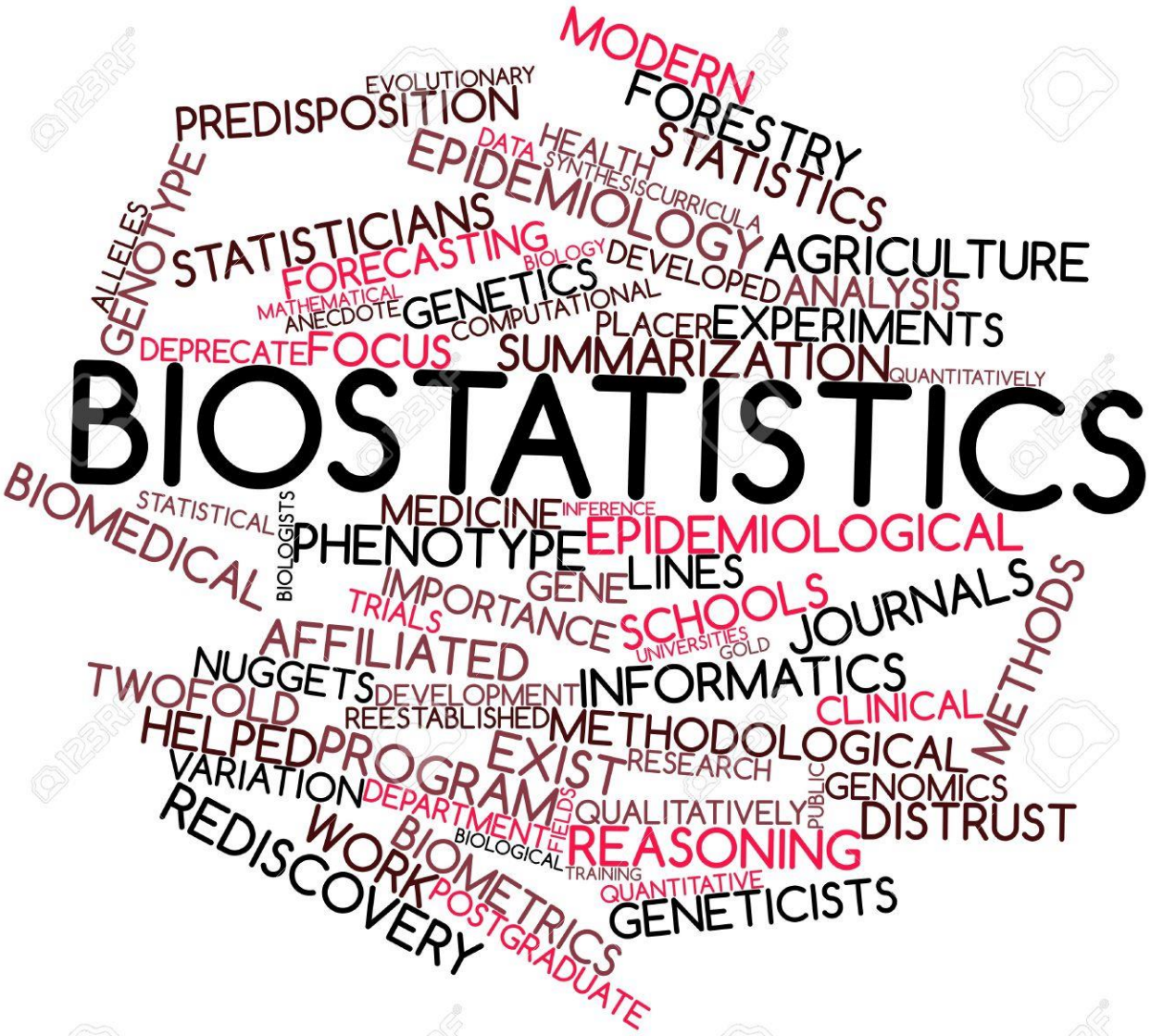
Министерство науки и высшего образования Российской Федерации
ФГБОУ ВО «Удмуртский государственный университет»
Институт естественных наук
Кафедра физиологии, клеточной биологии и биотехнологии

Алгоритмы биологической статистики
Учебно-методическое пособие



Издательский центр «Удмуртский университет»
Ижевск 2018

УДК 57: 519.2 (075.8)
ББК 28.0 в631.8 я73
А 456

Рекомендовано к изданию Учебно-методическим отделом УдГУ

Рецензент: к.м.н., доцент кафедры социальной гигиены и организации здравоохранения ИГМА Д.А. Толмачев
Автор: С.П. Кожевников

А 456 Алгоритмы биологической статистики: учебн.-метод. пособие / сост. С.П. Кожевников. Ижевск: Изд. центр «Удмуртский университет», 2018. – 75с.

ISBN 978-5-4312-0652-8

В пособии представлены примеры решения типовых задач, с которыми сталкиваются биологи в ходе статистической обработки данных, полученных в результате наблюдений и экспериментов. Рассмотрены следующие методы: соответствие данных закону нормального распределения, расчет объема выборки, параметрические и непараметрический виды анализа, множественные сравнения, факторный анализ, корреляционный и регрессионный анализы и т.д. Представлены алгоритмы выбора и пошаговые инструкции решений этих задач с использованием пакета программ STATISTICA. Пособие предназначено для бакалавров, магистров и аспирантов биологических направлений подготовки.

УДК57: 519.2 (075.8)
ББК 28.0 в631.8 я73

ISBN 978-5-4312-0652-8

Оглавление:

Введение

В последние годы широкое распространение получили различные программные средства для статистического анализа данных. Несмотря на это, необходимость владения хотя бы основами математической статистики сохраняется. Исследователь должен уметь грамотно выбирать подходящие статистические методы, знать их возможности и ограничения, корректно и осмыслено интерпретировать полученные результаты. Произвольное применение даже самых сложных методов статистического анализа может привести к ложным выводам.

В связи с этим, целью данного пособия является разработка наглядного алгоритма, демонстрирующего последовательность действий, которые следует выполнить исследователю при описании и анализе результатов научного исследования. Кроме того, рассмотрены некоторые примеры проведения различных методов статистического анализа, реализованные в программе STATISTICA.

Данное учебно-методическое пособие обобщает целый спектр методов математической статистически, применяемых при сборе и анализе биологической информации. В пособии представлены примеры решения типовых задач, с которыми сталкиваются биологи в ходе статистической обработки данных, полученных в результате наблюдений и экспериментов.

Данное пособие может быть использовано студентами бакалавриата и магистратуры по направлениям подготовки «Биология», «Психология» и «Физическая культура», преподавателями вузов, а также всеми интересующимися вопросами биологических исследований и практического применения теоретических знаний.

Алгоритм выбора метода статистического анализа.

Множество методов математической статистики, сложное описание процедур их выбора и реализации часто смущают исследователя. Однако при выборе необходимого метода статистического анализа учитывается ограниченное число ключевых факторов:

1. Тип распределения данных. В том случае, если распределение данных, полученных в эксперименте, рассматривается как соответствующее закону нормального распределения, применяются параметрические методы анализа. Для непараметрических методов анализа тип распределения данных не имеет значения.

2. Взаимосвязанность исследуемых данных. Взаимосвязанными (зависимыми) считаются те выборки, в которых изучаемый признак исследуется на одних и тех же объектах. Если измерения исследуемого признака проводятся на разных объектах, выборки рассматриваются как независимые (невзаимосвязанные). Для математической обработки данных в таких задачах используются методы сравнения для зависимых, либо независимых переменных.

3. Количественные характеристики исследуемых данных. В том случае если на исследуемый признак оказывают влияние несколько факторов, а также при сравнении нескольких экспериментальных групп, применяются различные виды множественного или дисперсионного анализа.

Кроме того, очень важно иметь представление об ограничениях, которые имеет каждый вид статистического анализа. Если выбранный метод не подходит для анализа имеющихся данных, всегда можно найти какой-либо другой, возможно, изменив тип представления самих данных.

С учетом данных факторов алгоритм выбора метода статистического анализа можно представить в виде следующей схемы:

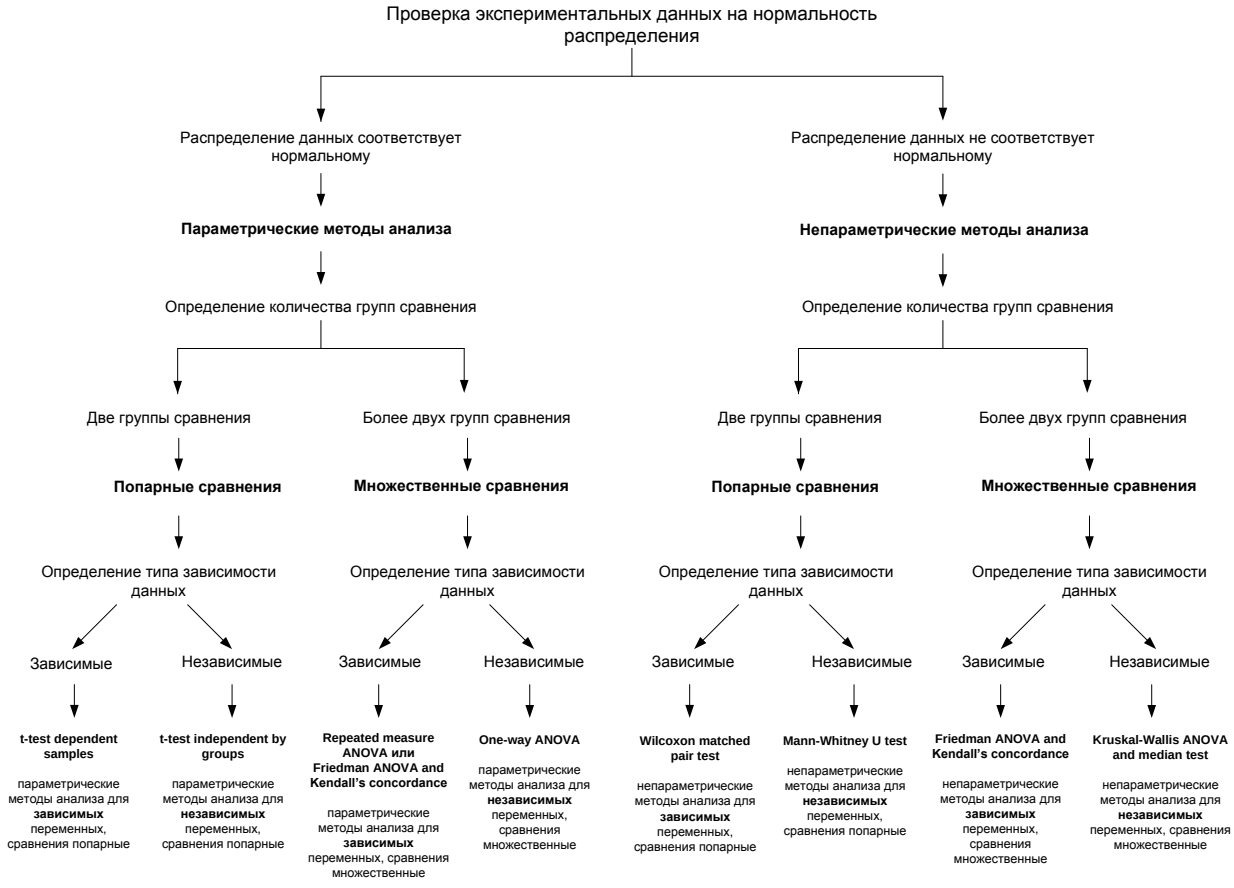


Рис. 1. Схема (алгоритм) выбора метода статистического анализа для биологических исследований

Раздел 1. Проверка соответствия анализируемых данных закону нормального распределения.

Существующие методы статистического анализа можно подразделить на две большие группы – параметрические и непараметрические. Важным условием, определяющим возможность применения того или иного метода анализа, является подчинение исследуемых данных закону нормального (Гауссова) распределения, графическое отображение которого имеет вид характерной колоколообразной кривой (рис.2).

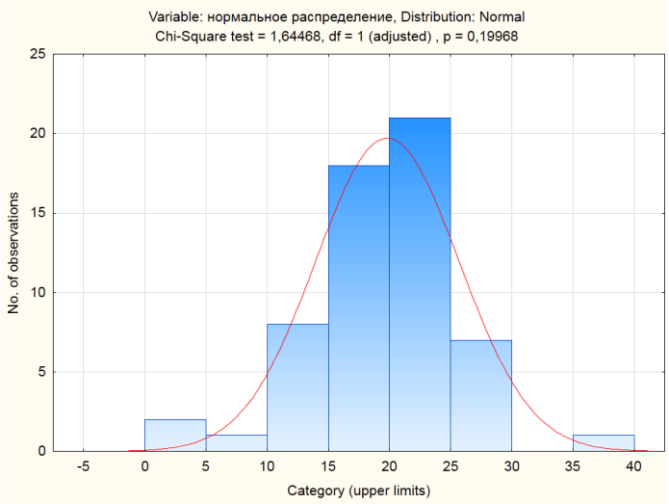


Рис. 2. Пример «нормального» (Гауссова) распределения данных

В случае подчинения исследуемых данных закону нормального распределения применяются параметрические методы анализа. В противном случае требуется применение непараметрических методов статистического анализа. Применение параметрических методов анализа для данных, не подчиняющихся закону нормального распределения признаков (распределение не соответствует критерию «нормальности»), приводит к выводам, не соответствующим действительности.

Установлено, что в подавляющем большинстве случаев (около 75%) распределение биологических признаков существенно отличается от «нормального». Во избежание ошибки, указанной выше, анализ любых биологических данных должен начинаться с проверки «нормальности» их распределения.

Рассмотрим некоторые подходы к оценке «нормальности» распределения данных, реализованные в программе STATISTICA.

