520	1,521939	0,148824

Результаты гребневой регрессии (Шаг 1)

N=20	Итоги Гребневой регрессии для зависимой переменной: Урожа I=,10000 R= ,90785426 R2= ,82419936 Скоррект. R2 ,79123674 F(3,16)=25,004 p<,00000 Станд. ошибка оценки: ,07168					
	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(16)	p-знач.
Св.член			-0,180690	0,367220	-0,492047	0,629369
Гумус	0,534386	0,127526	0,397883	0,094951	4,190418	0,000692
Порозность	0,305468	0,134377	0,015787	0,006945	2,273219	0,037148
Водопрочность	0,184751	0,116804	0,020597	0,013022	1,581725	0,133276

Результаты гребневой регрессии (Шаг 2)

N=20	Итоги Гребневой регрессии для зависимой переменной: Уро I=,10000 R= ,89258620 R2= ,79671013 Скоррект. R2 ,7727936 F(2,17)=33,312 p<,00000 Станд. ошибка оценки: ,07478					
	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(17)	p-знач.
Св.член			0,079837	0,342400	0,233167	0,818416
Гумус	0,565138	0,131485	0,420779	0,097898	4,298137	0,000487
Порозность	0,379189	0,131485	0,019598	0,006796	2,883906	0,010308

Результаты гребневой регрессии (Шаг 3)

N=20	Итоги Гребневой регрессии для зависимой переменной: Уро I=,10000 R= ,83501771 R2= ,69725458 Скоррект. R2 ,6804353 F(1,18)=41,456 p<,00000 Станд. ошибка оценки: ,08869					
	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(18)	p-знач.
Св.член			0,543150	0,358599	1,514644	0,147225
Гумус	0,796158	0,123653	0,592787	0,092067	6,438625	0,000005

Рис. 6.46. **Результаты гребневой регрессии методом исключения по шагам**

Так, если на нулевом шаге, на котором применялась стандартная процедура, в итоговой таблице регрессии представлены все четыре

независимые переменные, то после первого шага гребневой регрессии методом исключения в итоговой таблице остались три переменные: *гумус, порозность и водопрочность,* рис. 6.46. Из-за очень тесной зависимости между независимыми переменными плотность и порозность для устранения эффекта мультиколлинеарности удалена переменная *плотность*. При этом, результаты гребневой регрессии на этом шаге точно такие же, как в таблице после завершения всех шагов гребневой регрессии методом включения – переменные *гумус* и *порозность* значимы в уравнении множественной регрессии, на что указывает значение ***t*** и ***p***. Из-за того, что статистические показатели корреляции и регрессии переменной *водопрочность* не значимы на *05%* уровне, эта переменная исключена на втором шаге. На третьем, заключительном шаге исключены все переменные, у которых значение $p<0,01$

Таким образом, после проведения пошагового регрессионного анализа методом включения переменных и методом исключения переменных мы приходим к следующему выводу: с вероятностью *95%* значимыми признаками в множественной регрессионной модели, оказывающими существенное влияние на урожайность озимой пшеницы, являются содержание гумуса ($t_ф$ ***=4,298; p = 0,000487***) и порозность почвы ($t_ф$ ***=2,88; p = 0,010308***). В то время как с вероятностью 99% существенно влияние только гумуса ($t_ф$ ***=6,438; p = 0,000005***).

На основании проведенного регрессионного анализа можно составить следующее уравнение множественной регрессии:

$$Y = a + b_1 X_1 + b_2 X_2 = 0,079837+0,420779 \cdot X_1 + 0,019598 \cdot X_2$$

После округления до сотого знака уравнение регрессии принимает следующий вид: $Y = 0,08 + 0,42 \cdot X_1 + 0,02 \cdot X_2$

где Y *– урожайность,* b_1 *– коэффициент регрессии между гумусом и урожайностью*

b_2 *– коэффициент регрессии между порозностью и урожайностью –*

X_1 *– содержание гумуса,* **%;** X_2 *– порозность в %,* a **–** *свободный член.*

Результаты множественной регрессии между содержанием гумуса, порозностью и урожайностью можно представить на графике в трехмерном пространстве.

Графические инструменты программы Statistica предоставляют возможность построения различных двухмерных и трехмерных графиков, на основании которых можно провести более глубокий визуальный анализ.

Для построения графика выберем в галереи меню **Графика (Graphs)** модуль **Графики поверхностей (3D Surface Plot)** (рис. 6.47).

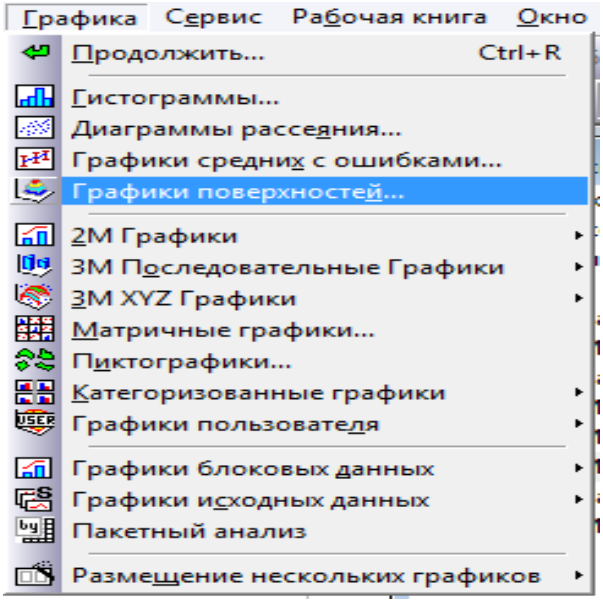


Рис. 6.47. **Меню галереи графиков**

После нажатия на кнопку **Ок** попадаем в диалоговое окно для выбора параметров представления результатов анализа (рис. 6.48). В появившемся диалоговом окне нажмем на вкладку **Быстрый**, по умолчанию оставляем **Линейный** тип **Подгонки**. Щелкнем по кнопке **Переменные (Variablets),** в качестве независимых переменных укажем: *X – Гумус* (номер переменной – 1), *Y– Порозность* (номер переменной – 3) и зависимой переменной *Z –Урожайность ячменя* (номер переменной – 5) и нажмем на клавишу **Ок.**

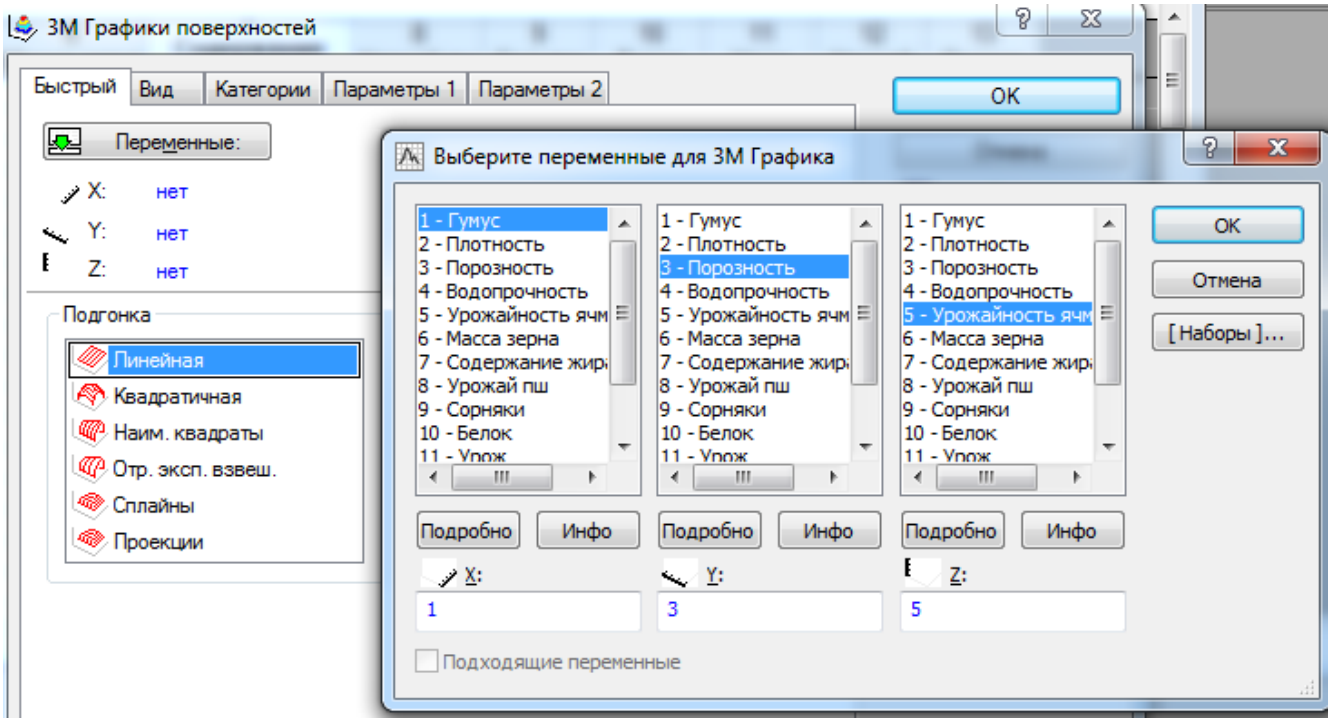


Рис. 6.48. **Диалоговое окно выбора параметров графика и выбора переменных**

После нажатия на кнопку **Ок** в рабочей книге получим трехмерный график поверхностей влияния содержания гумуса и порозности на урожайность ячменя (рис. 6.49). В верхней части графика представлено уравнение множественной регрессии:

Урожайность ячменя = ***0,0839 + 0,4628X + 0,0196Y***

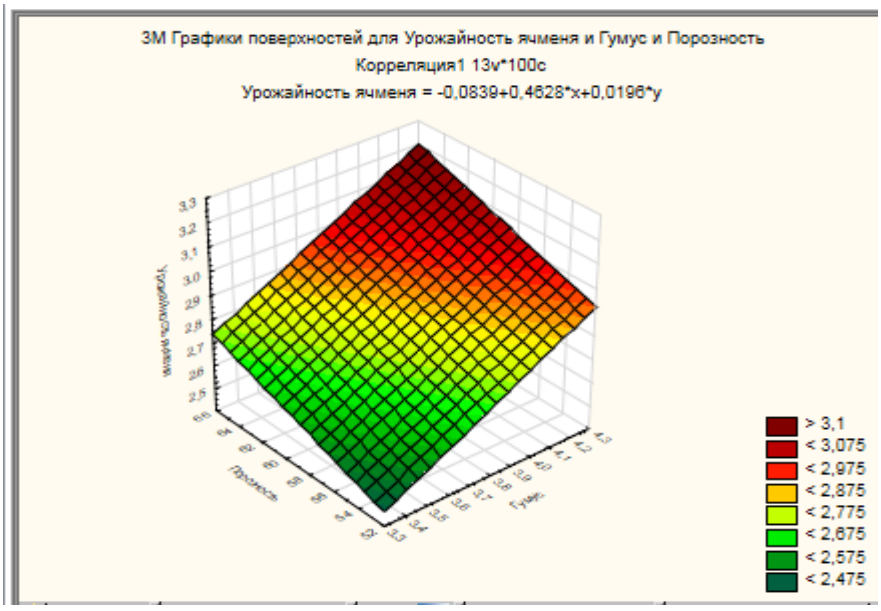


Рис. 6.49. **График поверхностей зависимости содержания гумуса, порозности и урожайности**

6.3.2 Оценка адекватности уравнения множественной регрессии

Качество модели множественной регрессии можно сравнить по нормальному вероятностному графику остатков. Остатки – это разности между опытными и предсказанными значениями зависимой переменной в построенной регрессионной модели. Чем меньше разброс значений остатков около линии регрессии тем, очевидно, лучше прогноз.