На рис. 3 представлены результаты подсчета количества нейроглиальных клеток в черной субстанции мозга крыс. Необходимо установить, подчиняется ли распределение этих данных закону нормального распределения.

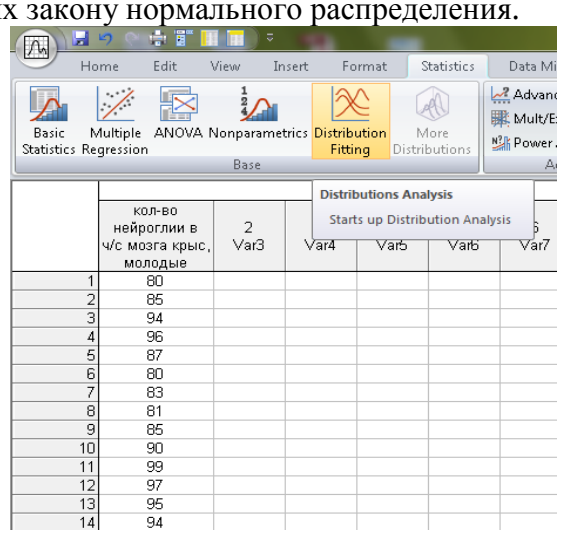


Рис. 3. Данные о количестве нейроглии у крыс

1. Из раздела главного меню **Statistics** запустить специальный модуль – **Distribution fitting** (Настройка распределения). Этот модуль позволяет проверить данные на соответствие целому ряду математических распределений (рис. 3).

2. Так как нам необходимо проверить подчинение данных закону нормального распределения, в списке непрерывных распределений **Continuous distributions** выбрать **Normal** (Нормальное) и нажать кнопку **ОК** (рис. 4).

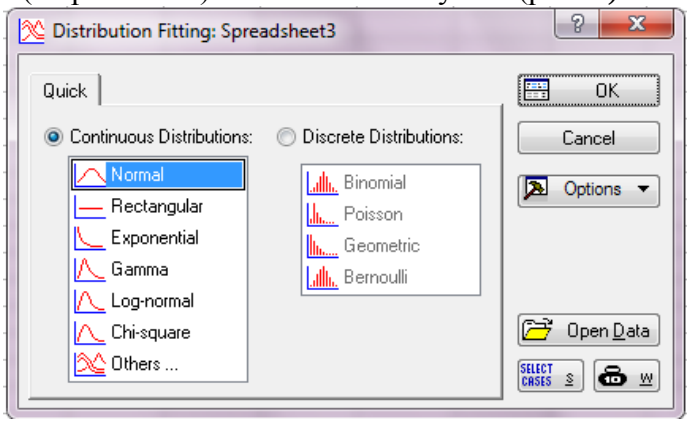


Рис. 4. Диалоговое окно модуля – Distribution fitting

3. В следующем окне, нажать кнопку **Variable** (переменные), указать, какую именно переменную мы хотим проанализировать. Затем нажать кнопку **Plot of observed and expected distributions** (График наблюдаемого и ожидаемого распределений) (рис. 5).

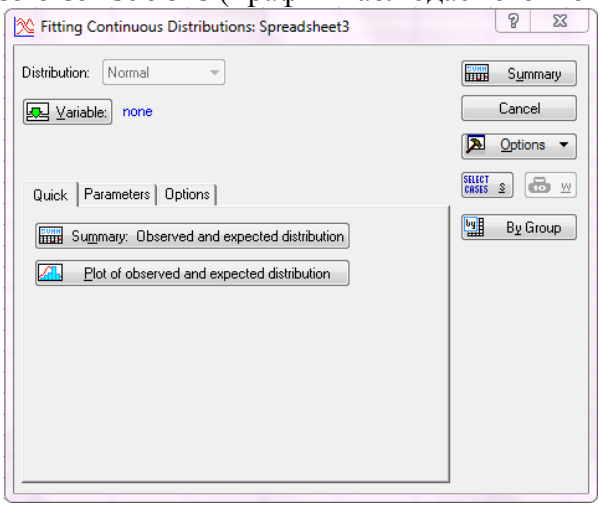


Рис. 5. Диалоговое окно для выбора анализируемых переменных

Полученная гистограмма отражает распределение данных исследуемого параметра (рис. 6).

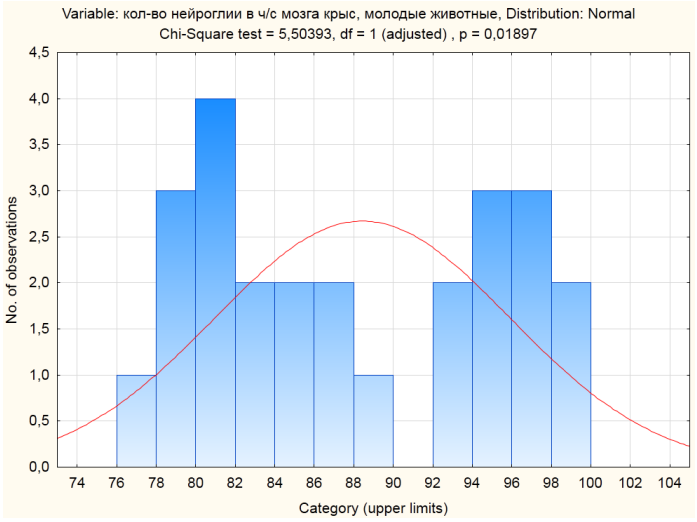


Рис. 6. Результат анализа распределения исследуемых данных

Полученный рисунок показывает, что распределение значений исследуемого параметра отличается от «нормального» (столбики гистограммы не формируют колоколообразную кривую). Это заключение основано на визуальном анализе, однако оно имеет и более строгое подтверждение. В верхней части гистограммы представлены результаты теста $\chi 2$ Chi-square test (тест хи-квадрат). Данный тест проверяет гипотезу о том, что наблюдаемое распределение не отличается от теоретически ожидаемого, «нормального». Если вероятность ошибки при отклонении этой гипотезы оказалась намного больше 0.05 ($p>0.05$), то гипотеза верна. Иными словами, распределение значений, составляющих данную выборку, статистически не отличается от «нормального». **В нашем случае вероятность ошибки менее 0.05 (p=0.01897), следовательно, распределение значений не подчиняется «нормальному» закону.**

Однако необходимо отметить, что применение теста хи-квадрат достаточно часто приводит к ошибочному выводу о «нормальности» распределения (мощность данного теста относительно невысока). В связи с этим лучше воспользоваться другими тестами, которые можно найти в модуле **Basic Statistics** (Основная статистика) (рис. 7).

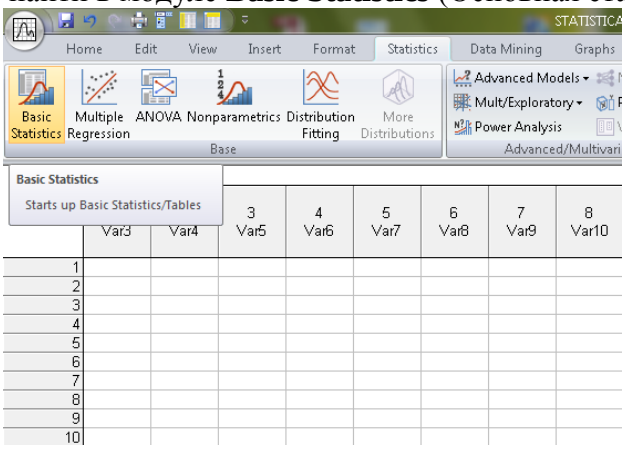


Рис. 7. Диалоговое окно модуля Basic Statistics (Основная статистика)

1. В разделе **Basic Statistics** (Основная статистика), выбрать модуль **Descriptive Statistics** (Описательная статистика) (рис. 8).

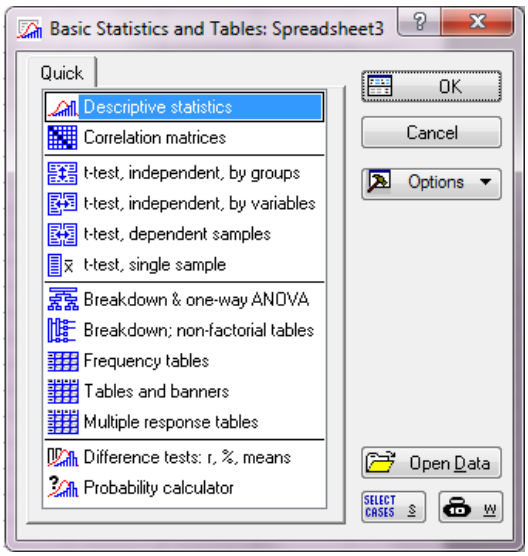


Рис. 8. Диалоговое окно модуля Описательная статистика

2. Открыть закладку **Normality** и выбрать опции **Kolmogorov-Smirnov and Lilliefors test for normality** (Тест Колмогорова-Смирнова и Лиллифорса) или **Shapiro-Wilk's W test** (W-тест Шапиро-Уилка) (рис. 9).

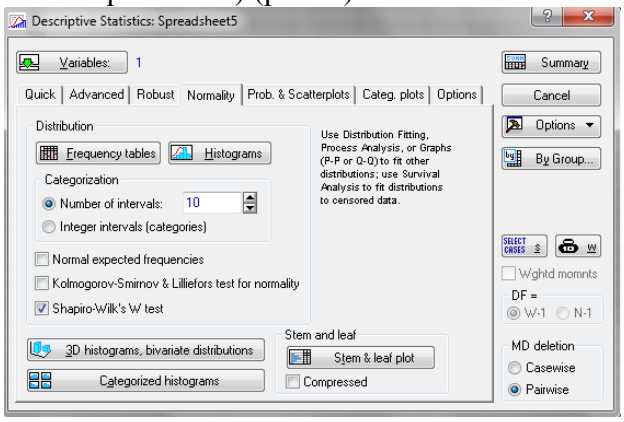


Рис. 9. Диалоговое окно модуля Описательная статистика

Эти тесты также проверяют гипотезу об отсутствии различий между наблюдаемым и теоретически ожидаемым, «нормальным» распределением. Наибольшей мощностью, особенно при небольших выборках (n < 50), обладает тест Шапиро-Уилка (Shapiro-Wilk's W test). Для выбора этого теста, необходимо поставить «галочку» рядом с его названием.

3. Далее нажать кнопку кнопка **Variables** (переменные) и выбрать переменную для анализа. После нажатия кнопки **Histograms** (гистограмма), программа создаст гистограмму распределения значений признака и ожидаемую нормальную кривую (рис.10).

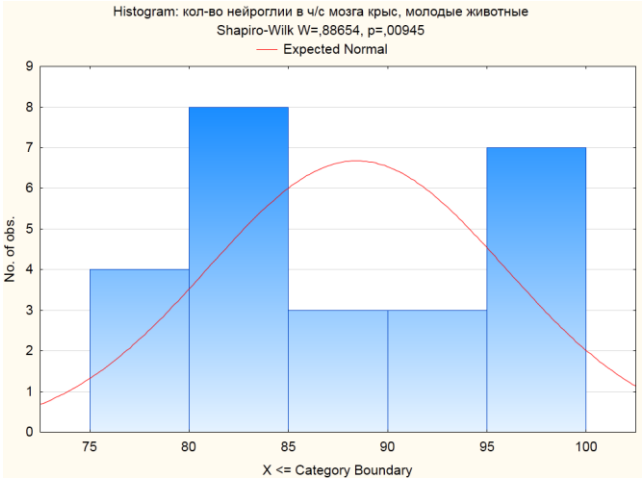


Рис. 10. Результат анализа распределения исследуемых данных

Результаты выбранных тестов на «нормальность» автоматически располагаются в заголовке этого графика. В нашем примере использование теста Шапиро-Уилка показывает P = 0.00945, что подтверждает сделанный ранее вывод о неподчинении данных закону нормального распределения.

Проверить данные на соответствие закону нормального распределения также можно с использованием графика нормальных вероятностей. На данном графике отображается зависимость реальных частот значения признака от ожидаемых, «нормальных». Если между наблюдаемым и ожидаемым распределениями нет никакой разницы, точки на этом графике выстроятся строго вдоль прямой. Иначе, они образуют фигуру, отличную от прямой. Для построения графика такого типа необходимо:

1. Из главного меню выбрать раздел **Basic Statistics** (Основная статистика) и модуль **Descriptive Statistics** (Описательная статистика) (рис. 7 и 8).

2. В появившемся диалоговом окне выбрать закладку **Prob. & Scatterplots** (Вероятностные графики и диаграммы рассеяния) и нажать на кнопку **Normal probability plot** (График нормальных вероятностей) (рис. 11).

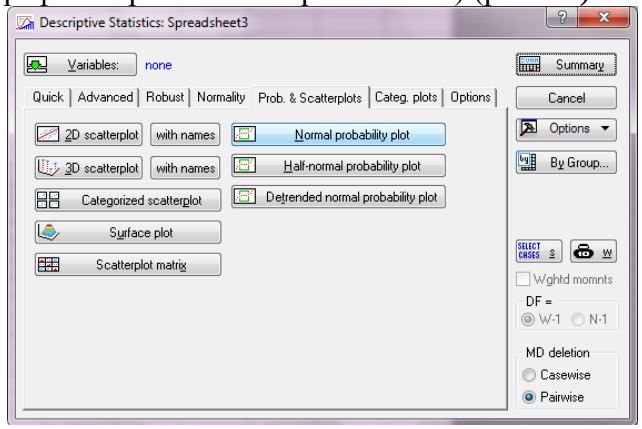


Рис. 11. Диалоговое окно модуля Вероятностные графики и диаграммы рассеяния

В результате появится график (рис. 12), точки на котором, в случае «нормального» распределения данных, плотно выстраиваются вдоль теоретически ожидаемой прямой. В нашем случае точки значительно отклоняются от прямой, что еще раз подтверждает предположение о несоответствии данных закону нормального распределения.