В заключении проведем оценку адекватности модели множественной регрессии по остаткам. Для этого в диалоговом окне результатов множественной регрессии (рис.6.50) активируем вкладку **Остатки/предсказанные** и нажмем на кнопку **Анализ остатков.**

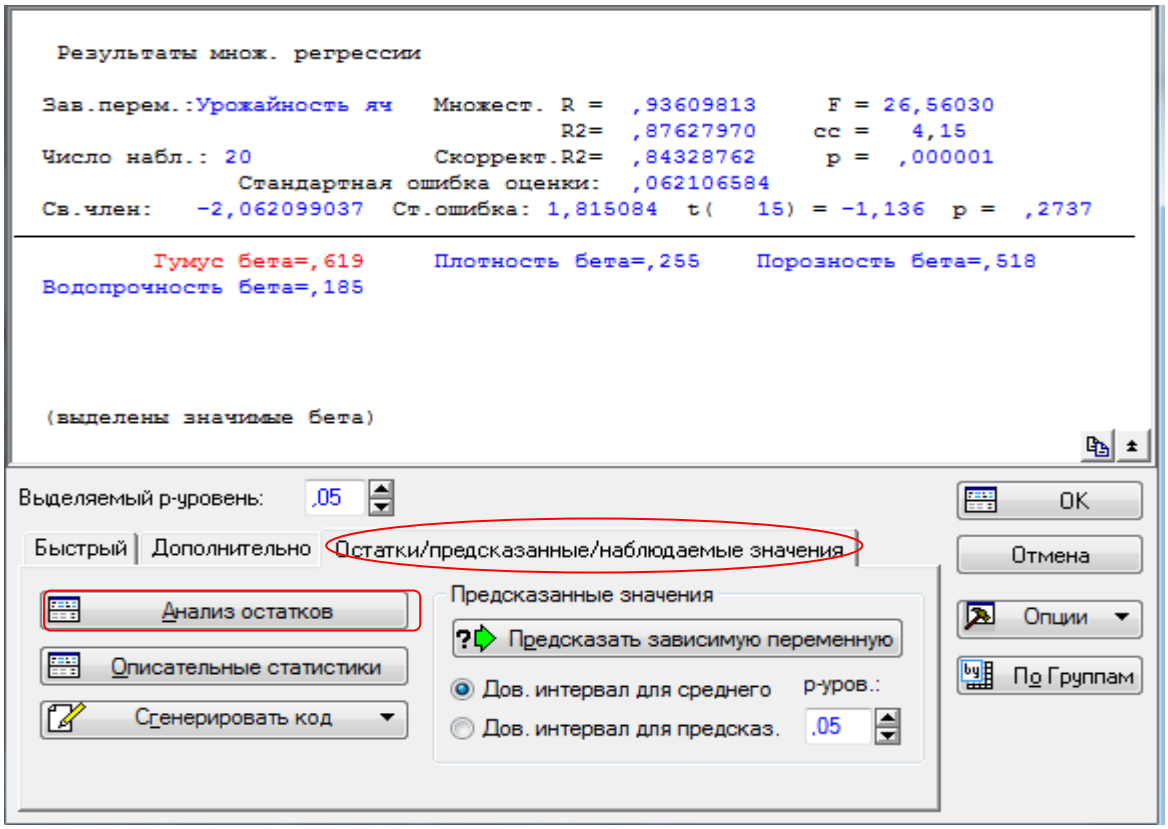


Рис. 6.50. **Диалоговое окно для выбора опции** ***Анализ остатков***

В появившемся окне (рис. 6.51) после активации вкладки **Вероятностные графики** нажмем на кнопку **Нормальный график остатков**

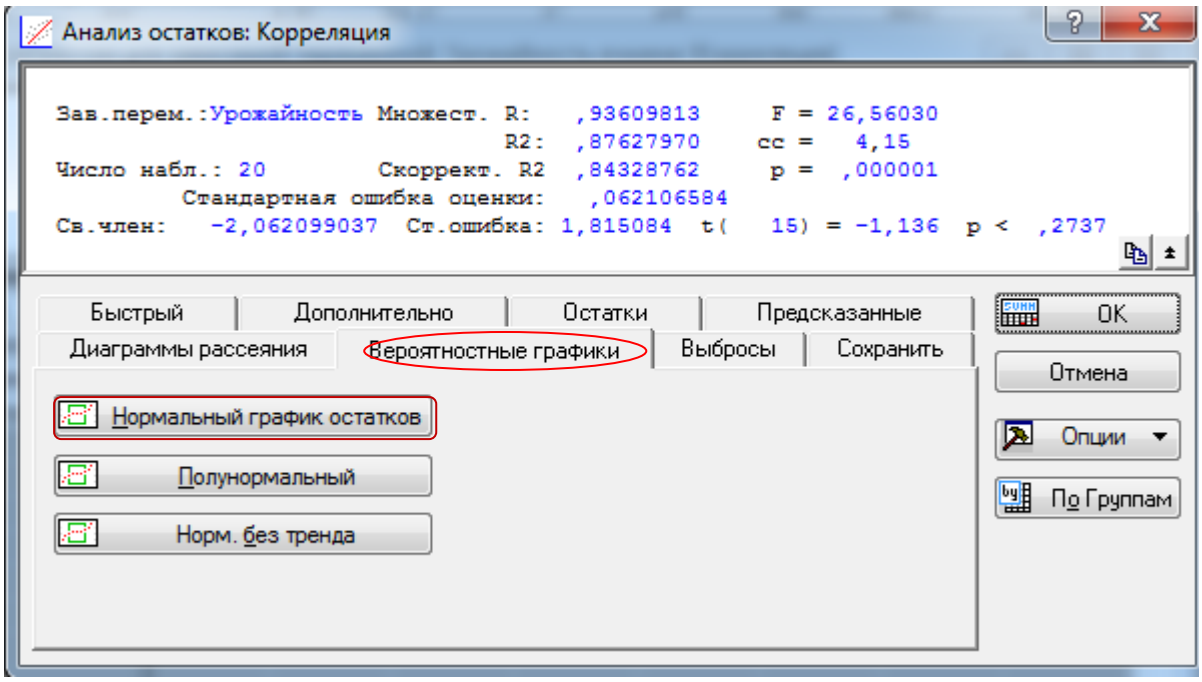


Рис. 6.51. **Диалоговое окно для построения нормального графика остатков**

После нажатия на кнопку **Ок** в рабочей книге получаем **Нормальный вероятностный график остатков.**

Ниже представлены два графика остатков: слева нормальный вероятностный график остатков множественной регрессии проведенного стандартным методом (без исключения незначимых переменных) (рис. 6.52) и справа – нормальный вероятностный график остатков после заключительного пошагового регрессионного анализа (рис. 6.53).

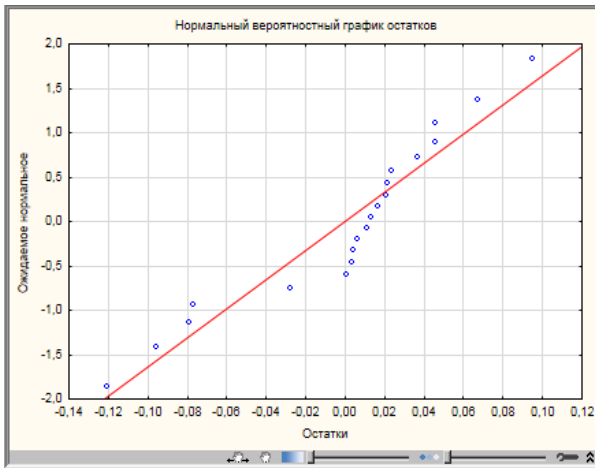


Рис. 6.52. **График остатков до исключения переменных**

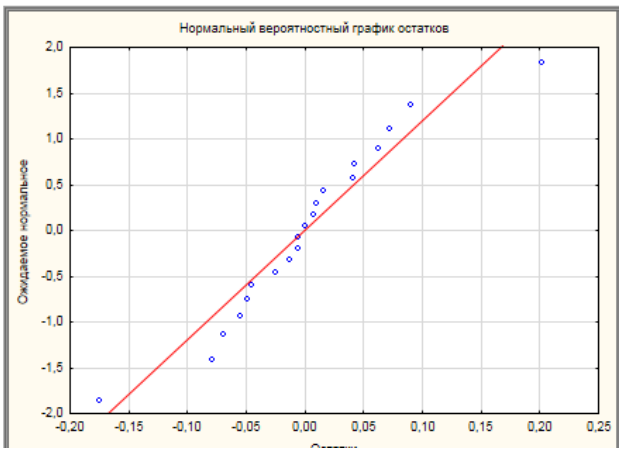


Рис. 6.53. **График остатков после исключения незначимых переменных**

Как видно из двух графиков пошаговая регрессия с исключением или же включением значимых переменных позволила описать множественную регрессию более точно. Так, если на первом графике остатков в верхней и нижней части точки приближены к прямой, а в средней части отмечается значительный разброс точек от теоретической линии регрессии, в то время как после проведения пошаговой (гребневой регрессии) эти точки имеют незначительный разброс от прямой линии.

Контрольные вопросы:

1. Независимые и зависимые признаки в агрономических исследованиях.
2. Корреляционный анализ в агрономических исследованиях.
3. Регрессионный анализ в агрономических исследованиях.
4. Виды корреляции.
5. Линейная и криволинейная зависимости.
6. Коэффициент корреляции. Что показывает коэффициент корреляции?
7. Коэффициент детерминации. Что показывает коэффициент детерминации?
8. Коэффициент регрессии. Что показывает коэффициент регрессии?
9. Как оценить качество уравнения регрессии?
10. Какие критерии служат для оценки параметров регрессии?
11. Как проводится анализ остатков?
12. В чем отличие парной и частной корреляции?
13. Как отличить прямолинейную зависимость от криволинейной?
14. Как подобрать уравнение регрессии для криволинейной зависимости?
15. Множественная корреляция. Статистические показатели множественной зависимости.
16. Множественный регрессионный анализ.
17. Бета-коэффициент. Какова роль бета-коэффициента?
18. Что такое мультиколлинеарность? Методы устранения мульти-коллинеарности.
19. Для чего проводится гребневая регрессия?

Глава 7. КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ относится к многомерной статистической процедуре, на основании которой производится разделение множества исследуемых объектов с их признаками на однородные *группы, которые называются кластерами*. С помощью кластерного анализа производится разбиение объектов на кластеры не по одному признаку, а по ряду признаков. При этом объекты, которые относятся к одному кластеру, должны быть однородными (сходными), а объекты, принадлежащие к разным кластерам, – разнородными.

Если объекты кластеризации представить как точки в *k-мерном пространстве признаков (k* – количество признаков, характеризующих объекты), то сходство между объектами определяется через понятие расстояния между *точками* X_i *и* X_j *– метрики* $d\ (X_i\ ,\ X_j)$*. В качестве метрики чаще всего используется евклидово расстояние или квадрат евклидового расстояния.* Евклидово расстояние – наиболее популярная метрика, является геометрическим расстоянием в многомерном пространстве. Чем меньше расстояние между объектами, тем они более схожи.

Кластерный анализ в отличие от большинства статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет исследовать множество исходных данных практически произвольной природы.

Основные этапы кластерного анализа:

1. Отбор объектов для кластеризации.
2. Выбор множества переменных, по которым будет производиться кластеризация и описание объектов по этим переменным.
3. Определение меры сходства или меры различия между объектами в соответствии с избранной метрикой.
4. Распределение объектов в кластеры с помощью той или иной процедуры объединения.
5. Проверка качества выбранного метода кластерного анализа.

В пакете Statistica предлагается 3 метода кластерного анализа

1. Иерархический кластерный анализ (tree clustering);

2. Метод k средних (k-means clustering).

3. Двухвходовое объединение (two-way joining);

Пример 1. В многолетнем многофакторном опыте изучаются 5 вариантов систем обработки почвы: обычная, глубокая, плоскорезная, минимальная и сочетание обычной обработки с минимальной (сочетание ОМ) и на каждом варианте систем обработки почвы 3 варианта удобрений: без удобрений, NPK и NPK +навоз, всего 15 вариантов или объектов. Все объекты описываются 10-ью признаками, к которым относятся: агрохимические показатели (содержание гумуса, азота, P_2O_5, K_2O, гидролизуемого азота, pH), агрофизические показатели (плотность, твердость, водопрочная структура почвы) и интегральный показатель плодородия почвы – урожайность озимой пшеницы.

Ввод данных

В программе Statistica создадим файл исходных данных «Кластер_16». Введем в таблице исходных данных в качестве переменных 10 признаков, а наблюдений по строкам – 15 объектов: системы обработки почвы в сочетании с удобрениями, как представлено ниже (рис. 7.1).

	1 Обработка	2 Гумус	3 Азот	4 pH	5 $P_2 0_5$	6 $K_2 0$	7 Гидролизуемый азот	8 Плотность почвы	9 Твердость почвы	10 Структура почвы	11 Урожай
1	Обычная без удоб	3,21	0,12	4,12	11,5	9,5	10,6	1,32	18,3	29,6	20,5
2	Обычная с NPK	3,19	0,22	4,35	15,7	10,6	15,8	1,34	21,3	31,2	41,5
3	Обычная с NPK+ навоз	3,61	0,32	5,04	21	14,2	23,4	1,2	15,4	39,5	56,8
4	Глубокая без удоб	2,98	0,1	4	10,6	9	8,4	1,38	19,5	25,3	19,6
5	Глубокая с NPK	3,01	0,16	4,2	16,3	11,4	11,5	1,38	19,4	28,4	38,7
6	Глубокая с NPK+ навоз	3,52	0,28	4,98	19,7	14,6	19,7	1,31	16,7	30,2	49,1
7	Плоскорезная без удоб	2,98	0,14	4,1	10,3	8,8	9,3	1,4	21,2	26,7	16,4
8	Плоскорезная с NPK	3,01	0,19	4,16	18,6	9,4	13,6	1,36	20,8	29,8	39,5
9	Плоскорезная с NPK+ навоз	3,45	0,29	4,98	20,1	12,5	17,7	1,29	17,7	34,3	46,8
10	Минимальная без удоб	3,33	0,2	4,52	12,4	13,1	12,9	1,36	18	31,6	22,1
11	Минимальная с NPK	3,41	0,23	4,63	19,6	14,8	20,4	1,3	18,4	34,9	49,9
12	Минимальная с NPK+ навоз	3,8	0,45	5,51	25,4	19,6	25,3	1,24	13,6	41,1	61,3
13	Сочетание ОМ без удоб	3,41	0,23	4,52	12	12,9	12,4	1,34	16,5	32,9	23,1
14	Сочетание ОМ с NPK	3,46	0,28	4,98	21	15,8	21,3	1,29	17,2	38,7	51,3
15	Сочетание ОМ с NPK+ навоз	3,96	0,49	5,98	28,7	19,9	24,8	1,23	11	43,3	64,8

Рис. 7.1. **Таблица исходных данных**

Так как изучаемые признаки измеряются в разных шкалах с различными диапазонами: содержание азота измеряется в десятых долях, а структура почвы

и урожайность изменяются от 20 до 50 единиц, то и вклад этих переменных в евклидовое расстояние будет разным, прежде всего из-за разных единиц измерения. Поэтому, перед классификацией данных, их необходимо стандартизировать с тем, чтобы привести все переменные к единой шкале.

Для стандартизации переменных выберем в строке **Меню Данные** и нажмем на опцию **Стандартизировать (Standartize)** (рис. 7.2).

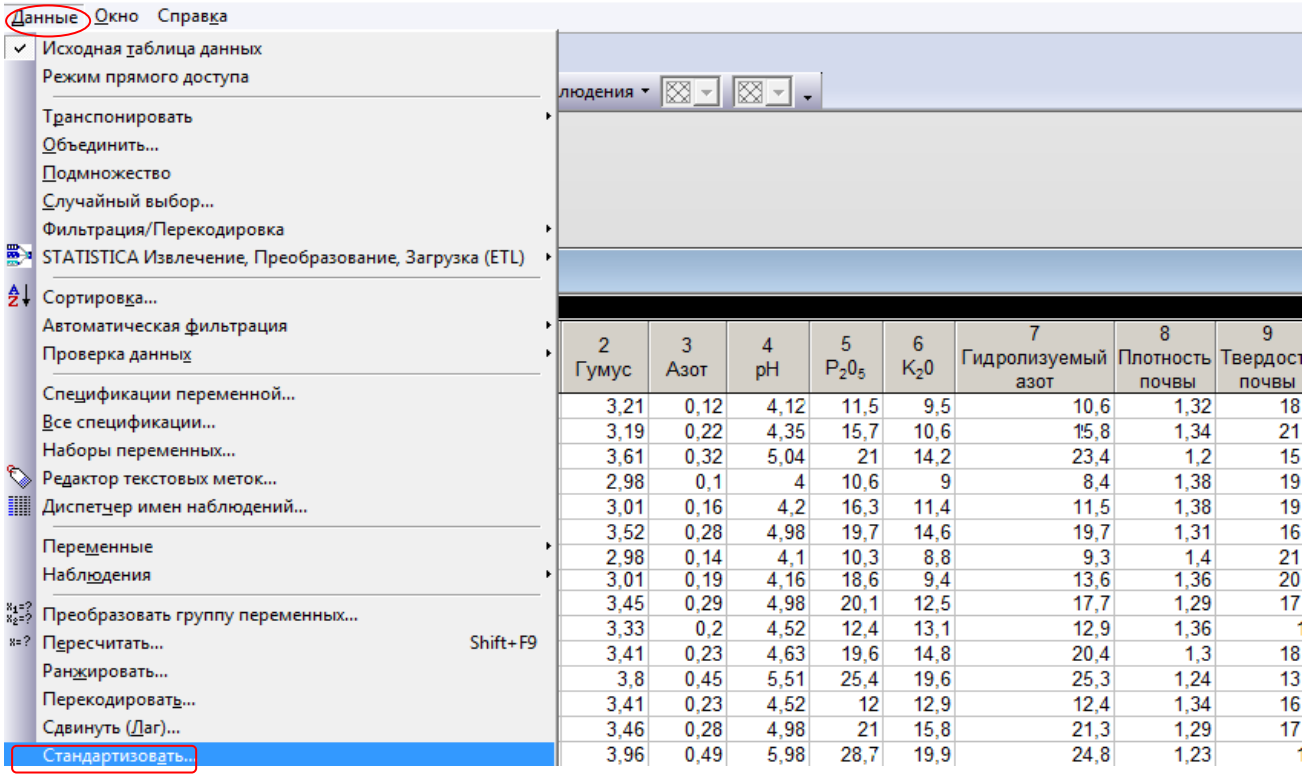


Рис. 7.2. **Рабочая панель для выбора опции *Стандартизация***

В появившемся окне **Стандартизация значений** (Рис. 7.3.) нажмем на клавишу **Переменные,** укажем диапазон переменных для стандартизации от №2 (Гумус) до №11 (Урожай), и после нажатия на клавишу **Ок** получим таблицу со стандартизированными данными. Сохраним файл со стандартизированными данными как «Кластер_стан_16».

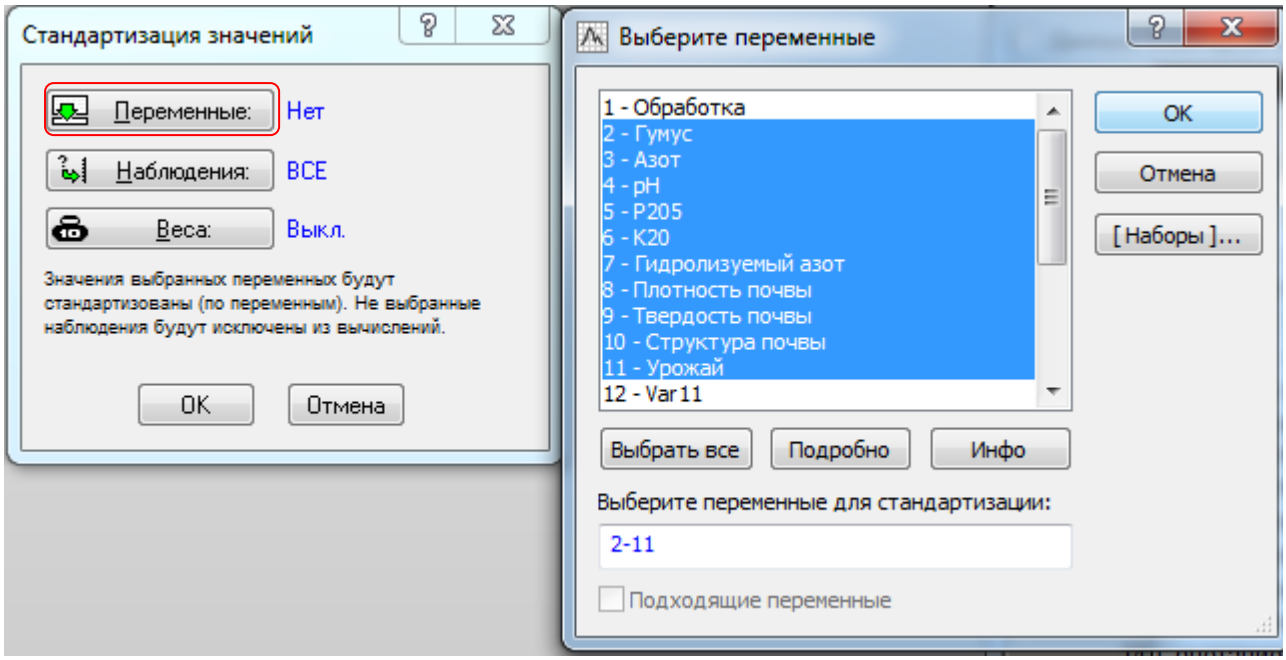


Рис. 7.3. **Диалоговое окно выбора переменных для стандартизации**

Таблица со стандартизованными переменными «Кластер_стан_16» приведена ниже (рис.7.1.1).

	1 Обработка	2 Гумус	3 Азот	4 pH	5 $P_2 0_5$	6 $K_2 0$	7 Гидролизуемый азот	8 Плотность почвы	9 Твердость почвы	10 Структура почвы	11 Урожай
1	Обычная без удоб	-0,4863	-1,139	-0,9668	-1,093	-1,015	-1,02487104	0,067918	0,224365	-0,6620579	-1,21029
2	Обычная с NPK	-0,5532	-0,2398	-0,5635	-0,331	-0,702	-0,117493729	0,4075078	1,2871465	-0,36506	0,08689
3	Обычная с NPK+ навоз	0,85208	0,65944	0,64646	0,6297	0,3199	1,20867311	-1,969621	-0,80299	1,17561692	1,031977
4	Глубокая без удоб	-1,2558	-1,3189	-1,1772	-1,256	-1,157	-1,40876144	1,0866875	0,6494776	-1,46024	-1,26588
5	Глубокая с NPK	-1,1554	-0,7793	-0,8265	-0,222	-0,475	-0,867824967	1,0866875	0,6140515	-0,8848064	-0,08607
6	Глубокая с NPK+ навоз	0,55095	0,29975	0,54125	0,394	0,4335	0,563039255	-0,101877	-0,342452	-0,5506837	0,556345
7	Плоскорезная без удоб	-1,2558	-0,9592	-1,0018	-1,31	-1,214	-1,25171537	1,4262773	1,2517205	-1,2003667	-1,46355
8	Плоскорезная с NPK	-1,1554	-0,5096	-0,8966	0,1946	-1,043	-0,50138413	0,7470976	1,1100163	-0,6249332	-0,03665
9	Плоскорезная с NPK+ навоз	0,31674	0,38967	0,54125	0,4665	-0,163	0,214047981	-0,441467	0,0118087	0,21037355	0,414273
10	Минимальная без удоб	-0,0848	-0,4196	-0,2654	-0,929	0,0076	-0,623531075	0,7470976	0,1180868	-0,2908105	-1,11145
11	Минимальная с NPK	0,18291	-0,1499	-0,0725	0,3759	0,4903	0,6851862	-0,271672	0,259791	0,32174779	0,605761
12	Минимальная с NPK+ навоз	1,4878	1,82845	1,47061	1,4274	1,8534	1,54021482	-1,290441	-1,440659	1,47261488	1,309943
13	Сочетание ОМ без удоб	0,18291	-0,1499	-0,2654	-1,002	-0,049	-0,710778894	0,4075078	-0,413304	-0,0494997	-1,04968
14	Сочетание ОМ с NPK	0,3502	0,29975	0,54125	0,6297	0,7743	0,842232273	-0,441467	-0,165322	1,02711794	0,69224
15	Сочетание ОМ с NPK+ навоз	2,02314	2,18814	2,29475	2,0257	1,9386	1,452967	-1,460236	-2,361737	1,88098707	1,52614

Рис. 7.1.1. **Таблица исходных данных**

В последующим все методы кластерного анализа будем проводить со стандартизированными данными «Кластер_стан_16».

В меню **Анализ (Statistics)** выберем **Многомерный разведочный анализ (Multivariate Exploratory Analysis)**, а затем модуль **Кластерный анализ (Cluster Analysis)** (рис. 7.4.)

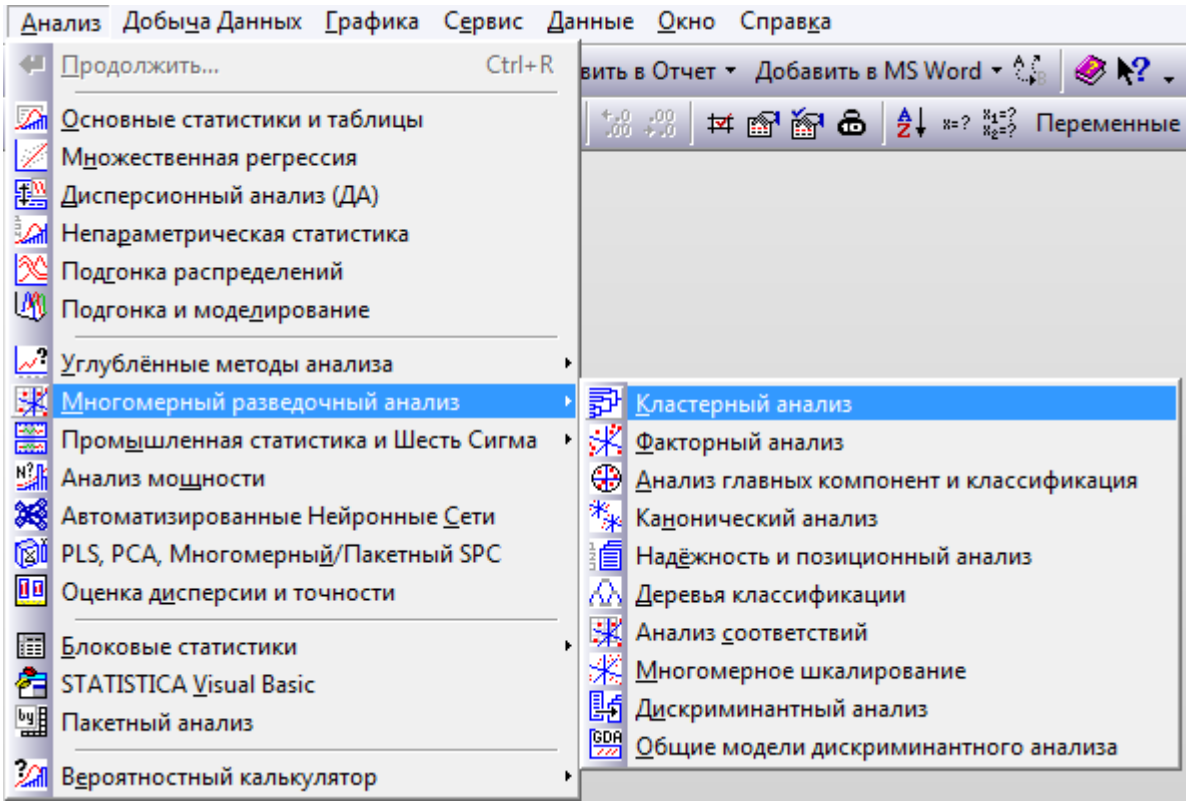


Рис. 7.4. **Диалоговое окно выбора модуля *Кластерный анализ***

Появляется стартовый модуль **Методы кластеризации (Clustering Method),** в котором предлагается 3 метода кластеризации (рис. 7.5)**: Иерархическая классификация (Joining tree clustering), Кластеризация методом К – средних (k-means clustering) и (Двухвходовое объединение (Two-way joining).**

Выберем метод **Иерархическая классификация (Joining tree clustering).**

7.1 Иерархический кластерный анализ (tree clustering)

При выборе данного метода строятся *дендрограммы (от греческого dendron* - "дерево") – дерево объединения кластеров. Дендрограмма описывает близость отдельных точек и кластеров друг к другу и показывает в графическом виде последовательность объединения (разделения) кластеров. Дендрограмма может быть представлена как в горизонтальной, так и вертикальной плоскости. Существует два метода иерархической кластеризации: *агломеративный* и *дивизивный*. Иерархическая агломеративная процедура

кластеризации приводит к последовательному объединению объектов сначала самых близких, а затем всё более отдалённых друг от друга. При иерархической дивизивной процедуре осуществляется последовательное разделение групп объектов сначала самых далёких, а затем всё более близких друг от друга.