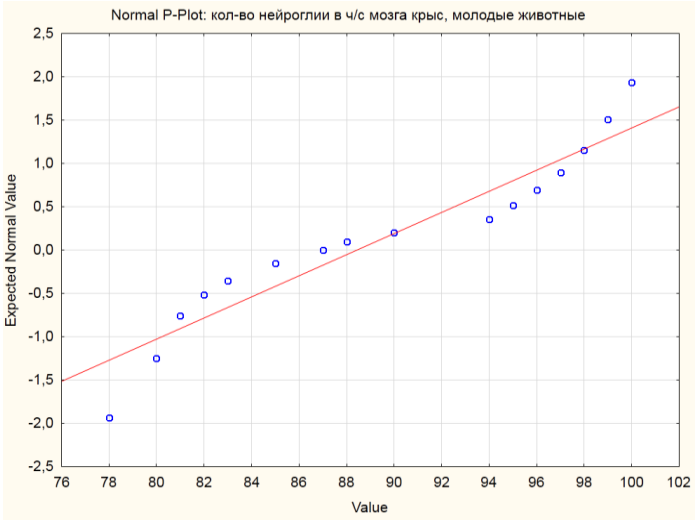


Рис. 12. График нормальных вероятностей

Раздел 2. Сравнение в двух группах.

Сравнение двух «независимых» групп, распределение данных в которых соответствует «нормальному» (T-test, independent, by groups).

В биологических исследованиях одной из наиболее часто встречаемых задач является сравнение арифметических средних двух групп. Важной характеристикой сравниваемых групп является их «зависимость» или взаимосвязанность.

Зависимые выборки содержат данные, полученные при исследовании одной и той же экспериментальной группы, но в разные временные периоды. Например, «до» и «после» какого-либо воздействия: лечения (введения препарата), обучения или тренировки, хирургической операции и т.д. В данном случае, результаты измерений полученные «до» и «после» экспериментального воздействия, будут взаимосвязаны друг с другом (взаимозависимы). Обратите внимание, количество объектов в этих выборках всегда одинаковое.

Независимые выборки получаются при исследовании двух различных групп. Результаты измерения в одной выборке не оказывают влияния на результаты, полученные в другой выборке. Например, «экспериментальная» и «контрольная» группы или сравнения в группах мужчин и женщин. Допускается, чтобы количество объектов в них было различным.

Классическим методом, позволяющим решить подобную задачу, является t-тест Стьюдента, или просто «t-тест». В ходе данного теста проверяется гипотеза о том, что наблюдаемые различия между средними значениями сравниваемых выборок случайны и не вызваны действием изучаемого фактора (нулевая гипотеза).

Данный тест относится к группе параметрических методов анализа, его корректное применение требует выполнения трех условий:

1. Обе выборки должны быть независимыми;
2. Обе выборки должны подчиняться закону нормального распределения;
3. Обе выборки должны быть однородны (разброс данных внутри выборок не должен быть слишком большим).

Наиболее важным является условие соблюдения требования о подчинении закону нормального распределения. **Несоблюдение этого требования делает применение данного теста невозможным.**

Алгоритм выбора данного метода статистического анализа можно представить в виде следующей схемы (рис. 13):

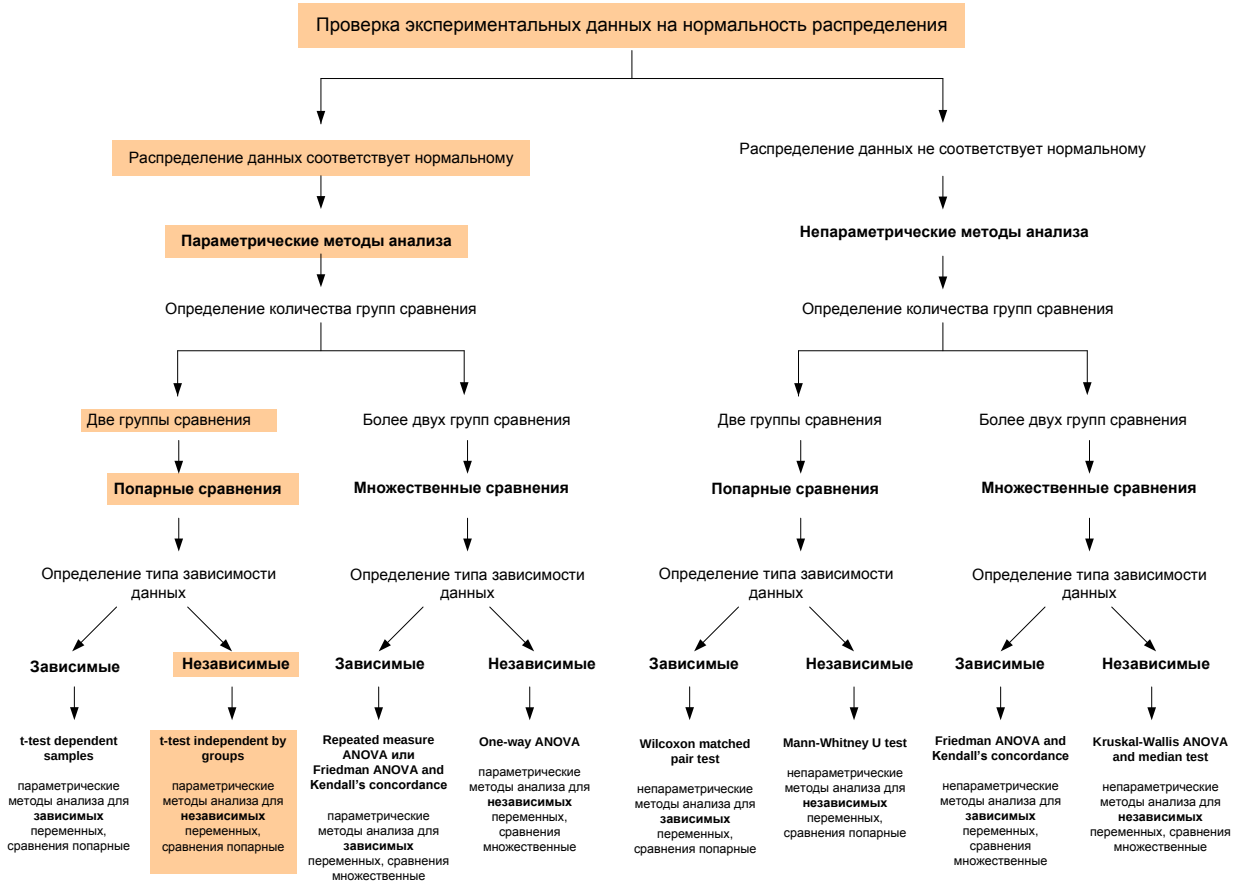


Рис. 13. Схема выбора метода статистического анализа для сравнения двух независимых групп, при условии соответствия данных закону нормального распределения

Рассмотрим применение t-теста с использованием следующего примера.

Известно, что процессы старения сопровождаются гибелью нервных клеток в различных отделах головного мозга. Предполагается, что возможной причиной гибели нейронов является усиление воспалительных процессов в нервной ткани. Для оценки интенсивности воспаления подсчитывали количество глиальных клеток (они являются маркерами воспаления) у животных (крыс) разных возрастных групп. Статистический анализ должен ответить на вопрос, различается ли среднее количество данных клеток и, как следствие, интенсивность воспалительных процессов у животных разного возраста.

На рис. 14 представлены данные о количестве глиальных клеток в двух экспериментальных группах: группа 1 – молодые животные; группа 2 старые животные. **Обратите внимание на оформление данных (рис. 14)** - таблица имеет 2 переменные. Первая переменная **– группирующая** (Grouping variables) содержит коды, указывающие на принадлежность данных к конкретной группе. Вторая переменная **зависимая** (Dependent variables) содержит собственно данные. Сами данные (в нашем случае – о количестве клеток) располагаются по вертикали (друг под другом в один столбец). Аналогичным образом оформляются данные во всех случаях, когда производится сравнение независимых групп.

1 Grouping variable	2 Кол-во клеток глии
1	23
1	49
1	42
1	47
1	46
2	87
2	109
2	41
2	79
2	106
2	126
2	85
2	77

Рис. 14. Пример оформления данных при сравнении двух независимых групп

В наиболее простом варианте данные для каждой группы («молодые» и «старые») можно просто внести в отдельные столбцы, однако при сравнении независимых групп первый вариант их оформления является предпочтительным.

Допустим, что данные в обеих выборках распределены «нормально», а дисперсии различаются незначительно. Для выполнения t-теста необходимо:

1. Запустить соответствующий модуль из меню: **Statistics / Basic statistics / t-test, independent, by groups** (рис. 15).

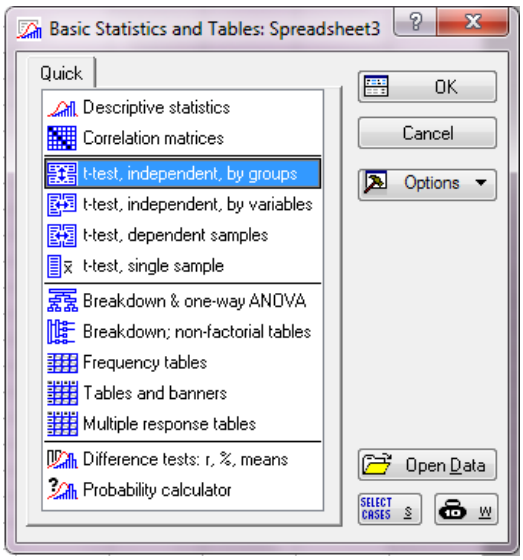


Рис. 15. Диалоговое окно модуля Основная статистика

2. В открывшемся окне указать коды групп в соответствующих окнах (**Code for Group 1 и 2**) или нажать кнопку **Variables** и указать переменные, которые необходимо сравнить.

В правом окне необходимо выбрать групповую переменную, в левом – зависимую переменную (рис. 16).

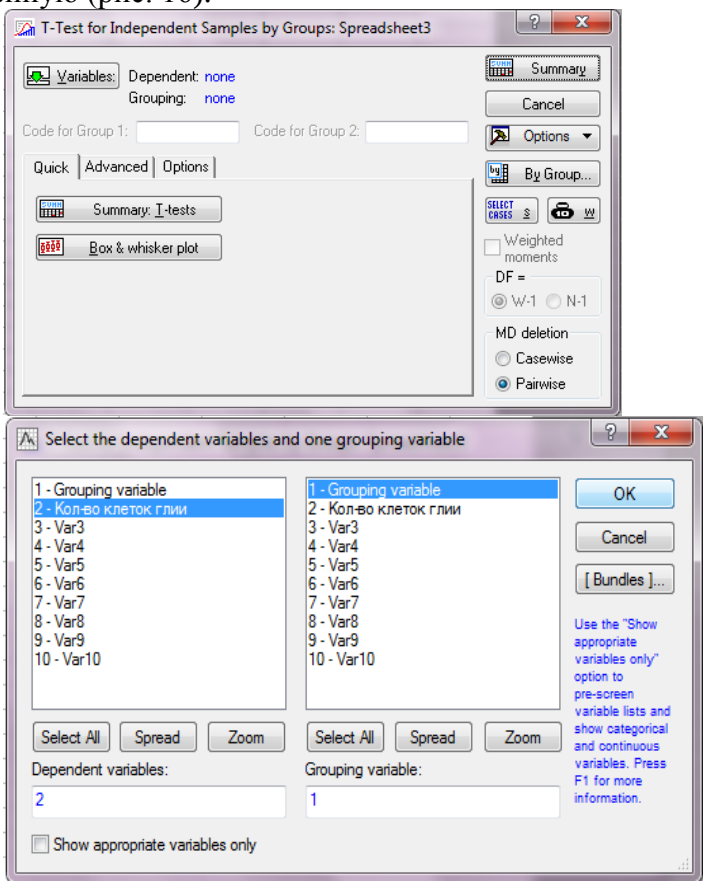


Рис. 16. Диалоговые окна для выбора исследуемых переменных

3. Нажать на кнопку **Summary: T-tests**. В итоге программа создаст таблицу, содержащую следующие результаты (рис.17).

Variable	T-tests; Grouping: Grouping variable (Spreadsheet3) Group 1: 1 Group 2: 2 Mean 1	Mean 2	t-value	df	p	Valid N 1	Valid N 2	Std.Dev. 1	Std.Dev. 2	F-ratio Variances	p Variances
Кол-во клеток глии	41,40000	88,75000	-3,86696	11	0,002622	5	8	10,59717	25,70575	5,884111	0,106096

Рис. 17. Таблица с результатами t - теста

Данная таблица содержит следующие показатели: Std. dev. стандартное отклонение выборки 1; Std. dev. стандартное отклонение выборки 2; P, Variances – вероятность ошибки для F-теста Фишера, если $P > 0.05$, условие однородности дисперсий выполняется.

Главный показатель – это значение P (вероятность ошибочно отклонить нулевую гипотезу об отсутствии различий между средними). В нашем случае $P < 0.05$, следовательно, между средними значениями количества клеток глии у молодых и старых животных есть статистически значимые различия.

Сравнение двух «независимых» групп, распределение данных в которых не соответствует «нормальному» (Mann-Whitney U- test).

Если распределение значения признака в двух сравниваемых группах отличается от «нормального», применение параметрического t-теста для их сравнения будет приводить к искаженным результатам. В таких случаях следует воспользоваться соответствующим непараметрическим аналогом теста Стьюдента.

Сравнение двух независимых групп, распределение данных в которых не соответствует «нормальному», производится с использованием U-теста Манна-Уитни (Mann-Whitney U- test).