Важным этапом кластеризации является выбор процедуры объединения (правило объединения) и выбор расстояния или метрики.

Существуют следующие процедуры объединения (Amalgamation [linkage] rule):

1. Метод одиночной связи (Single Linkage) – принцип ближайшего соседа
2. Метод полной связи (Complete Linkage) – принцип дальнего соседа
3. Невзвешенное попарное среднее (Unweighted pair-group average)
4. Взвешенное попарное среднее (Weighted pair-group average)
5. Невзвешенный центроидный метод (Unweighted pair-group centroid)
6. Взвешенный центроидный метод (Weighted pair-group centroid)
7. Метод Уорда (Ward's method).

В программе Statistica предлагаются следующие меры расстояния (метрики):

1. Евклидово расстояние (Euclidean distances)
2. Квадрат евклидова расстояния (Squared Euclidean distances)
3. Манхэттенское расстояние (City-block (Manhattan) distance)
4. Расстояние Чебышева (Chebychev distance metric)
5. Процент несогласия (Percent disagreement)
6. 1 – коэффициент корреляции Пирсона

Проведем кластерный анализ с использованием наиболее распространенного метода – иерархического агломеративного метода.

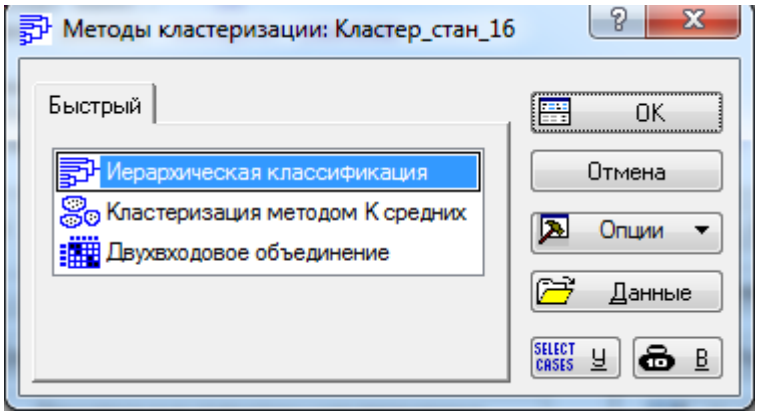


Рис. 7.5. **Стартовый модуль методов кластерного анализа**

В окне выбора переменных выберем признаки для классификации – переменные, начиная с переменной №2 – гумус и заканчивая переменной №11 – урожай. Переменная №1 – обработка включает в себя по строкам объекты для кластеризации – наименование систем обработки почвы в сочетании с удобрениями.

Открываем вкладку **Дополнительно (Advanced)** и в поле **Объекты (Cluster)** выберем **Наблюдения (строки) (Cases-rows)** (рис. 7.6).

Окошко **Правило объединения (Amalgamation linkage)** содержит установки для выбора мер сходства между объектами, выберем **Метод одиночной связи (Single linkage).** В окошке **Мера расстояния (Distance measure)** предлагаются различные виды расстояний, выберем мера близости **Евклидовое расстояние (Euclidean distances)**

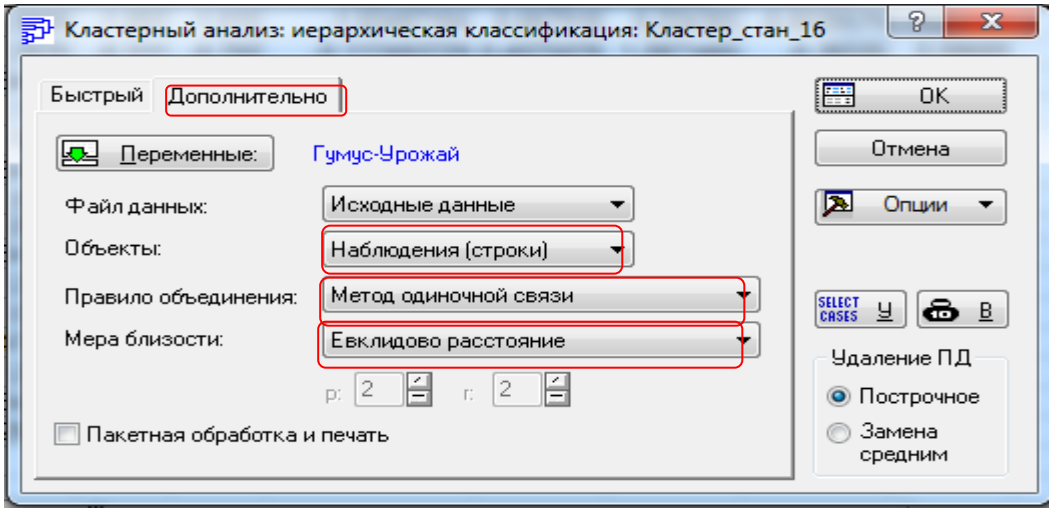


Рис. 7.6. **Диалоговое для выбора процедуры объединения и метрики расстояния**

После установки всех необходимых параметров для проведения кластерного анализа щелкнем по кнопке **Ok,** появится расширенное диалоговое окно с результатами иерархической классификации (рис. 7.7). В верхней части окна представлена информация о количестве наблюдений (объектов – 15) и переменных (10), а также выбранные правило объединения и мера близости (метрика расстояния). Ниже располагаются функциональные кнопки, с помощью которых можно просмотреть разные варианты результатов иерархической классификации.

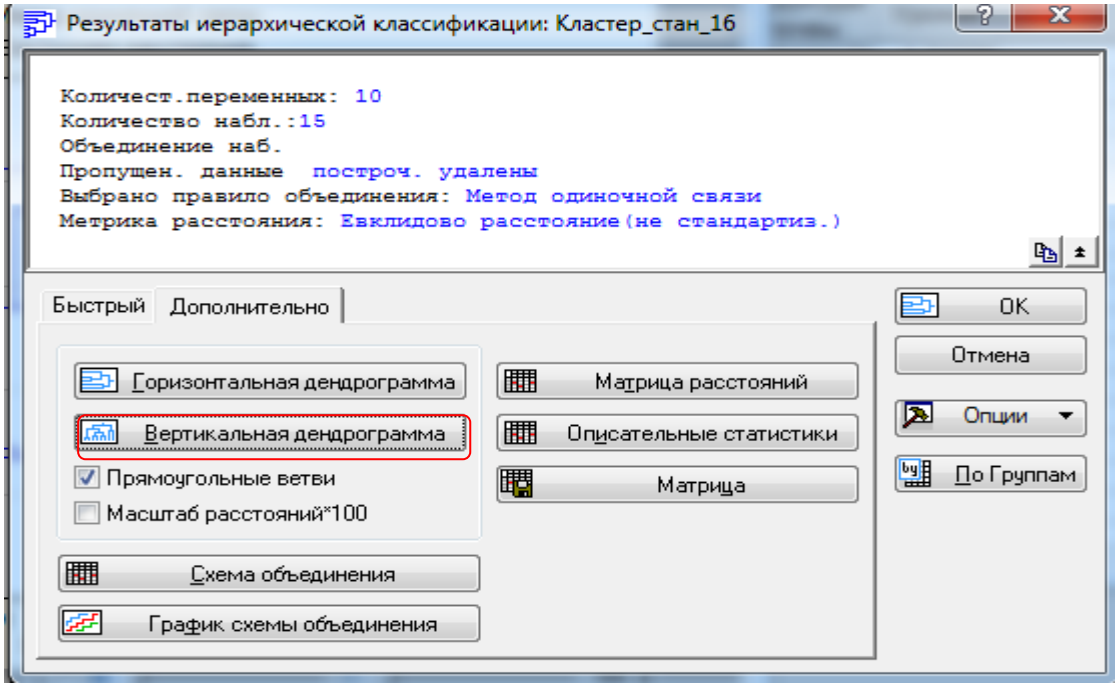


Рис. 7.7. **Расширенное диалоговое окно результатов иерархической классификации**

Для построения дендрограммы нажмем на кнопку **Вертикальная дендрограмма (Vertical icicle plot),** отметим **Прямоугольные ветви** и после нажатия на кнопку **Ок** получаем вертикальную дендрограмму для 15 объектов методом одиночной связи (рис. 7.8).

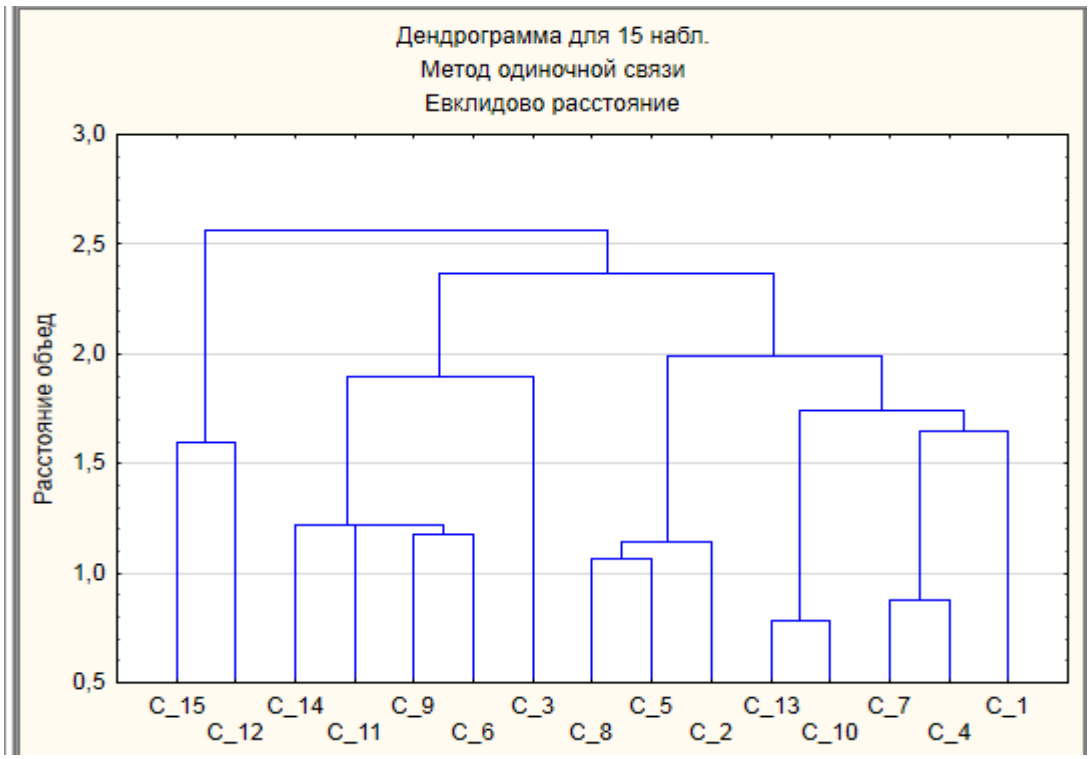


Рис. 7.8. **Вертикальная дендрограмма иерархического агломеративного кластерного анализа для 15 объектов**

Для замены номеров строк на наименование изучаемых объектов (систем обработок почвы в сочетании с удобрениями) наведем мышку на график вертикальной дендрограммы (рис. 7.8) и дважды щелкнем на правую клавишу мышки.

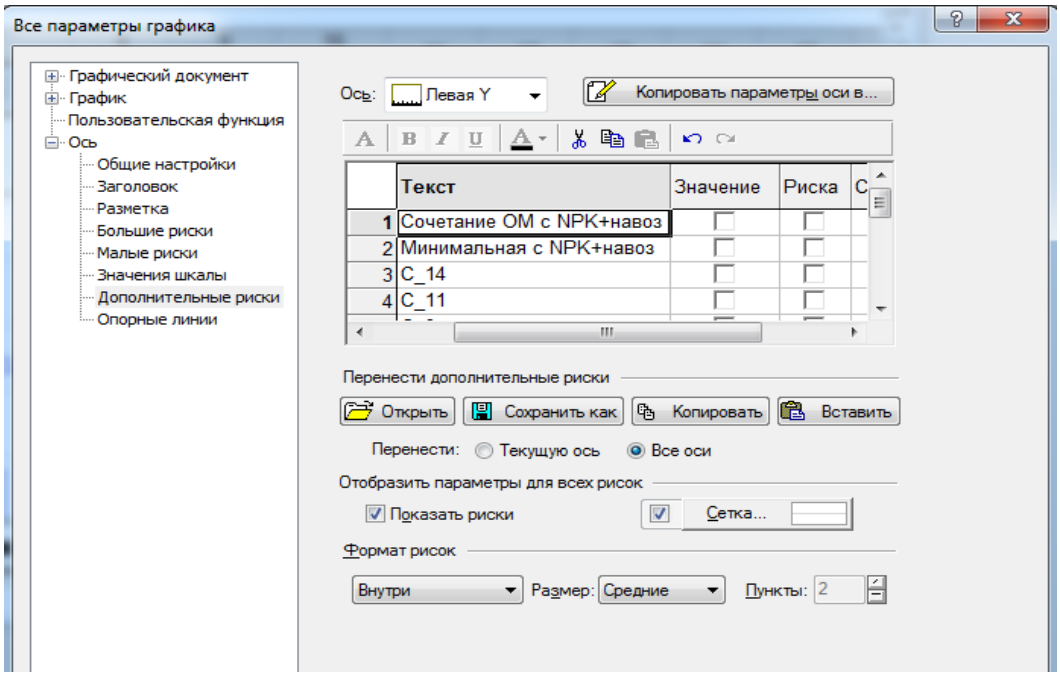


Рис. 7.9. **Редактирование параметров дендрограммы**

В появившемся окне активируем параметр графика «**Дополнительные риски**» и в поле **Текст** последовательно заменяем номера строк на наименование систем обработки почвы и удобрений: С_15 на Сочетание ОМ с NPK+навоз, С_12 на Минимальная с NPK+навоз и т.д.(рис. 7.9).