Алгоритм выбора данного метода статистического анализа можно представить в виде следующей схемы (рис. 18):

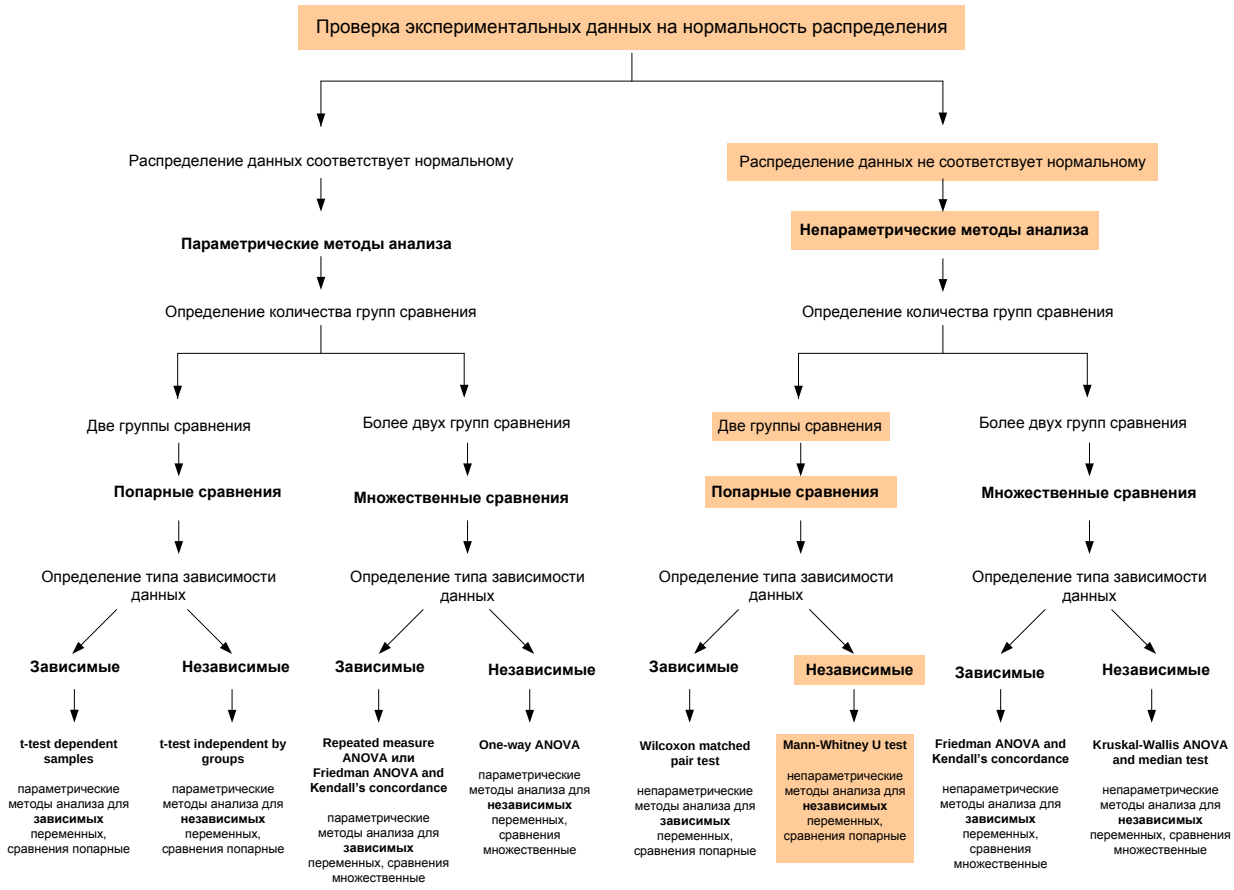


Рис. 18. Схема выбора метода статистического анализа при сравнении двух независимых групп, данные в которых, не подчиняются закону нормального распределения

В программе STATISTICA этот тест выполняется следующим образом: внести в таблицу результаты исследования в соответствии с правилами оформления данных для независимых групп (см. стр. 13-14).

1. В меню **Statistics** выбрать пункт **Nonparametrics** (Непараметрическая статистика) (рис. 19).

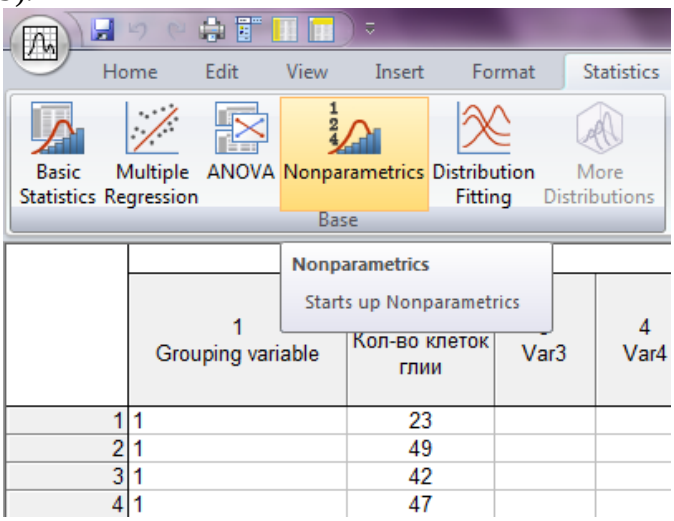


Рис. 19. Раздел главного меню Непараметрическая статистика

2. Далее необходимо выбрать пункт **Comparing two independent samples** (Сравнение двух независимых выборок) (рис. 20).

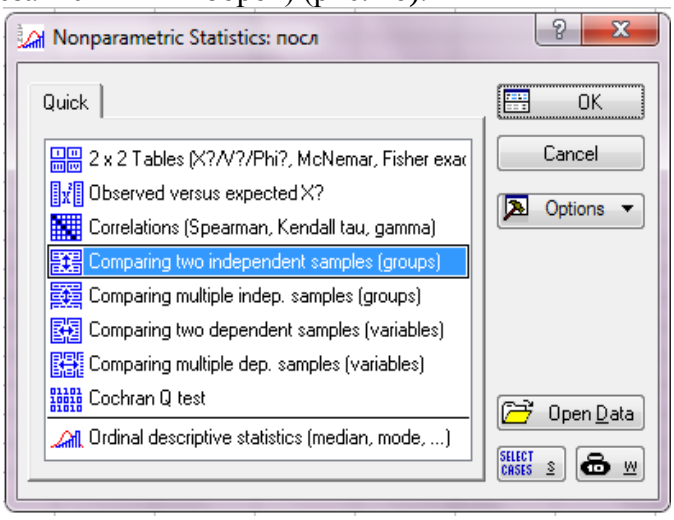


Рис. 20. Диалоговое окно модуля Основная статистика

3. В появившемся окне нажать на кнопку **Variables** и выбрать зависимую и групповую переменные. В правом окне выбрать групповую переменную, в левом – зависимую и нажать ОК (рис. 21).

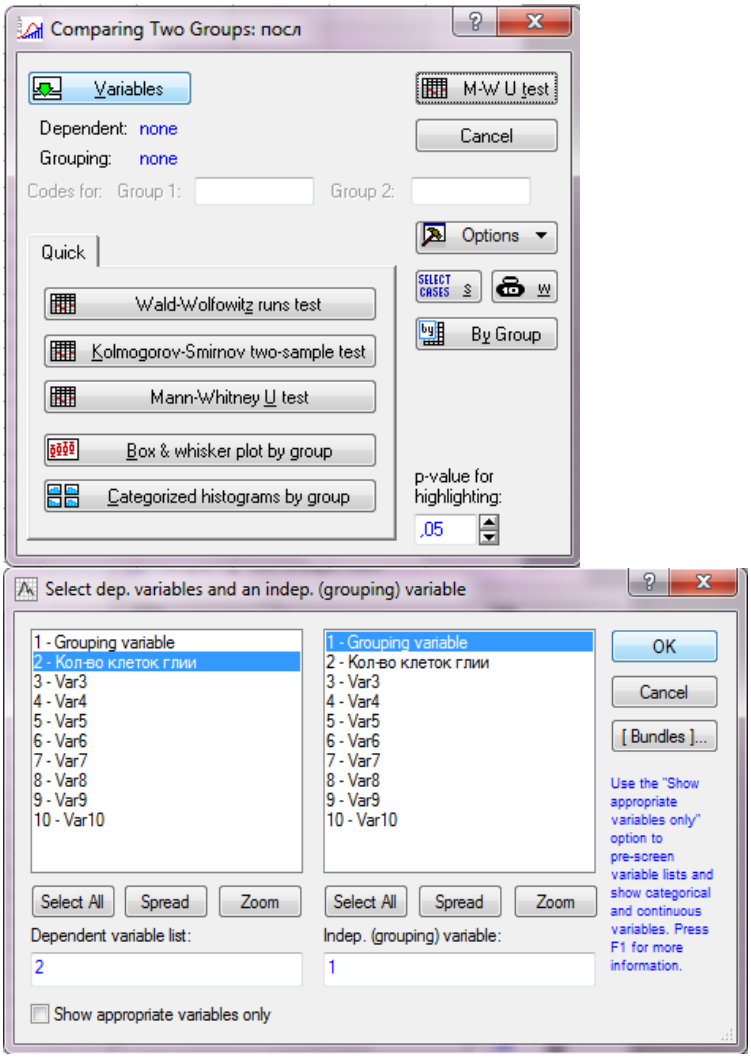


Рис. 21. Диалоговые окна для выбора исследуемых переменных

4. Нажать на кнопку **Mann-Whitney U-test** или M-W U test (рис. 21), после чего появится таблица с результатами рис. 22.

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

variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2	2*1sided exact p
Кол-во клеток глии	19,00000	72,00000	4,000000	-2,26897	0,023271	-2,26897	0,023271	5	8	0,018648

Рис. 22. Таблица с результатами t - теста

Главный показатель, на который необходимо обратить внимание это величина вероятности ошибки (p-value). Поскольку $P < 0.05$ между сравниваемыми выборками имеются статистически значимые различия (Примечание: в отличие от t-теста, тест Манна-Уитни сравнивает не средние значения выборок, а суммы рангов по каждой из них).

Сравнение двух «зависимых» групп, распределение данных, в которых соответствует «нормальному» (T-test, dependent samples).

Напомню, с зависимыми выборками исследователь имеет дело в том случае, если исследование выполняется на одних и тех же объектах. Рассмотрим следующий пример. Известно, что интенсивные физические нагрузки приводят к ослаблению иммунитета и частым простудным заболеваниям. Для выяснения причин данного феномена было проведено модельное исследование на животных. В качестве физической нагрузки лабораторные животные (мыши) плавали с грузом до истощения. «До» и «после» нагрузки у животных производили взятие крови и оценивали уровень иммуноглобулинов. Необходимо выяснить, различается ли среднее количество иммуноглобулинов «до» и «после» физической нагрузки. Поскольку исследование проводится на одних и тех же животных, то выборки являются зависимыми. При условии соблюдения требований о «нормальности» распределения данных, воспользуемся t-тестом для зависимых выборок.

Алгоритм выбора данного метода статистического анализа можно представить в виде следующей схемы:

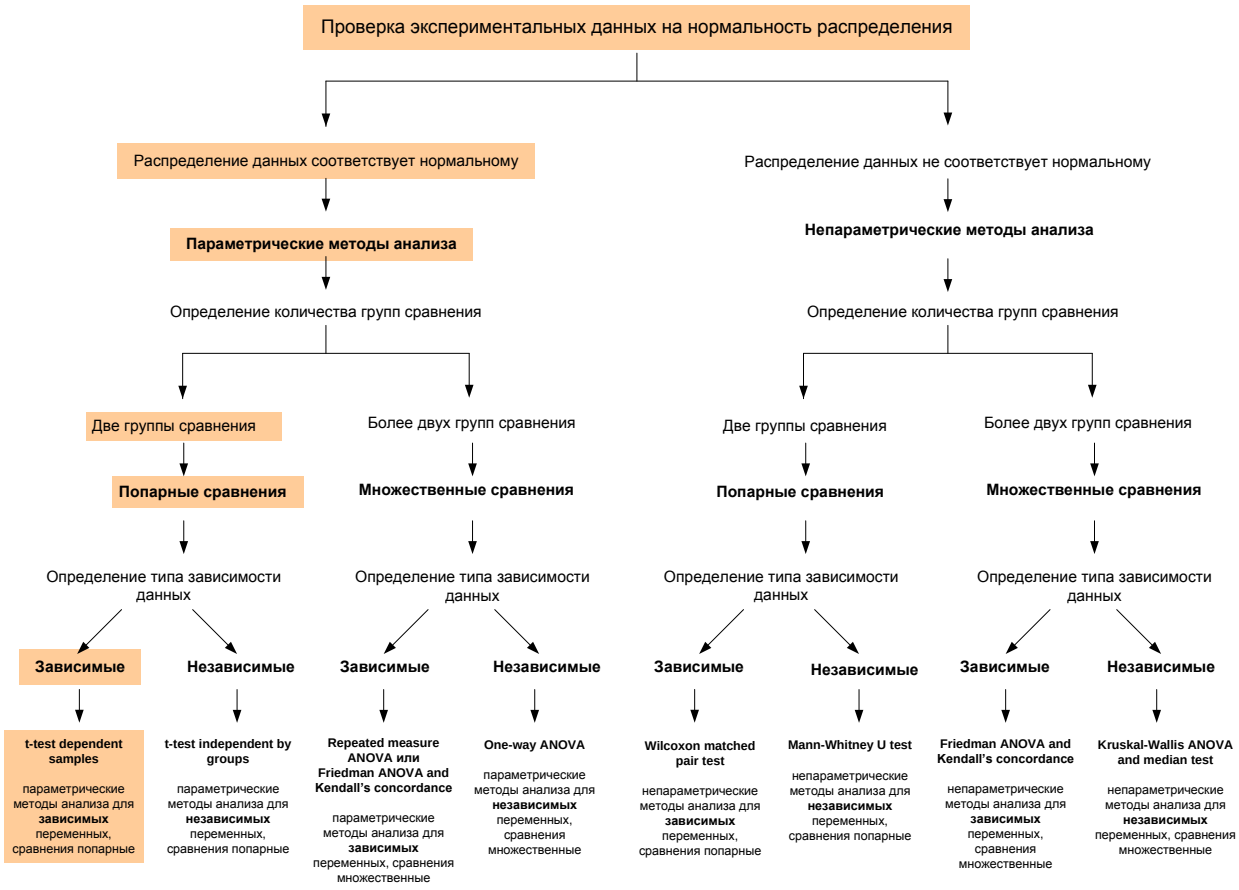


Рис. 23. Схема выбора метода статистического анализа при сравнении двух зависимых групп при условии соответствия данных закону нормального распределения

Обратите внимание: поскольку экспериментальные данные имеют **зависимый** характер – внесите результаты исследования для каждой переменной в отдельные столбцы (рис. 24). Включение в таблицу групповой переменной не требуется. Аналогичным образом оформляются данные во всех случаях, когда производится сравнение зависимых групп.