После заверешения редактирования параметров графика нажмем на кнопку **Ок** и получаем вертикальную дендрограмму с наименованниями изучаемых систем обработки почвы и удобрений (рис. 7.10.).

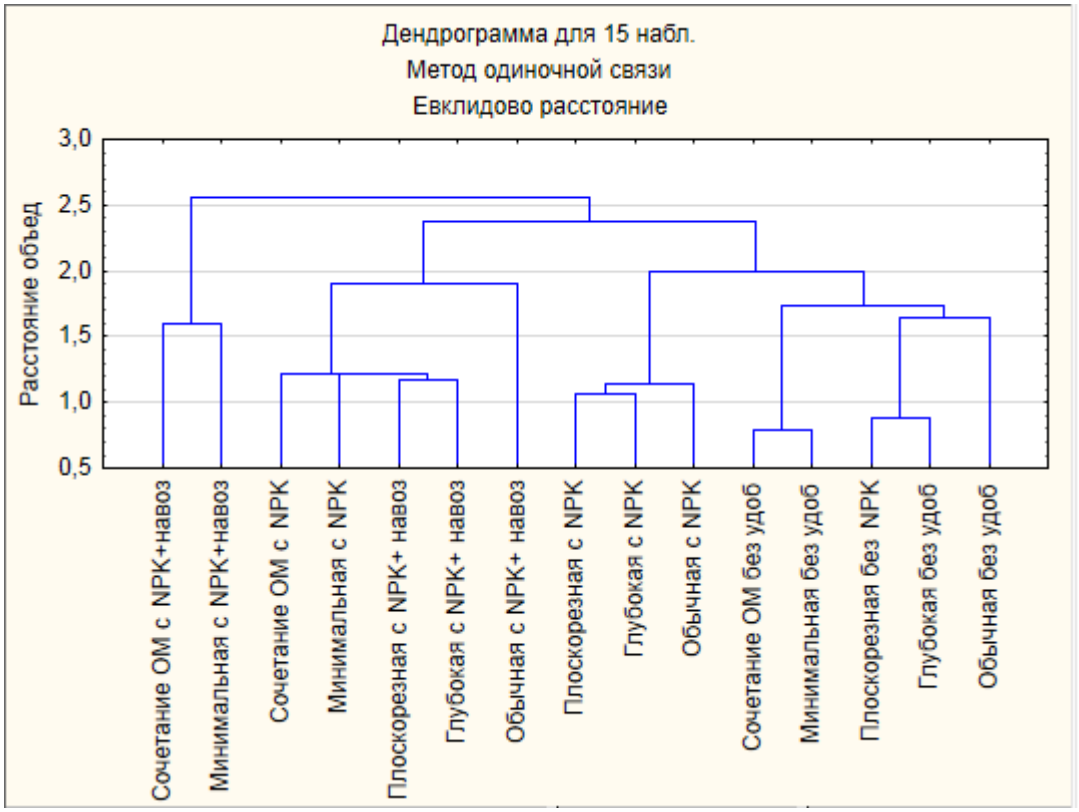


Рис. 7.10. **Дендрограмма для систем обработки почвы и удобрений**

На графике по оси абсцисс отложены объекты – разные системы обработки почвы и удобрений, по оси ординат – евклидово расстояние между объектами. Процесс объединения объектов (систем обработок почвы в сочетании с удобрениями) в кластеры происходит на основании сходства (близости) по евклидовому расстоянию. Так, в первый кластер объединяются минимальная система обработки почвы без удобрений и сочетание обычной с минимальной без удобрений (евклидовое расстояние – 0,7875), второй кластер – плоскорезная без удобрений и глубокая без удобрений (евклидовое расстояние – 0,8806), третий кластер – плоскорезная с NPK, глубокая с NPK и

обычная с NPK. Четвертый кластер объединяет плоскорезную систему обработки в сочетании с NPK+ навоз, глубокую с NPK и глубокую в сочетании с NPK +навоз, пятый кластер – минимальную с NPK и сочетание ОМ с NPK. Объекты с самыми высокими значениями по большинству агрохимических и агрофизических показателей плодородия почвы и урожайности: минимальная система обработки почвы и сочетание ОМ с NPK+ навоз входят в восьмой кластер. При этом в правой части расположены кластеры, которые объединяют системы обработки почвы без удобрений, в то время как в левой части – системы обработки почвы с NPK+навоз.

Данные дендрограммы (рис. 7.10) четко показывают о слиянии мелких кластеров в три крупных кластера, в которые входят следующие системы обработки почвы и удобрения:

I – обычная без удобрений, глубокая без удобрений, плоскорезная без удобрений, минимальная без удобрений, сочетание ОМ без удобрений, обычная с NPK, глубокая с NPK и плоскорезная с NPK;

II– обычная с NPK+навоз, глубокая с NPK+ навоз, плоскорезная с NPK+навоз, минимальная с NPK, сочетание ОМ с NPK;

III– минимальная с NPK+навоз, сочетание ОМ с NPK+навоз.

О последовательности объединения изучаемых обработок почвы в сочетании с удобрениями в кластеры можно судить, если в расширенном окне результатов иерархической классификации (рис. 7.7) выбрать опцию **Схема объединения (Amalgamation schedule).** После нажатия на кнопку **Ок** получаем таблицу, в которой представлена (рис. 7.11) пошаговая схема объединения систем обработки почвы и удобрений. В первом столбце указаны расстояния объединения (Евклидовое расстояние). Каждая строка показывает состав кластера на данном шаге классификации. Процесс кластеризации завершается за m – 1 шагов, когда в итоге все объекты будут объединены в один кластер, в нашем случае 15-1 = 14 шагов.

Результаты данной таблицы дополняют представленную выше дендрограмму (рис.7.8.)

расст. объедин.	Схема объединения (Кластер_стан_16) Метод одиночной связи Евклидово расстояние **Объект 1**	Объект 2	Объект 3	Объект 4	Объект 5	Объект 6	Объект 7	Объект 8	Объект 9	Объект 10	Объект 11	Объект 12	Объект 13	Объект 14	Объект 15
,7875684	C_10	C_13													
,8806806	C_4	C_7													
1,067561	C_5	C_8													
1,147125	C_2	C_5	C_8												
1,177215	C_6	C_9													
1,217314	C_6	C_9	C_11												
1,221108	C_6	C_9	C_11	C_14											
1,599597	C_12	C_15													
1,649783	C_1	C_4	C_7												
1,738869	C_1	C_4	C_7	C_10	C_13										
1,900632	C_3	C_6	C_9	C_11	C_14										
1,993361	C_1	C_4	C_7	C_10	C_13	C_2	C_5	C_8							
2,369678	C_1	C_4	C_7	C_10	C_13	C_2	C_5	C_8	C_3	C_6	C_9	C_11	C_14		
2,565353	C_1	C_4	C_7	C_10	C_13	C_2	C_5	C_8	C_3	C_6	C_9	C_11	C_14	C_12	C_15

Рис. 7.11. **Схема объединения**

7.2 Кластеризация методом k-средних

Кластеризация методом k-средних (k-means) относится к итеративному методу кластерного анализа, суть которого заключается в том, что исследователь может задать нужное количество кластеров, чтобы они были настолько различны, насколько это возможно. До сих пор нет определенного критерия определения заданного количества кластеров. В принципе можно задавать любое количество кластеров в интервале от *1* до *m–1*, где *m* – количество объектов. Вместе с тем, рекомендуются следующие решения:

- можно использовать результаты предшествующих исследований;
- ориентироваться на результаты иерархического кластерного анализа;
- менять число образуемых кластеров и выбирать наилучшее разбиение по задаваемому критерию качества.

Алгоритм кластеризации методом k-средних заключается в том, что вначале программа определяет центры запланированных кластеров, затем вычисляется расстояние между центрами кластеров и каждым объектом, далее объект приписывается к тому кластеру, к которому он ближе всего. В завершении вычисляются средние значения каждого кластера, число средних равно числу анализируемых переменных. Набор средних представляет собой координаты нового положения центра кластера. Сходство и разнородность объектов определяется посредствам m-мерного евклидового расстояния между векторами измерений.

Когда результаты кластерного анализа получены, можно рассчитать средние для каждого кластера по каждому признаку (переменной), чтобы оценить, насколько кластеры различаются друг от друга. В идеале мы должны получить сильно различающиеся средние для всех или большинства переменных, используемых в кластерном анализе. Оценить качество кластеризации можно с помощью дисперсионного анализа (ANOVA). Значения *F*-критерия и уровня значимости (*p*), полученные для каждого измерения, являются важным индикатором того, насколько хорошо соответствующее измерение дискриминирует кластеры.

Для кластеризации данных нашего примера (объекты – 15 систем обработки почвы в сочетании с удобрениями и 10 признаком) выберем в стартовом модуле (рис. 7.5.1) опцию **Кластеризация методом К средних (K-means clustering)**

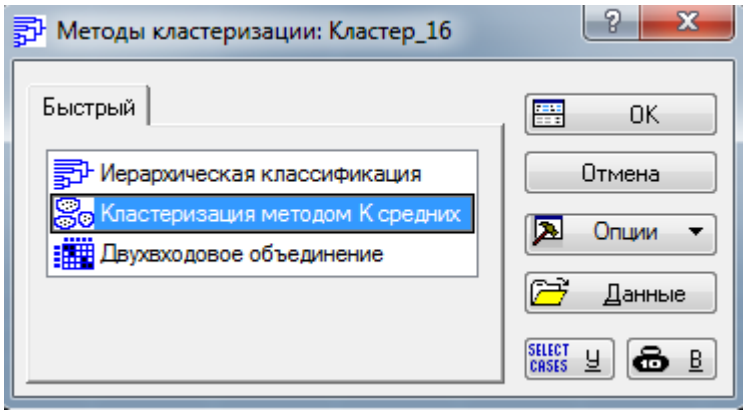


Рис. 7.5.1. **Стартовый модуль методов кластерного анализа**

После нажатия на кнопку **Ок** появляется диалоговое окно для выбора параметров кластеризации (рис. 7.12), в котором откроем вкладку **Дополнительно (Advanced).** Так как в методе k-средних в качестве метрики используют евклидову метрику, то данные предварительно необходимо стандартизировать, поэтому в окне выбора переменных укажем данные таблицы со стандартизированными переменными файла «Кластер_стан_16».

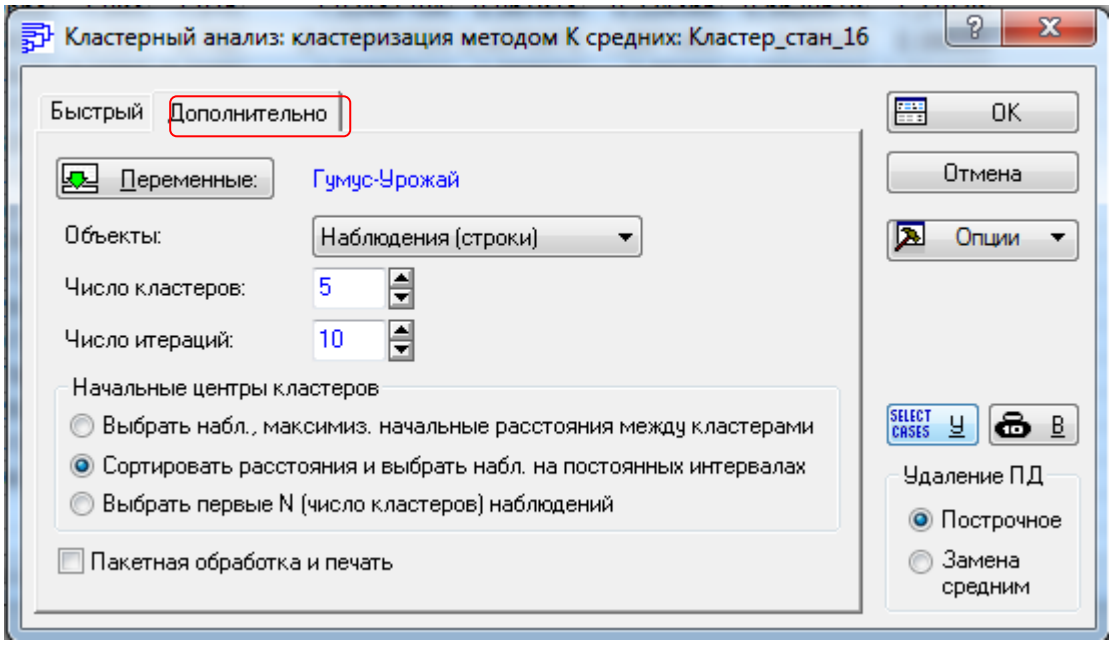


Рис. 7.12. **Диалоговое окно выбора параметров кластеризации**

Задача кластерного анализа сводится к разбиению 15 объектов (5 систем обработки почвы в сочетании с 3-мя вариантами удобрений) на m кластеров с тем, чтобы:

- каждая система обработки почвы в сочетании с удобрением должна принадлежать одному и только одному кластеру;
- системы обработки почвы, принадлежащие одному и тому же кластеру, должны быть сходными;
- системы обработки почвы, принадлежащие разным кластерам, должны быть разнородными.