	1 IgG мышей до физ. нагрузки	2 IgG мышей после физ.нагрузки	3 Var3
1	9	10	
2	7	9	
3	9	8	
4	6	8	
5	8	9	
6	9	10	
7	6	7,5	
8	6	7	
9	7	6	
10	6	6,5	
11	6	7	
12	6	6	
13	6	6	
14	6,5	6	
15	7	7	
16	7	6	
17	6	6	
18	7	6	
19	5	7	
20			

Рис. 24 Пример оформления данных при сравнении двух зависимых групп.

Для выполнения этого варианта t-теста необходимо:

1. Запустить из меню **Statistics** модуль **Basic statistics / t-test, dependent samples** (рис. 25).

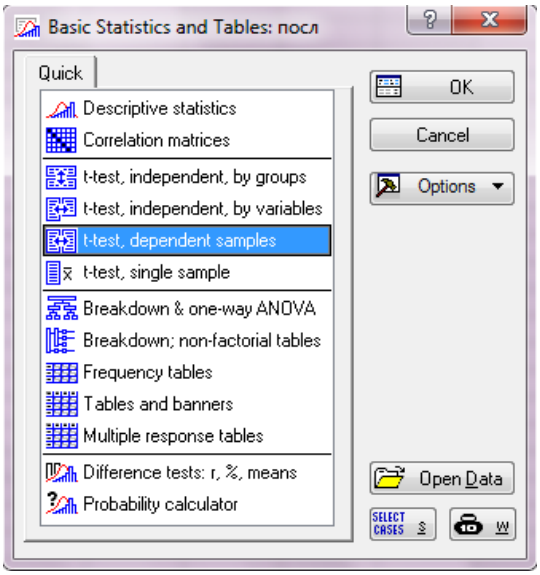


Рис. 25. Диалоговое окно модуля Основная статистика

2. Нажать на кнопку **Variables** и указать переменные, участвующие в анализе: первую (**First variable**) и вторую (**Second variable** (рис. 26).

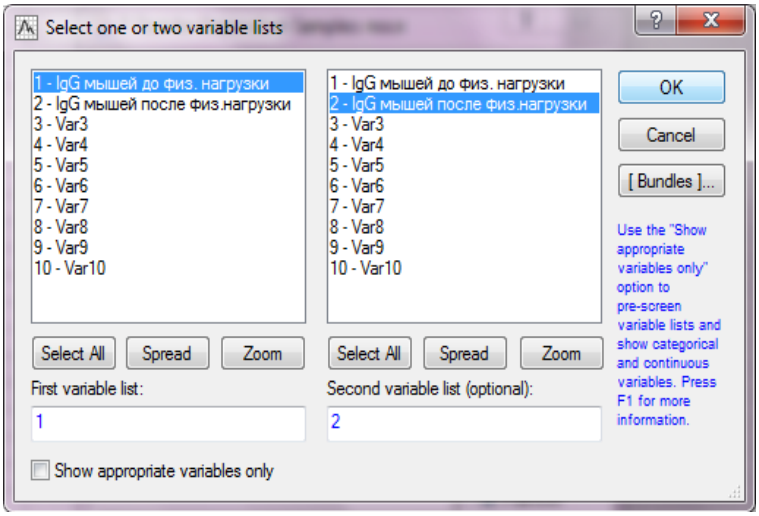


Рис. 26. Диалоговое окно для выбора исследуемых переменных

3. Нажать на кнопку **Summary: T-tests**. Появится таблица с результатами, аналогичная той, что мы видели при выполнении t-теста для независимых выборок (рис. 27).

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

Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p	Confidence -95,000%	Confidence +95,000%
IgG мышей до физ. нагрузки	6,815789	1,169170								
IgG мышей после физ.нагрузки	7,263158	1,378087	19	-0,447368	1,052705	-1,85240	18	0,080439	-0,954756	0,060019

Рис. 27. Таблица с результатами t- теста

Поскольку в нашем случае P > 0.05, можно сделать заключение о том, что среднее количество иммуноглобулинов до и после физической нагрузки достоверно не различается.

Сравнение двух зависимых групп, распределение данных в которых не соответствует «нормальному» (Wilcoxon matched pair test).

В том случае, если распределение данных в двух зависимых выборках отличается от «нормального», для их сравнения необходимо использовать тест Уилкоксона (Wilcoxon matched pair test).

Алгоритм выбора данного метода статистического анализа можно представить в виде следующей схемы:

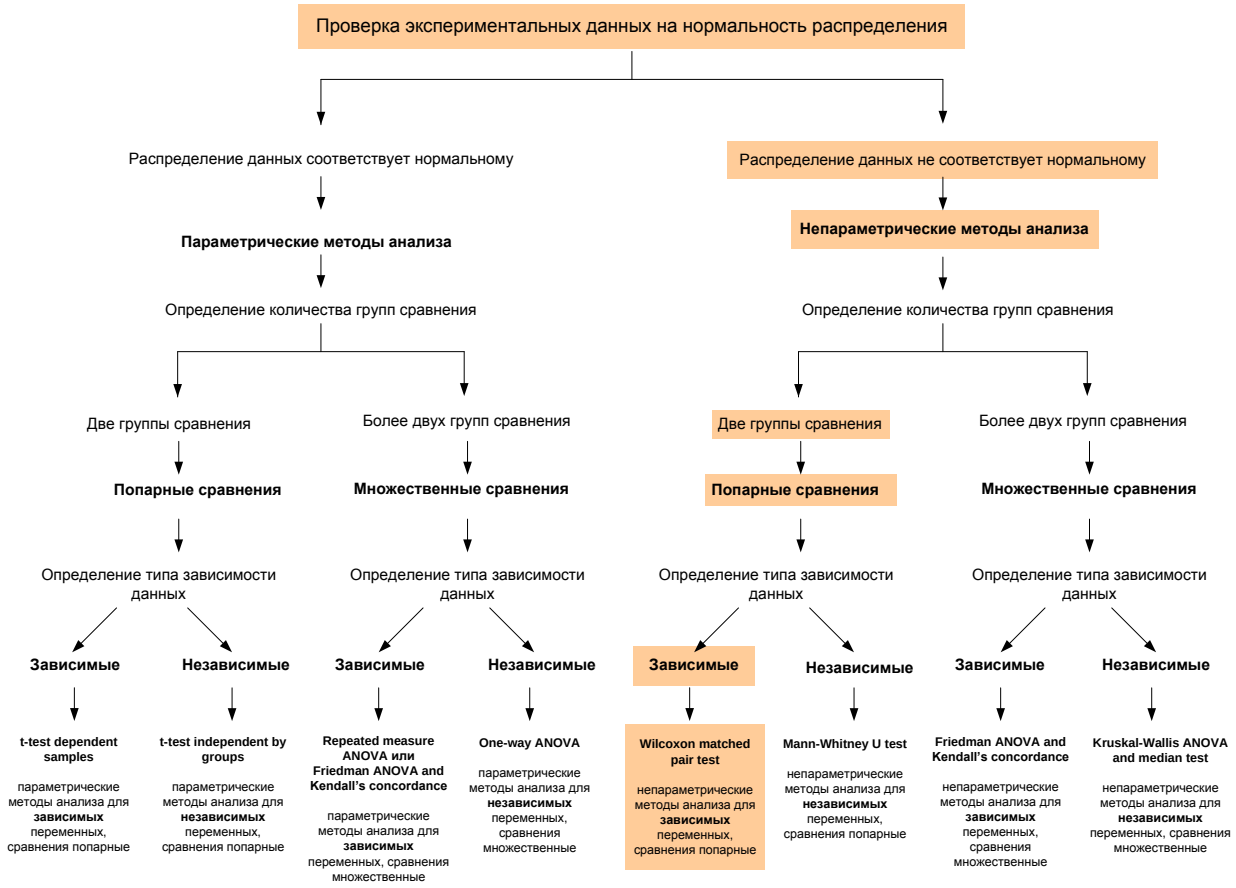


Рис. 28. Схема выбора метода статистического анализа при сравнении двух зависимых групп при условии несоответствия данных закону нормального распределения

Для демонстрации работы данного теста воспользуемся предыдущим примером, но допустим, что условие о «нормальности» распределения экспериментальных данных не выполняется.

Тест Уилкоксона можно запустить следующим образом: внести в таблицу результаты исследования в соответствии с правилами оформления данных для зависимых групп (см. стр. 19-20).

1. В разделе **Statistics / Nonparametrics / Comparing dependent samples** выбрать t-тест для зависимых выборок (рис. 29).

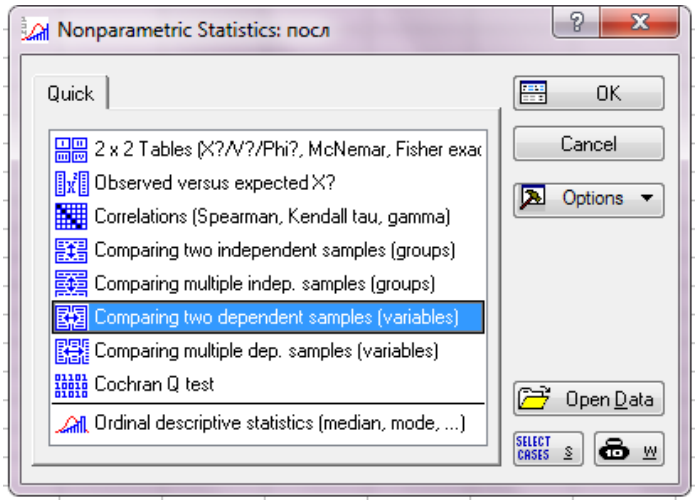


Рис. 29. Диалоговое окно модуля Основная статистика

2. Нажать кнопку **Variables**, задать переменные для анализа и нажать кнопку **Wilcoxon matched pair test** (тест Уилкоксона) (рис. 30).

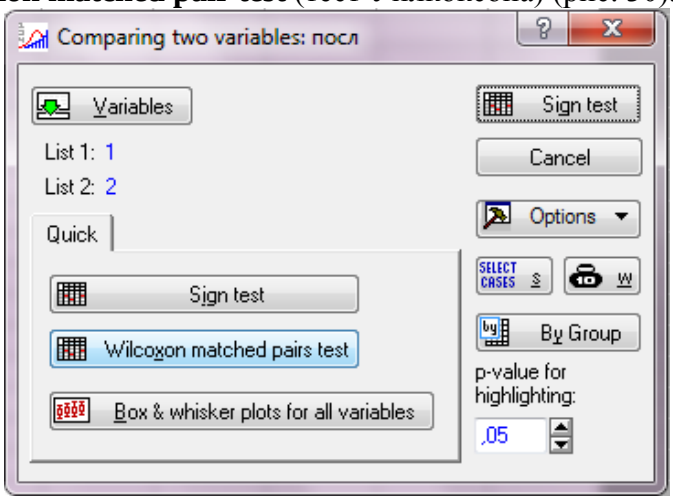


Рис. 30. Диалоговое окно для выбора исследуемых переменных

В результате появится таблица (рис. 31). Поскольку P>0.05, следовательно статистически значимые различия между сравниваемыми выборками отсутствуют.

Pair of Variables	Valid N	T	Z	p-value
	Wilcoxon Matched Pairs Test (посл) Marked tests are significant at p <,05000			
IgG мышей до физ. нагрузки & IgG мышей после физ.нагрузки	15	29,50000	1,732284	0,083224

Рис. 31. Таблица с результатами теста Уилкоксона.

Раздел 3. Множественные сравнения (сравнения нескольких групп).

Тест Стьюдента и его непараметрические аналоги, рассмотренные выше, предназначены для сравнения **исключительно двух выборок**. Однако очень часто данный тест используется для сравнений в 3 и более выборках, что резко повышает вероятность ошибки первого рода (ошибка 1-го рода – это вероятность ложно отклонить нулевую гипотезу, т.е. найти различия там, где их нет). Максимально допустимая вероятность этой ошибки равна 5%.

Допустим, необходимо провести сравнения 3 независимых групп. Для этого предполагается провести 3 попарных сравнения: гр. 1 х 2; гр. 1 х 3 и гр. 2 х 3. Это означает, что контроль ошибки первого рода можно обеспечить, только разделив значение номинального уровня значимости на количество попарных сравнений. В данном случае 3, получаем 0.05/3=0.017. Таким образом, нулевая гипотеза отвергается, если достигаемый уровень значимости при использовании парного критерия Стьюдента P< 0.017.