Так как в нашем опыте изучается 5 систем обработки почвы в сочетании с 3-мя вариантами удобрений, естественно, напрашиваются две гипотезы: первая, что разные системы обработки почвы распределятся в 5 групп (кластеров) и вторая – разные системы обработки распределятся в 3 кластера, в каждом из них будут разные удобрения. С целью проверки этих гипотез проведем две кластеризации с m = 5 и m=3. Выполняя последовательное разбиение на различное число кластеров, можно сравнивать качество получаемых решений.

Для проверки первого предположения в поле выбора **числа кластеров (number of clusters)** укажем 5, **число итераций (number of iterations)** оставляем по умолчанию – 10 и нажмем на кнопку **Ок** (рис.7.12).

После проведения вычислений появится диалоговое окно результатов кластеризации методом **k-средних (k-Means Clustering Results)**, в верхней части которого представлена информация о количестве переменных и наблюдений (объектов), о количестве итераций для получения запланированного числа кластеров (рис. 7.13). В нижней части окна расположены кнопки для вывода различной информации по кластерам.

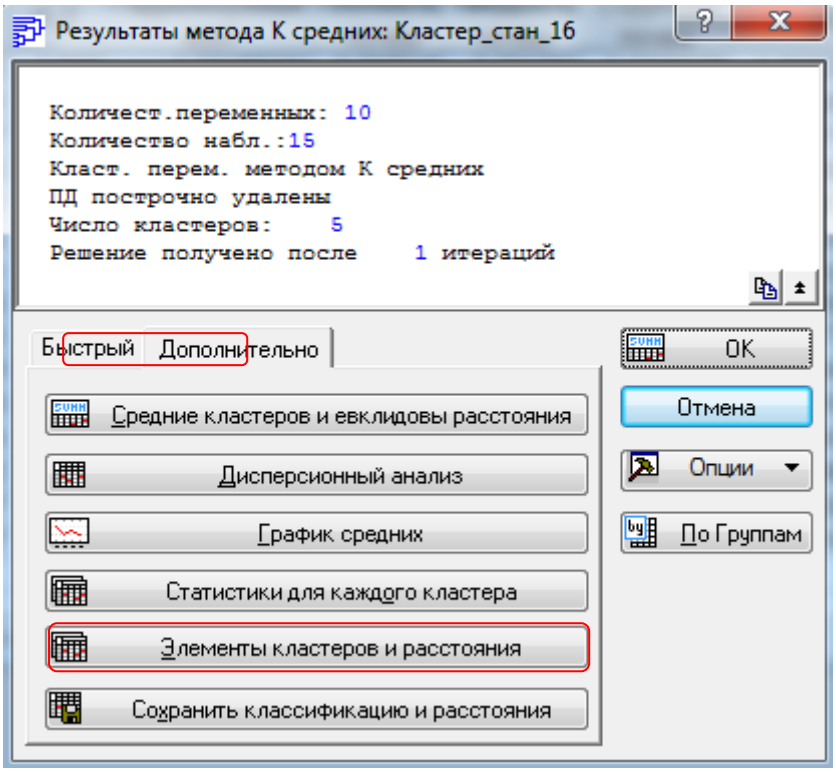


Рис. 7.13. **Расширенное диалоговое окно результатов кластеризации**

Для визуализации и удобства оценки параметров получаемых кластеров определить какие наблюдения (объекты) входят в запланированные кластеры – из каких элементов состоит тот или кластер.

Откроем вкладку **Дополнительно** и нажмем на кнопку **Элементы кластеров и расстояния,** появится рабочая книга с элементами пяти кластеров. Для того, чтобы понять, какие системы обработок почвы входят в тот или иной кластер проведем их идентификацию. Открываем последовательно каждый кластер и в колонке **Наблюдения** последовательно заменяем номера строк на наименование систем обработки почвы и удобрений, например в пятом кластере: С_1 на Обычная без удобрений, С_4 на Глубокая без удобрений и т.д.

и получаем удобные для последующего анализа кластеры не с номерами строк, а с наименованиями систем обработок почвы и удобрений (рис. 7.14).

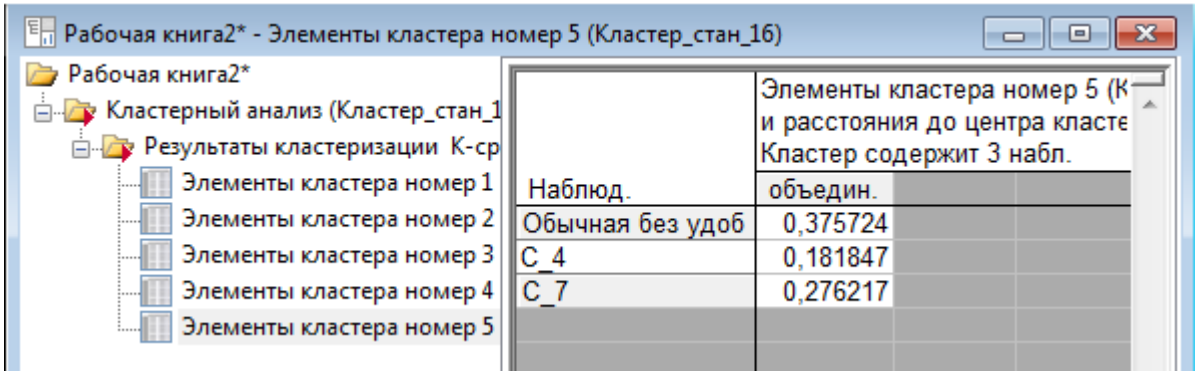


Рис. 7.14. **Рабочая книга с 5-тью таблицами элементов кластеров**

После замены во всех кластерах номеров строк на анализируемые системы обработок почвы получаем 5 таблиц, в каждой из которых приводятся перечень систем обработок, вошедших в тот или иной кластер, и величины их евклидового расстояния до центра кластера (рис. 7.15). Наиболее схожими оказались кластеры № 2 и №3, в которые вошли минимальная система и сочетание ОМ без удобрений и с полным набором удобрений, о чем свидетельствуют минимальные значения евклидового расстояния.

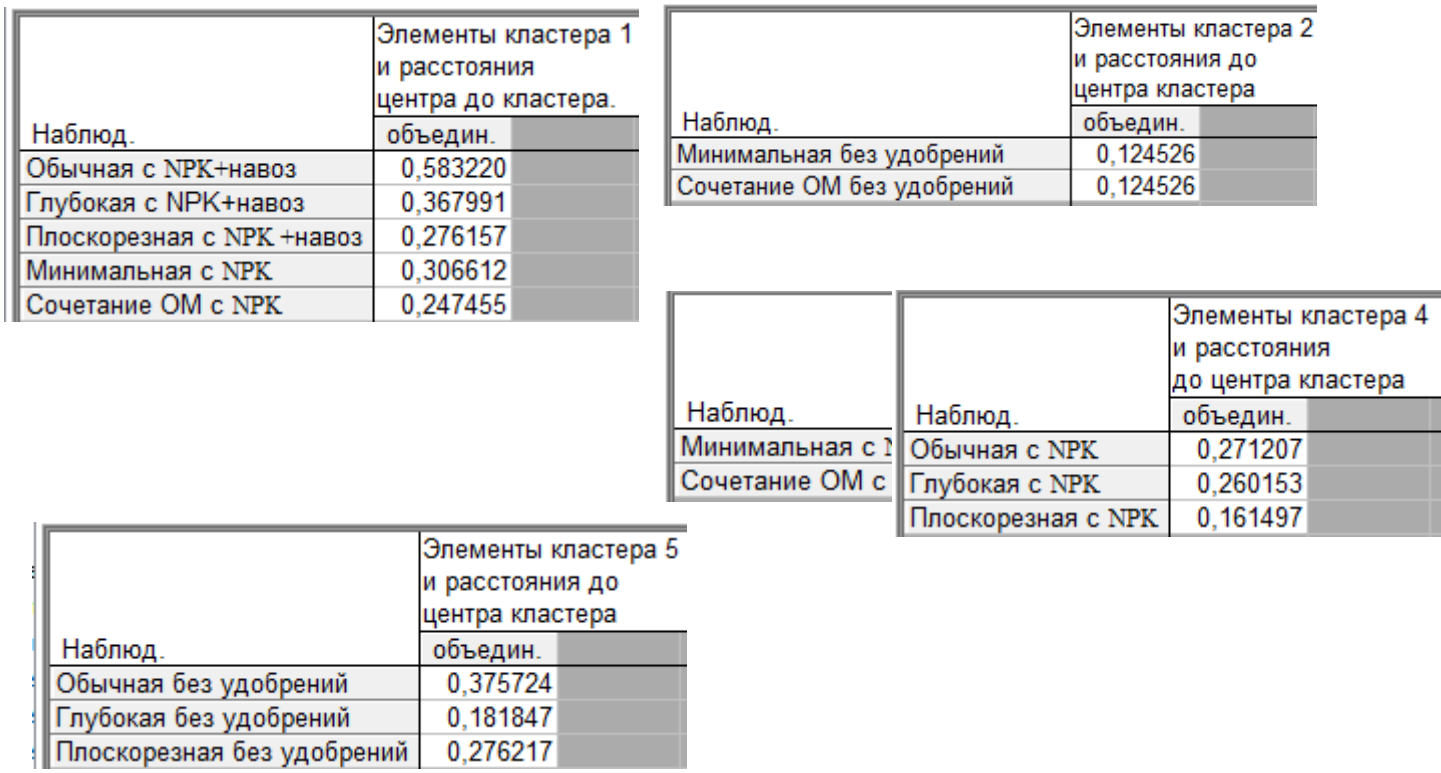


Рис. 7.15. **Результаты итоговой кластеризации методом k-средних 5-ти кластеров**

Для анализа разнородности полученных пяти кластеров в расширенном окне результатов кластеризации нажмем кнопку **Средние кластеров и евклидовые расстояния** и получим следующую таблицу (рис. 7.16):

Кластер Номер	Евклидовы расст. между кластерами (Кластер_стан_16) Расстояния под диагональю Квадраты расстояний над диагональю				
	Но. 1	Но. 2	Но. 3	Но. 4	Но. 5
Но. 1	0,000000	1,007209	1,667925	1,211298	2,470341
Но. 2	1,003598	0,000000	4,493022	0,527089	0,579424
Но. 3	1,291482	2,119675	0,000000	5,529288	7,835947
Но. 4	1,100590	0,726009	2,351444	0,000000	0,440392
Но. 5	1,571732	0,761199	2,799276	0,663620	0,000000

Рис. 7.16. **Евклидовы расстояния между кластерами**

В верхней строке и первом столбце таблицы указаны номера кластеров, под диагональю (0,000) на пересечении строк и столбцов приведены значения евклидового расстояния, а над диагональю квадрат евклидового расстояния между каждой парой кластеров. Евклидово расстояние является метрикой для оценки геометрического расстояния в многомерном пространстве. Чем меньше расстояние между объектами, тем они более схожи. Наибольшее расстояние между кластерами №3 и №5; №3 и №4; № 2 и №3, так как евклидовые расстояния для этих групп от 2,11 до 2,80. Менее схожи кластеры №4 и 5; №2 и №4 и №2 и №5, здесь величина евклидового расстояния меньше единицы (от 0,66 до 0,76).

Сравнение таблиц, в которых приводятся элементы кластеров и таблицы средних кластеров показывает, что задача кластерного анализа методом k-средних успешно выполнена, так как различия между кластерами намного больше, чем различия между системами обработок почвы с удобрениями в каждом кластере, о чем свидетельствуют значения евклидовых расстояний в этих таблицах.

Для визуального представления средних для каждого кластера откроем вкладку **Дополнительно** и нажмем на кнопку **График средних (Graph of means)** и получим графическое изображение результатов кластеризации (рис. 7.17). На графике показаны средние значения переменных для каждого кластера. По горизонтали отложены участвующие в классификации

переменные, а по вертикали – средние значения переменных в разрезе получаемых кластеров в евклидовом расстоянии.

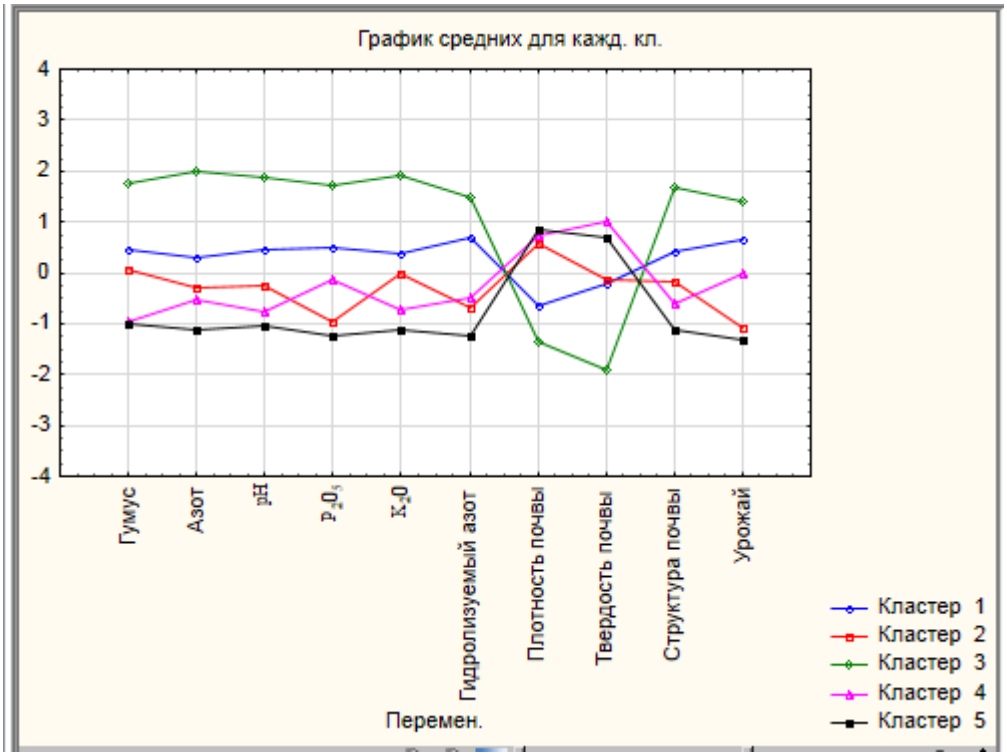


Рис. 7.17. **График средних для 5 кластеров**

Данные графика полностью подтверждают степень различий или схожести между кластерами, представленными в таблице средних и евклидового расстояния. Средние всех показателей плодородия по кластерам заметно отличаются друг от друга. Причем, наибольшее различие между пятыми и третьими кластерами, что видно по евклидовому расстоянию. В то время как второй и четвертый кластеры являются более похожими.

Для статистической оценки качества кластеризации проведем дисперсионный анализ. В диалоговом расширенном окне результатов кластеризации (рис. 7.13.) нажмем на кнопку **Дисперсионный анализ (Analysis of Variance),** появится таблица дисперсионного анализа, в которой приведена межгрупповая и внутригрупповая суммы квадратов, степени свободы, *F* – критерий и *p*- уровень значимости (рис. 7.18).

перемен.	Дисперсионный анализ (Кластер_стан_16)					
	Между SS	cc	Внутри SS	cc	F	значим. p
Гумус	13,00828	3	0,991715	11	48,09549	0,000001
Азот	12,38706	3	1,612938	11	28,15932	0,000018
pH	12,79076	3	1,209239	11	38,78428	0,000004
P_20_5	11,96681	3	2,033189	11	21,58103	0,000065
K_20	12,46527	3	1,534730	11	29,78113	0,000014
Гидролизуемый азот	12,24312	3	1,756882	11	25,55177	0,000029
Плотность почвы	10,88344	3	3,116557	11	12,80450	0,000656
Твердость почвы	12,27311	3	1,726887	11	26,05927	0,000027
Структура почвы	11,02634	3	2,973660	11	13,59601	0,000509
Урожай	11,51863	3	2,481365	11	17,02087	0,000192

Рис. 7.18. **Дисперсионный анализ результатов разделения на 5 кластеров**

Так как фактическое значение *F*- критерия по всем переменным больше $F_{05} = 6,22$ (при числе степеней свободы *3* и *11*), а уровень значимости *p* намного меньше принятого $p=0,05$, что свидетельствует о том, что средние в группах с вероятностью 95% существенно отличаются друг от друга. Это дополнительно подтверждает о качественном распределении изучаемых систем обработки почвы и удобрений по запланированным пяти кластерам.

Предположим, что разделение систем обработки почвы произойдет по трем вариантам удобрений (без удобрений, NPK, NPK+навоз). Для проверки данной гипотезы вернемся к диалоговому окну выбора параметров кластеризации (рис. 7.19) и зададим программе в строке **Число кластеров** 3 кластера.

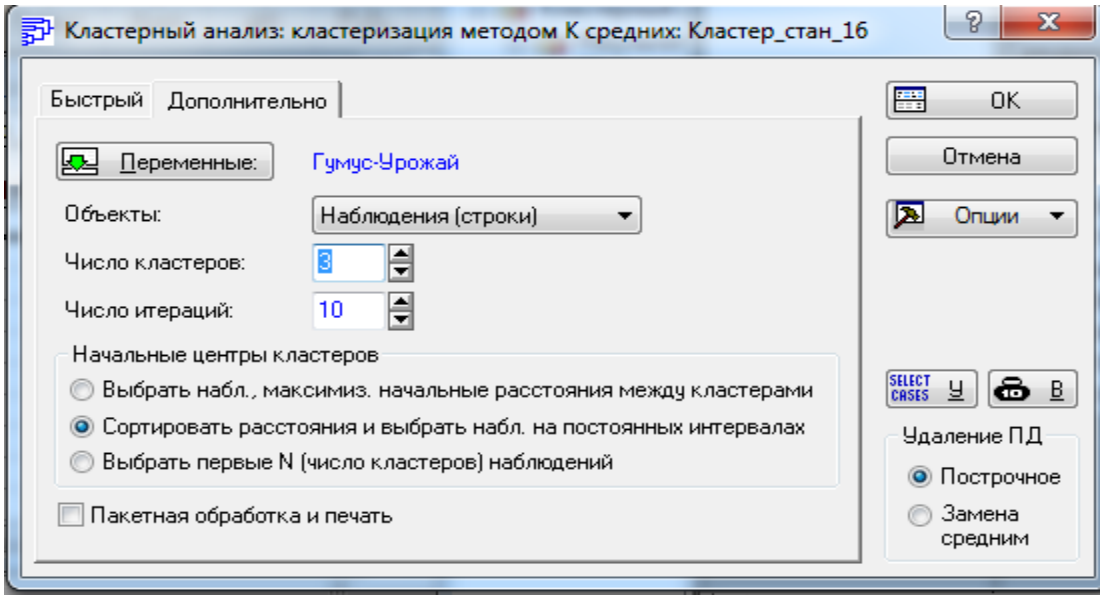


Рис. 7.19. **Диалоговое окно выбора параметров кластеризации**

Для перехода к открытому диалоговому окну результатов кластеризации нажмем в нижней части экрана на **Переход на опции** 

Далее в расширенном диалоговом окне результатов кластеризации (рис. 7.13) как и при анализе пяти кластеров выбираем следующие опции: **Элементы кластеров и расстояния, Средние кластеров и евклидовые расстояния, График средних (Graph of means) и Дисперсионный анализ (Analysis of Variance)** и получаем следующие результаты (рис.7.20):

Наблюд.	Элементы кластера номер 1 и расстояния до центра клас Кластер содержит 5 набл. объедин.
Обычная с NPK+ навоз	0,583220
Глубокая с NPK+ навоз	0,367991
Плоскорезная с NPK+ навоз	0,276157
Минимальная с NPK	0,306612
Сочетание ОМ с NPK	0,247455

Наблюд.	Элементы кластера номер 2 и расстояния до центра клас Кластер содержит 8 набл. объедин.
Обычная без удоб	0,370352
Обычная с NPK	0,483135
Глубокая без удоб	0,514326
Глубокая с NPK	0,338823
Плоскорезная без удоб	0,530898
Плоскорезная с NPK	0,461361
Минимальная без удоб	0,424727
Сочетание ОМ без удоб	0,590909

Наблюд.	Элементы кластера номер 3 и расстояния до центра клас Кластер содержит 2 набл. объедин.
Минимальная с NPK+ навоз	0,252918
Сочетание ОМ с NPK+ навоз	0,252918

Рис. 7.20. **Результаты итоговой кластеризации методом k-средних 3-х кластеров**

Таблица евклидовых расстояний между 3-мя кластерами (рис. 7.21).

Кластер Номер	Евклидовы расст. между кластерами (К Расстояния под диагональю Квадраты расстояний над диагональю		
	Но. 1	Но. 2	Но. 3
Но. 1	0,000000	1,466751	1,667925
Но. 2	1,211095	0,000000	5,969553
Но. 3	1,291482	2,443267	0,000000

Рис. 7.21. **Результаты итоговой кластеризации**

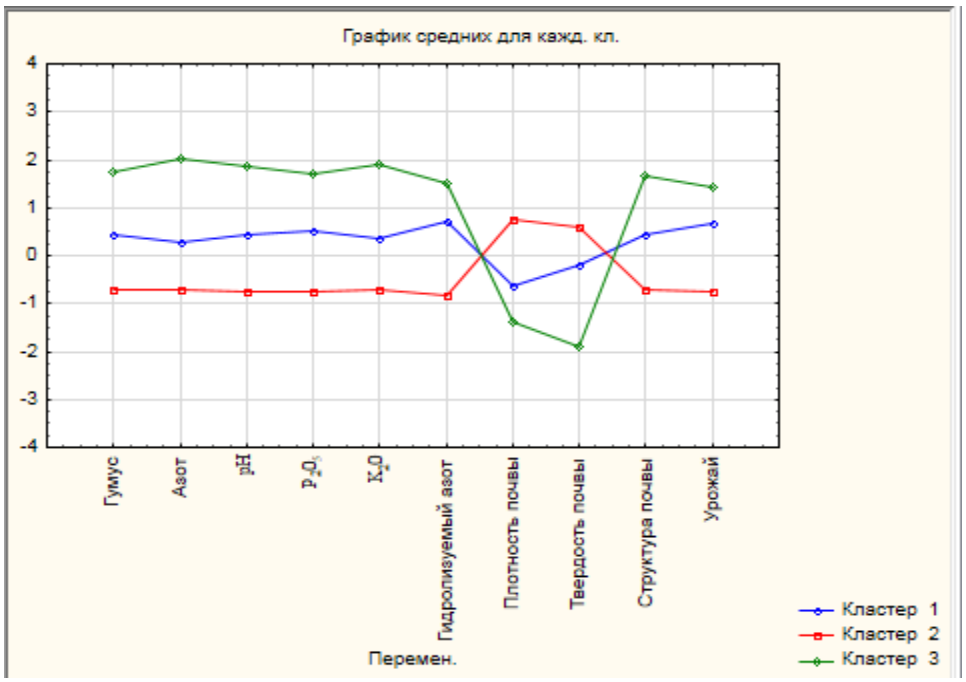


Рис. 7.22. **График средних 3-х кластеров**

Разбиение 15 объектов на 3 кластера по сравнению с разделением их на 5 кластеров привело к тому, что между кластерами стало еще большее различий. Так, евклидовое расстояние между кластерами № 1 и №2; №1 и №3 равно, соответственно, 1,21 и 1,29, а между кластерами №2 и №3 различие почти в 2 раза больше (рис. 7.21). Об этом же свидетельствует график средних (рис.7.22). Как показывает график, расстояние между средними значениями переменных по кластерам и общее расстояние между центрами кластеров значительное. Средние значения по всем агрофизическим и агрохимическим показателям плодородия почвы значительно отличаются друг от друга. Это свидетельствует о качественном разбиении на кластеры (группы).

	Дисперсионный анализ (Кластер_стан_16)					
перемен.	**Между SS**	cc	Внутри SS	cc	F	значим. p
Гумус	11,33114	2	2,668856	12	25,47416	0,000048
Азот	12,31806	2	1,681941	12	43,94230	0,000003
pH	12,49976	2	1,500238	12	49,99110	0,000002
P_20_5	11,63129	2	2,368713	12	29,46229	0,000023
K_20	11,86449	2	2,135510	12	33,33486	0,000013
Гидролизуемый азот	12,23965	2	1,760355	12	41,71765	0,000004
Плотность почвы	10,32990	2	3,670098	12	16,88767	0,000325
Твердость почвы	10,37517	2	3,624831	12	17,17349	0,000301
Структура почвы	10,41081	2	3,589187	12	17,40363	0,000284
Урожай	10,90783	2	3,092175	12	21,16535	0,000116

Рис. 7.23. **Дисперсионный анализ результатов разделения на 3 кластера**

В таблице (рис. 7.23) дисперсионного анализа приведены значения межгрупповых (*Between SS*) и внутригрупповых (*Within SS*) сумм квадратов отклонений, а также фактическое значение критерия Фишера ($F_ф$) и уровень значимости (p) для каждой переменной. Чем больше значения критерия $F_ф$ и меньше значения (p), тем лучше переменная характеризует принадлежность объектов к кластеру и тем «качественнее» наша кластеризация. Признаки с малыми значениями $F_ф$ ($F_{05}>F_ф$) и большими значениями p ($p>0,05$) можно из процедуры кластеризации исключить.

Дисперсионный анализ итогов кластеризации методом *k-средних* (для всех переменных $F_{фак}>F_{01}$, уровень значимости $p<0,0002$) показал, что все 10 показателей плодородия почвы оказали существенный вклад в разделение объектов на группы (кластеры). Это является дополнительным подтверждением успешного распределения изучаемых систем обработки почвы и удобрений на запланированные 3 кластера.

7.3 Двухвходовое объединение

Если при иерархической кластеризации и методе k-средних разделение на кластеры осуществляется отдельно или по объектам (наблюдениям) или по переменным, то суть метода двухвходового **объединения (Two-way joining)** состоит в том, что одновременно классифицируются как объекты, так и переменные. При этом предполагается, что и наблюдения (объекты) и переменные одновременно вносят вклад в формируемые кластеры и это может привести к достаточно интересным результатам. Вместе с тем, если по своей природе формируемые кластеры неоднородны по своей природе, это может привести к проблемам с интерпретацией полученных результатов.