Во избежание данной ошибки необходимо использовать специальные методы статистического анализа для множественных сравнений. Алгоритм выбора данного метода статистического анализа можно представить в виде следующей схемы:

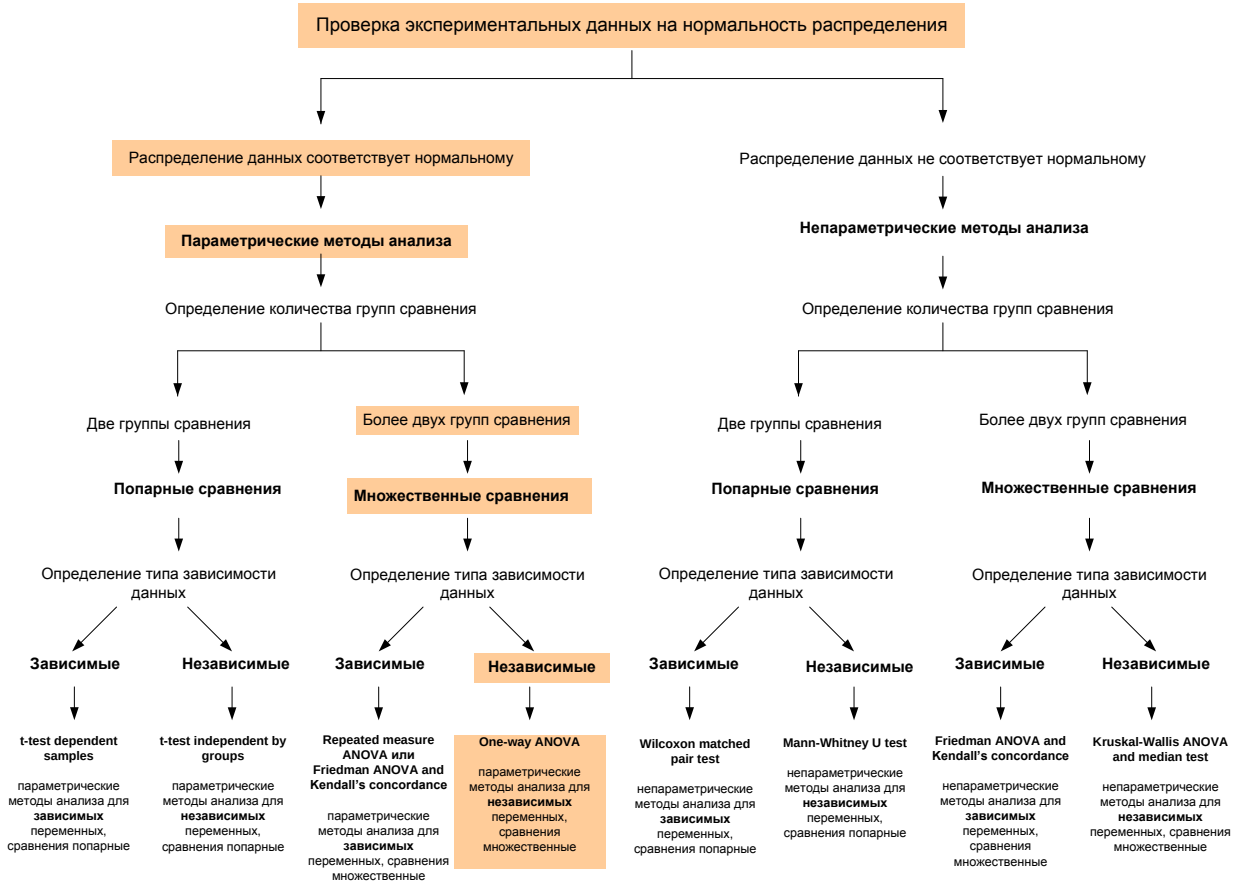


Рис. 32. Схема выбора метода статистического анализа при сравнении трех и более независимых групп при условии соответствия данных закону нормального распределения

Однофакторный дисперсионный анализ (One-way ANOVA).

В качестве примера, требующего использования дисперсионного анализа, можно рассмотреть исследование биохимических показателей крови в нескольких группах больных, страдающих болезнью Паркинсона, а также в группе условно здоровых лиц. Первая группа – условно здоровые лица (контроль), вторая группа больные на ранней стадии заболевания, третья группа – больные на поздней стадии заболевания. В указанных группах исследовали такой биохимический показатель, как количество альфа-синуклеина в плазме крови.

Поскольку количество исследуемых групп и, как следствие, групп сравнения более двух, воспользуемся однофакторным дисперсионным анализом. Для выполнения анализа данного типа необходимо внести в таблицу результаты исследования в соответствии с правилами оформления данных для независимых групп (см. стр. 14).

1. Из меню **Statistics** запустить модуль **One-way ANOVA** (рис. 33). Далее выбрать тип анализа **One-way ANOVA**.

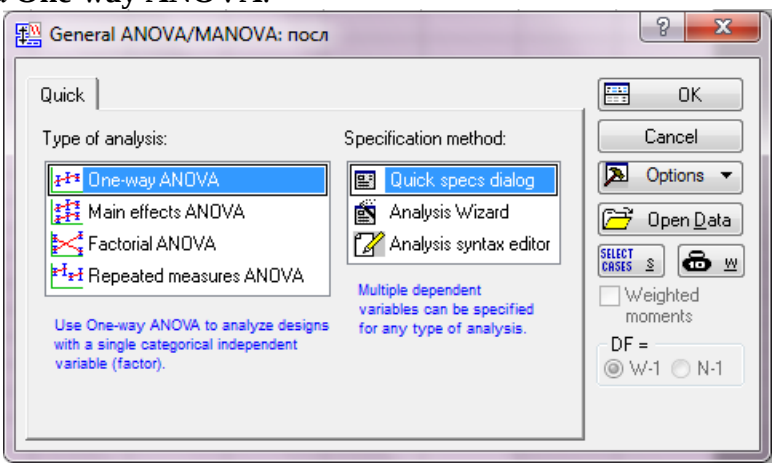


Рис. 33. Диалоговое окно модуля дисперсионный анализ

2. Нажать на кнопку **Variables** и выбрать зависимую и группирующую переменные

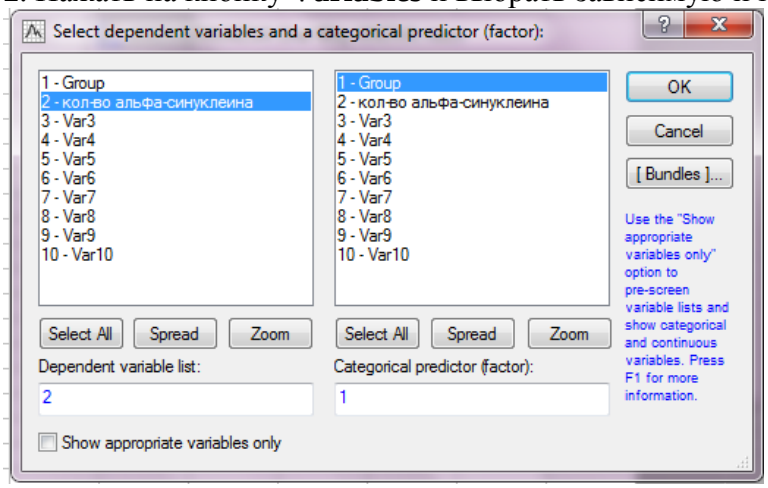


Рис. 34. Диалоговое окно для выбора исследуемых переменных

3. Нажать на кнопки: **Factor codes / All** (это укажет программе, что необходимо проанализировать все экспериментальные группы) **/ ОК / ОК** (рис. 35).

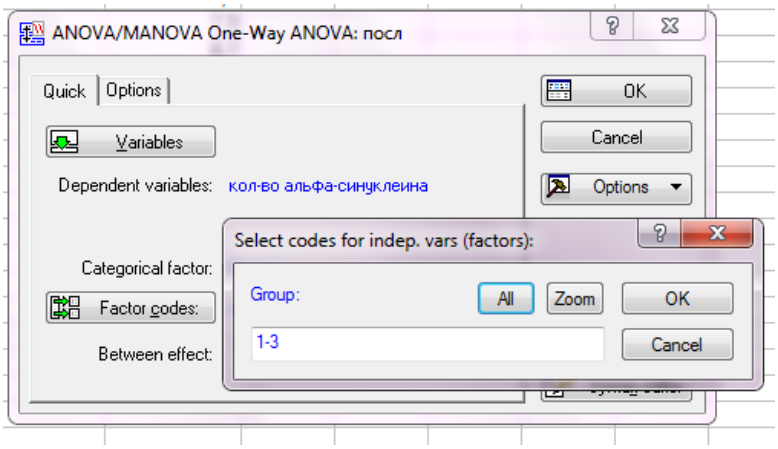


Рис. 35. Диалоговое окно для выбора кодов исследуемых групп

В результате появится окно с 8 закладками (рис. 36), автоматически открытое на закладке **Quick** (Быстро).

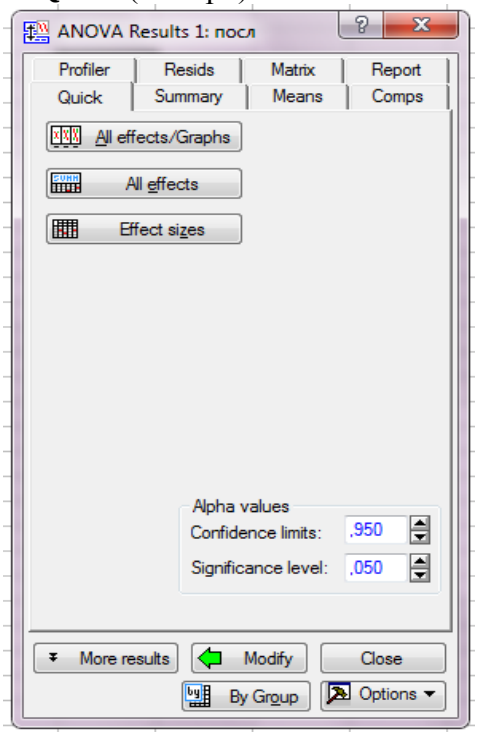


Рис. 36. Диалоговое окно выбора результатов дисперсионного анализа

Нажав на кнопку **All effects** (Все эффекты) можно быстро получить результаты анализа. Однако рассматриваемый вариант анализа является параметрическим, следовательно, требует выполнения ряда обязательных условий:

1) Однородность дисперсий (отсутствие статистически значимой разницы между показателями разброса данных в группах);

2) Подчинение данных (во всех группах) закону нормального распределения;

3) Независимый характер выборок.

В связи с этим следует провести проверку выборки на предмет соответствия данным требованиям. Для проверки однородностей дисперсий необходимо:

1. Нажать на кнопку **More results** (Дополнительные результаты), расположенную в нижней части окна **ANOVA Results**.

2. В появившемся окне (рис. 37) открыть закладку **Assumptions** (Допущения).

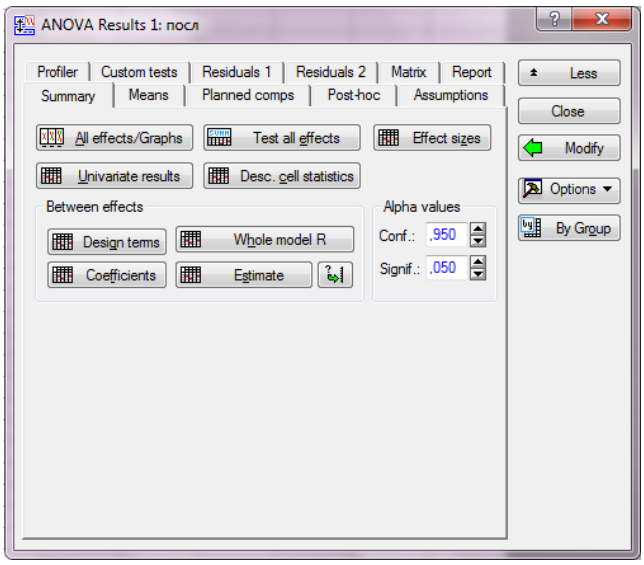


Рис. 37. Диалоговое окно дополнительных результатов дисперсионного анализа

3. В разделе **Homogeneity of variances/covariances** нажать на кнопку **Levene's test** (тест Левена) (рис. 38).

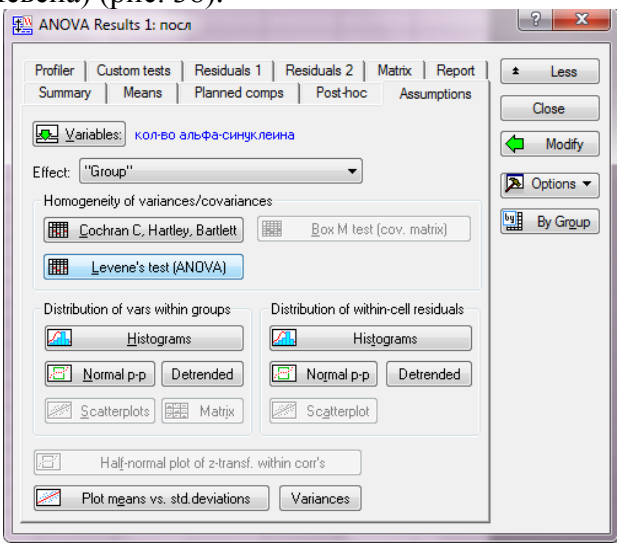


Рис. 38. Диалоговое окно дополнительных результатов дисперсионного анализа

В появившейся таблице представлены результаты теста на сравнение дисперсий (рис. 39). Если различия между дисперсиями отсутствуют (P>0.05), то применение параметрического варианта дисперсионного анализа обосновано. В нашем случае различия отсутствуют (P = 0.12).

	Levene's Test for Homogeneity of Variances (посл) Effect: "Group" Degrees of freedom for all F's: 2, 21			
	MS Effect	MS Error	F	p
кол-во альфа-синуклеина	35,19565	15,09321	2,331887	0,121741

Рис. 39. Результаты теста на различия между дисперсиями

Для проверки «нормальности» распределения анализируемых данных можно использовать опцию, доступную в поле **Distribution of variables within groups** (Распределение переменных внутри групп). Однако данную операцию лучше выполнить заранее, воспользовавшись специальным модулем – **Distribution fitting** (Настройка распределения).

При соблюдении всех условий необходимо на закладке **Summary** (Итоги) нажать кнопку **Test all effects** (Проверить все эффекты) (рис. 40).

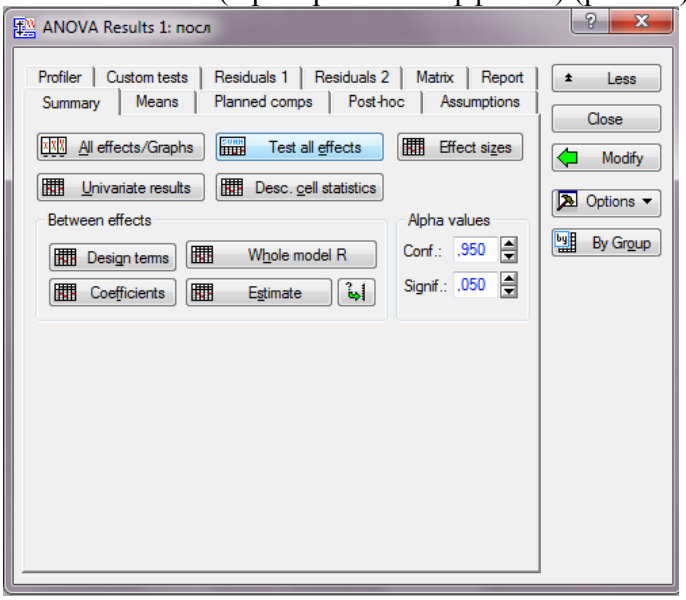


Рис. 40. Диалоговое окно выбора результатов дисперсионного анализа

В появившейся таблице (рис. 41) необходимо разыскать ячейку с величиной ошибки P. Поскольку в нашем примере $P < 0.05$, можно заключить, что количество альфа-синуклеина в плазме больных различных групп статистически значимо различается.

Univariate Tests of Significance for кол-во альфа-синуклеина (
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	1834,555	1	1834,555	44,44760	0,000001
Group	498,408	2	249,204	6,03772	0,008482
Error	866,766	21	41,275		

Рис. 41. Результаты дисперсионного анализа

Апостериорный анализ (Post-hoc analysis).

При проведении дисперсионного анализа важно понимать, что он позволяет проверить лишь гипотезу об отсутствии различий между сравниваемыми группами в целом. Узнать, какие именно группы различаются между собой, с его помощью невозможно. Для выяснения этого вопроса используют методы множественных сравнений, являющихся частью так называемого апостериорного анализа **Post-hoc analysis**. Данные методы позволяют провести попарные сравнения средних значений всех групп, включенных в дисперсионный анализ.

Для выполнения апостериорных сравнений в окне **ANOVA Results** нужно нажать кнопку **More results** (дополнительные результаты) (рис. 36). Далее необходимо открыть закладку **Post hoc** (апостериорные сравнения) (рис. 42).

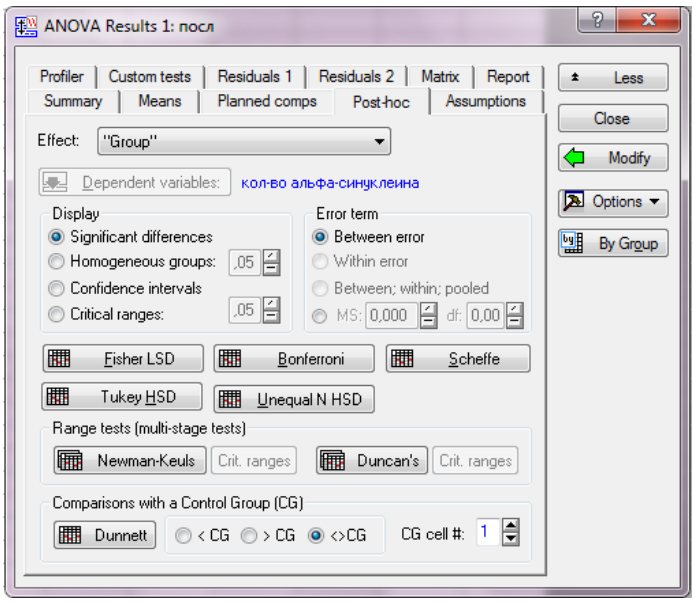


Рис. 42. Диалоговое окно выбора методов апостериорного анализа

В программе STATISTICA предложено несколько разновидностей тестов для множественных сравнений: **Fisher LSD, Bonferroni, Scheffe, Tukey HSD, Newman-Keuls, Duncan's, Dunnet** (они несколько отличаются по мощности). Наиболее часто используемыми являются тесты Тьюки (**Tukey HSD**) и Ньюмена-Кейлса (**Newman-Keuls**).

Нажав на кнопку соответствующего теста, можно получить таблицу с матрицей значений P (рис. 43).

LSD test; variable кол-во альфа-синуклеина (посл)
Probabilities for Post Hoc Tests
Error: Between MS = 41,275, df = 21,000

Cell No.	Group	{1} 15,159	{2} 5,0133	{3} 6,0562
1	1		0,004737	0,009943
2	2	0,004737		0,748631
3	3	0,009943	0,748631	

Рис. 43. Результаты апостериорного анализа

Из рисунка 43 видно, что статистически значимая разница в количестве альфа-синуклеина наблюдается между контрольной группой и больными на первой стадии заболевания, а также между контрольной группой и больными на второй стадии заболевания. Между группами больных статистически значимая разница в количестве альфа-синуклеина отсутствует.

Дисперсионный анализ Фридмана (Friedman ANOVA and Kendall's concordance).

Дисперсионный анализ Фридмана (Friedman ANOVA) применяется в том случае, если исследуемые выборки являются взаимосвязанными (зависимыми). Важно отметить что, являясь непараметрическим, он не требует соблюдения условий о «нормальности» распределения и однородности дисперсий в исследуемых группах.

Алгоритм выбора данного метода статистического анализа можно представить в виде следующей схемы:

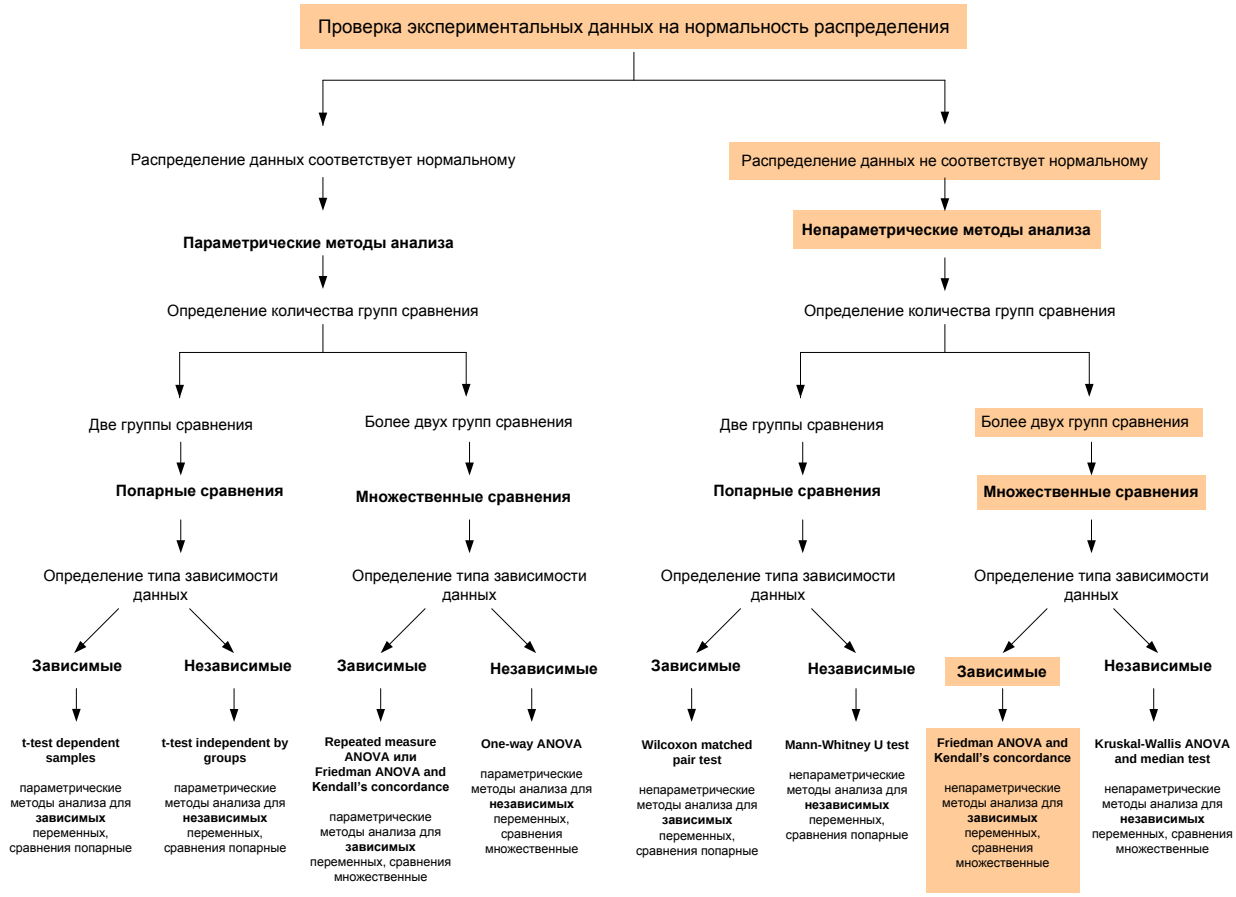


Рис. 44. Схема выбора метода статистического анализа при сравнении трех и более зависимых групп при условии несоответствия данных закону нормального распределения

На рис. 44 представлены данные об изменении количества фосфолипидов крови у спортсменов во время тренировочного процесса, а также в соревновательный период 2011-2012 годов. Необходимо выяснить, произошли ли существенные изменения количества фосфолипидов к концу спортивного сезона. Поскольку в исследовании принимали участие одни и те же спортсмены, полученные выборки являются взаимосвязанными (зависимыми). Используем для их сравнения дисперсионный анализ Фридмана.

Внесем в таблицу результаты исследования в соответствии с правилами оформления данных для зависимых групп (см. стр. 20-21) (рис. 45).

	1 24.11.2011	2 25.12.2011	3 22.01.2012	**4** **23.02.2012**
1	1,19	1,77	1,56	1,22
2	1,8	1,09	1,39	2,21
3	1,9	1,21	1,34	2,89
4	2,12	1,02	2,24	2,3
5	1,06	0,72	1,7	2,1
6	1,5	1,3	1,47	1,11
7	1,7	1,4	1,6	1,4
8				

Рис. 45. Пример оформления данных при сравнении трех и более зависимых групп

1. Запустить модуль анализа из меню: **Statistics / Nonparametric / Comparing multiple dependent samples** (Сравнение нескольких зависимых выборок) (рис. 46).

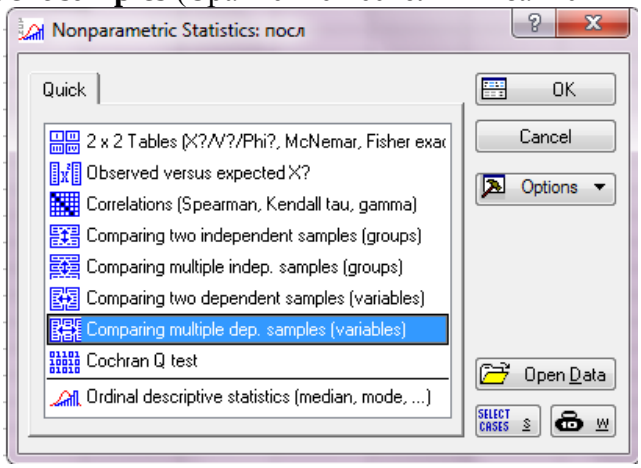


Рис. 46. Диалоговое окно модуля дисперсионный анализ

2. Нажать кнопку **Variables** и выбрать переменные, которые необходимо проанализировать (рис. 47).

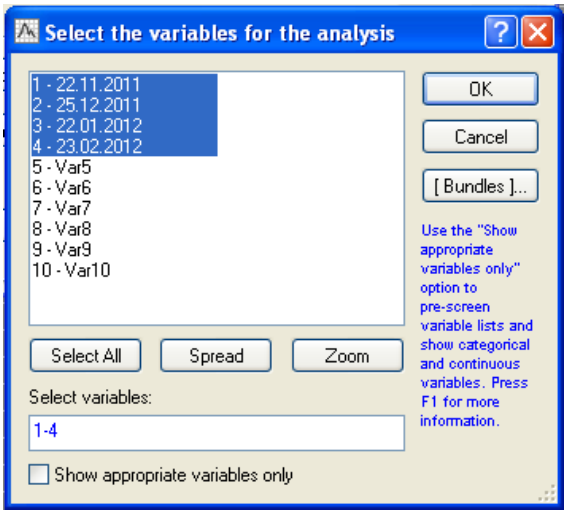


Рис. 47. Диалоговое окно для выбора исследуемых переменных

3. Нажать кнопку **Summary: Friedman ANOVA and Kendall's concordance** (Итоги: ANOVA по Фридману и критерий согласованности Кендалла) (рис. 48).