Для кластеризации данных нашего примера файл «Кластер_стан_16» (объекты – 15 систем обработки почвы в сочетании с удобрениями и 10 признаком) выберем в стартовом модуле (рис. 7.5.2) опцию Д**вухвходовое объединение (Two-way joining)**

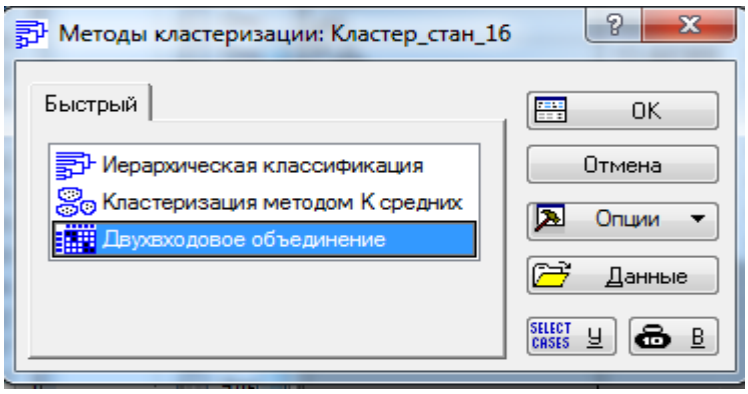


Рис. 7.5.2. **Стартовый модуль методов кластерного анализа**

В появившемся диалоговом окне (рис. 7.24) укажем диапазон переменных 2–11, в группе **Значение порога (Threshold Value)** выберем режим **Вычисление по данным (Стд.откл./2) Computed from data (Std.Dev./2)**

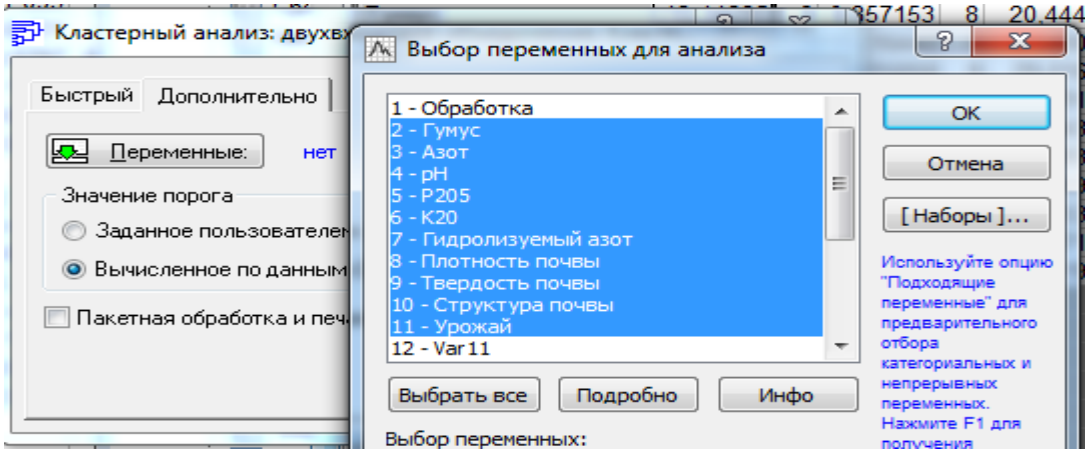


Рис. 7.24. **Диалоговое окно выбора параметров кластеризации двухвходового объединения**

После задания всех парамтеров нажмем на кнопку **Ок** и получим расширенное дилоговое окно результатов двухвходового объединения (рис. 7.25), в верхней части которого приводятся данные по количеству переменных, наблюдений, значение порога и стандартного отклонения. В нижней части располагаются кнопки для вывода разных форм результатов кластерного анализа.

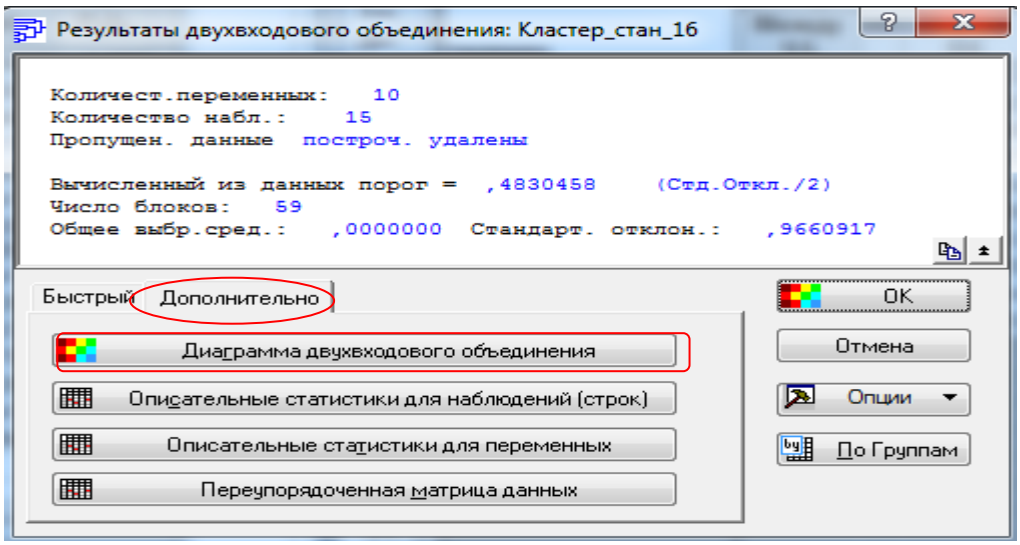


Рис. 7.25. **Расширенное диалоговое окно результатов двухвходового объединения**

Откроем вкладку **Дополнительно (Advanced)**, далее нажмем на **кнопку Диаграмма двухвходового объединения (Two-way joining graph)** и получим графическое изображение результатов двухвходового объединения (рис. 7.26).

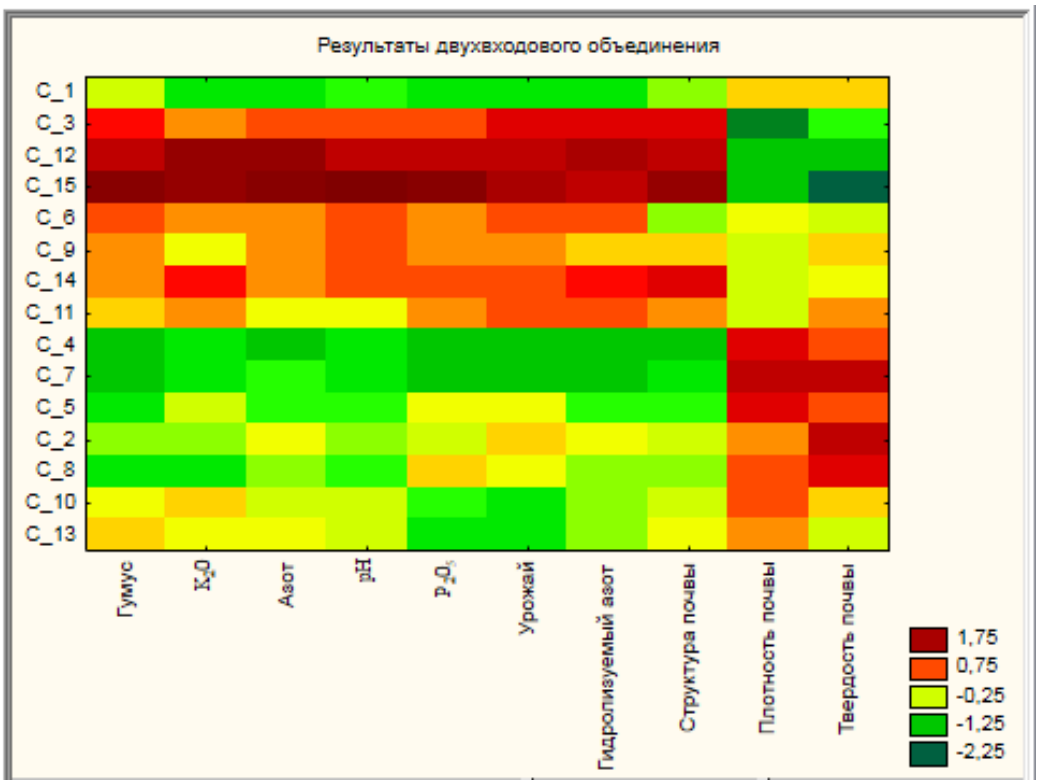


Рис. 7.26. **Результаты кластеризации методом двухвходового объединения**

На графике по горизонтали отложены участвующие в классификации переменные, в данном случае – агрохимические, агрофизические показатели и урожайность, а по вертикали – наблюдения (объекты). Цвета ячеек,

находящихся на пересечении, указывают на принадлежность элементов матрицы к определенному кластеру.

Контрольные вопросы:

1. Какие задачи решаются с помощью кластерного анализа?
2. Этапы кластерного анализа.
3. Методы кластеризации.
4. Что оценивается с помощью евклидового расстояния?
5. Иерархический кластерный анализ.
6. Что описывает дендрограмма?
7. Методы объединения при иерархической классификации.
8. Какие меры расстояния?
9. В чем суть кластеризации методом *k*-средних?
10. В чем суть метода двухвходового объединения ?
11. Визуальное представление средних при кластеризации.кластеров?
12. Как оценить качество кластеризации?

Библиографический список

1. Богданов Ю.И., Руднев А.В. Основы прикладной статистики: Уч. пособие. М.: МГИЭТ (ТУ), 2001, 113с.: ил.

2. Боровиков В.П. Популярное введение в программу STATISTICA. М. Компьютер – пресс. 1998. 267 с.

3. Боровиков В.П. Популярное введение в современный анализ данных и машинное обучение на STATISTICA . Учебное пособие для вузов. М. 2018. 354 с.

4. Буреева Н.Н. Многомерный статистический анализ с использованием ППП “STATISTICA”. Учебно-методический материал по программе повышения квалификации «Применение программных средств в научных исследованиях и преподавании математики и механики»/Н.Н. Буреева: Н. Новгород, 2007.–112 с.

5. ВуколовЭ.А. Основы статистического анализа. Практикум постатистическим методам и исследованию операций сиспользованием пакетов STATISTICA и EXCEL : учебное пособие / Э.А. Вуколов. — 2-еизд., испр. идоп. — М. : ФОРУМ, 2012. — 464 с.

6. Диденко Т.Н. Многомерные методы статистического анализа данных в экологии/Т.Н. Давиденко, О.Н. Давиденко, В.В. Пискунов, В.А. Болдырев. Учеб. пособие для студ. Биол. Фак., обучающихся по спец. 013100 «Экология», 011600 «Биология». – Саратов: Изд-во Сарат. ун-та, 2006. – 56 с.:ил.

7. Доспехов Б.А. Методика полевого опыта (с основами статистической обработки результатов исследований). Изд-во «АЛЬЯНС», 2011.– 352 с.

8. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл. – М.: Книга по Требованию, 2012. –128 с.

9. Кирюшин Б.Д. Основы научных исследований в агрономии/. Б.Д. Кирюшин, Р.Р. Усманов, И.П. Васильев И.П. – М.: КолосС, 2009. – 398 с.

10. Литтл Т.М., Сельскохозяйственное опытное дело. Планирование и анализ/Т.М. Литтл, Ф.Дж. Хиллз/Пер. с англ. Б.Д. Кирюшина; Под ред. и с предисловием Д.В. Васильевой. – М.: Колос, 1981. – 320 с.

11. Макарова Н.В. Статистический анализ медико-биологических данных с использованием пакетов статистических программ Statistica, SPSS, NCSS, SYSTAT: методическое пособие/Н.В. Макарова; Всерос. Центр эксрен. и радиац. Медицины им. А.М. Никифорова МЧС России – СПб.: Политехника-сервис, 2012. – 178 с.
12. Мастицкий С. Э. Методическое пособие по использованию программы STATISTICA при обработке данных биологических исследований. – Мн.: РУП «Институт рыбного хозяйства». – 76 С.
13. Мешалкина Ю.Л., Самсонова В.П. Математическая статистика в почвоведении: Практикум.- М.: МАКС Пресс, 2008. – 84с.
14. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. - Москва : Изд-во Медиа Сфера, 2006. - 305 с. : ил.,
15. Регрессионный анализ в почвоведении/Е.В. Шеин Е.В. [и др.] : учеб. пособие/ Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир: Изд-во ВлГУ, 2016. – 88 с.
16. Стукач О.В. Программный комплекс Statistica в решении задач управления качеством: учебное пособие / О.В. Стукач; Томский политехнический университет. – Томск: Изд-во Томского поли-
технического университета, 2011 – 163 с.
17. Электронный учебник по статистике. StatSoft, Inc. 1999. Москва. StatSoft. web: http://www.statsoft.ru

Учебно-методическое издание

Усманов Раиф Рафикович

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ АГРОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ В ПРОГРАММЕ «STATISTICA»

Учебно-методическое пособие

Ответственный редактор Е.Е. Рытова

Подписано для размещения в Электронно-библиотечной системе РГАУ-МСХА имени К.А. Тимирязева 10.08. 2020 г.

Оригинал-макет подготовлен Издательством РГАУ-МСХА
127550, Москва, Тимирязевская ул., 44
Тел. 8 (499) 977-40-64