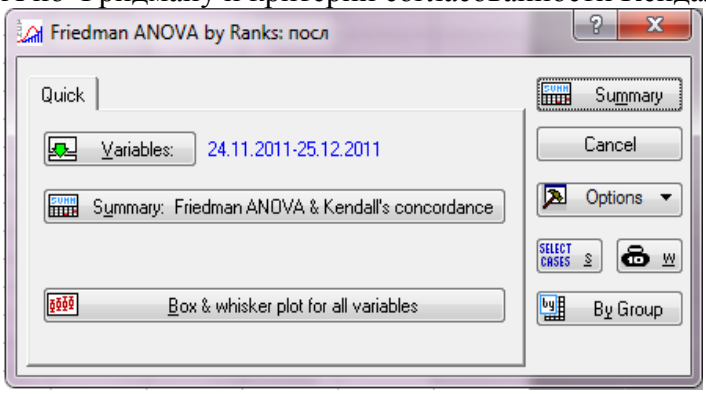


Рис. 48. Диалоговое окно для запуска дисперсионного анализа

4. В таблице с результатами найти величину ошибки P (расположена в заголовке таблицы) (рис. 49). Поскольку в нашем случае P>0.05, следовательно, достоверные отличия между количеством фосфолипидов в исследуемых группах отсутствуют.

В этом же заголовке приводится так называемый коэффициент согласованности Кендалла (Coeff. of Concordance). Чем ближе коэффициент Кендалла к 1, тем больше различия между группами.

Friedman ANOVA and Kendall Coeff. of Concordance (Spreadsheet1)
ANOVA Chi Sqr. (N = 7, df = 3) = 4,304348 p = ,23042
Coeff. of Concordance = ,20497 Aver. rank r = ,07246

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
22.11.2011	2,714286	19,00000	1,610000	0,382840
25.12.2011	1,642857	11,50000	1,215714	0,328677
22.01.2012	2,714286	19,00000	1,614286	0,302316
23.02.2012	2,928571	20,50000	1,890000	0,659798

Рис. 49. Результаты дисперсионного анализа

Однако данный метод содержит тот же недостаток, что и его параметрический аналог (однофакторный дисперсионный анализ). Он позволяет проверить лишь гипотезу об отсутствии различий между сравниваемыми группами в целом. Узнать, какие именно группы различаются между собой, данный вид анализа не позволяет.

Дисперсионный анализ Крускала-Уоллиса (Kruskal-Wallis ANOVA).

Как уже отмечалось выше, экспериментальные данные, полученные в ходе биологических исследований, достаточно редко подчиняются закону нормального распределения. Более того, очень часто объем выборок оказывается слишком малым, что не позволяет сделать какие-либо выводы относительно типа распределения. Все это делает применение параметрического дисперсионного анализа невозможным.

Одним из способов выхода из данной ситуации является применение непараметрического дисперсионного анализа Крускала-Уоллиса (или Н-теста) (Kruskal-Wallis ANOVA).

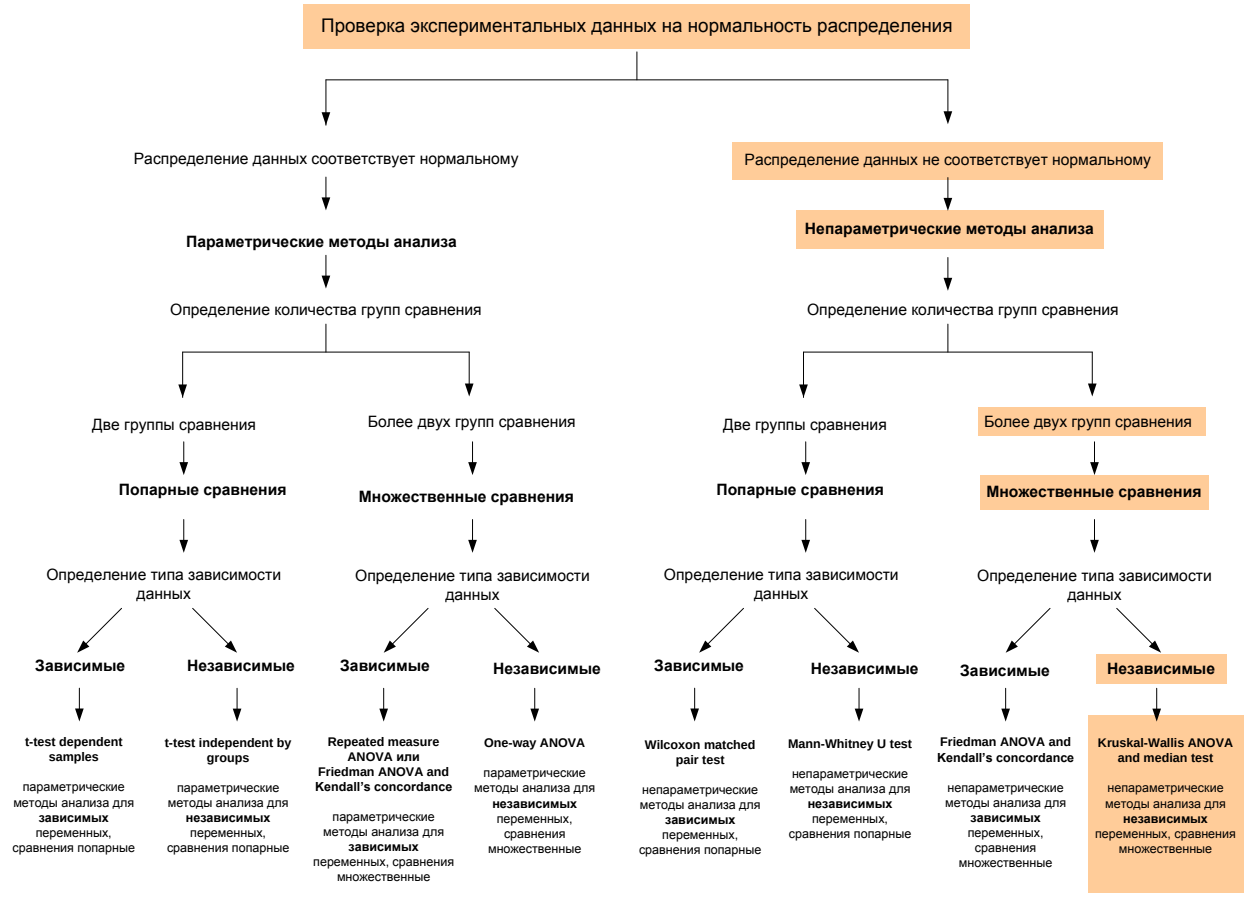


Рис. 50. Схема выбора метода статистического анализа при сравнении трех и более независимых групп, при условии несоответствия данных закону нормального распределения

Разберем применение дисперсионного анализа Крускала-Уоллиса на следующем примере: на рис. 51 представлены данные о количестве глиальных клеток в разных структурах головного мозга.

	1 Group	2 кол-во глиальных клеток
1	черн. субстанц.	87
2	черн. субстанц.	109
3	черн. субстанц.	41
4	черн. субстанц.	79
5	черн. субстанц.	106
6	черн. субстанц.	126
7	черн. субстанц.	85
8	черн. субстанц.	77
9	красн. ядра	32
10	красн. ядра	15
11	красн. ядра	20
12	красн. ядра	18
13	красн. ядра	24
14	красн. ядра	25
15	интраламинар. ядра	27
16	интраламинар. ядра	38
17	интраламинар. ядра	27
18	интраламинар. ядра	29
19	интраламинар. ядра	51
20		

Рис. 51. Пример оформления данных при сравнении трех и более независимых групп

Как видно, число наблюдений в каждом отделе невелико (n не более 8), что не позволяет корректно оценить характер распределения данных. Если прибегнуть к небольшой хитрости и объединить все имеющиеся данные в одну совокупность, окажется, что их распределение не соответствует закону нормального распределения.

Чтобы выяснить, различается ли количество клеток в разных отделах мозга, применим дисперсионный анализ Крускала-Уоллиса. Обратите внимание, данные внесены в таблицу в соответствии с правилами оформления для независимых групп (см. стр. 13-14).

1. Запустить из меню модуль **Statistics / Nonparametrics / Comparing multiple independent samples** (Сравнение нескольких независимых выборок) (рис. 52).

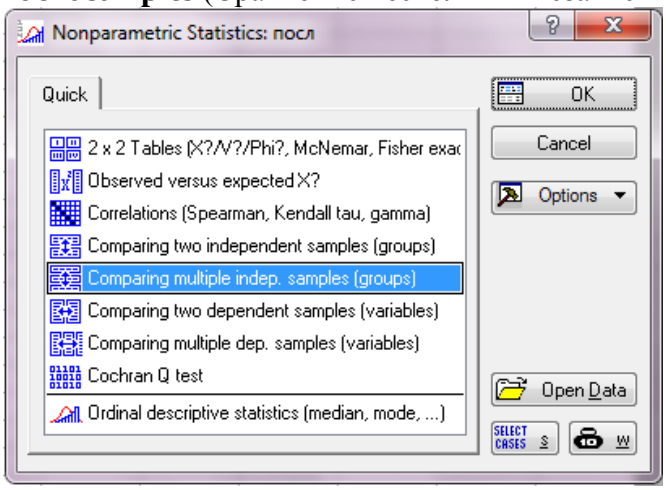


Рис. 52. Диалоговое окно для выбора дисперсионного анализа

2. Нажать кнопку **Variables** (Переменые) и выбрать зависимую («количество клеток») и группирующую переменные (рис. 53).

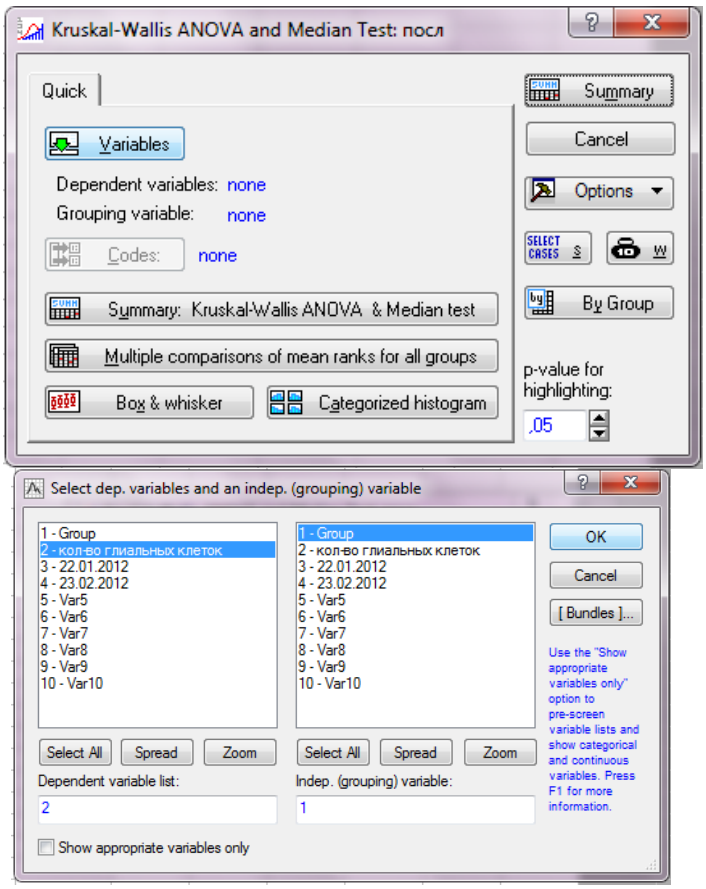


Рис. 53. Диалоговые окна для выбора исследуемых переменных

3. Далее нажать на кнопки: **Factor codes / All / ОК** (рис. 54).

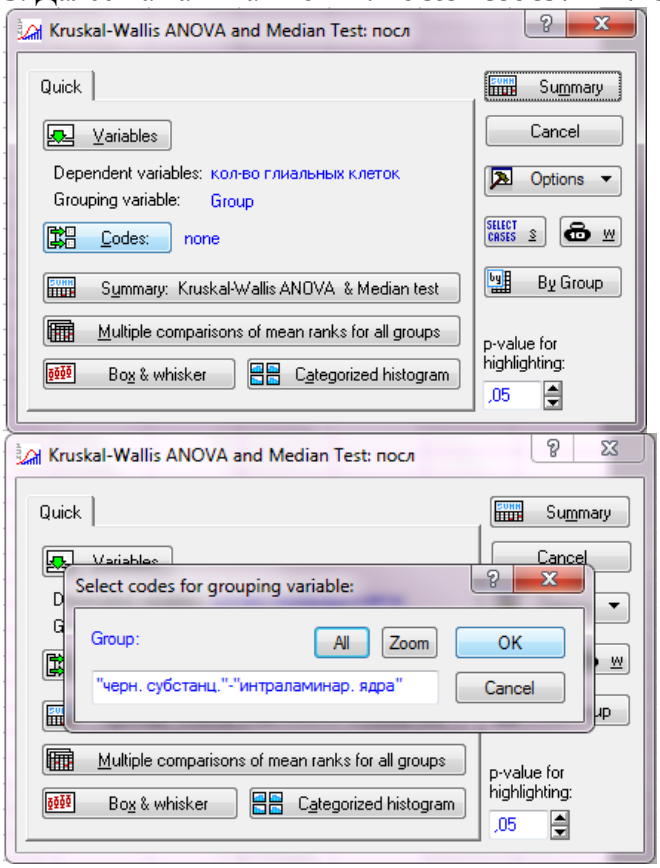


Рис. 54. Диалоговые окна для выбора кодов исследуемых групп

4. Нажать на кнопку **Summary: Kruskal-Wallis ANOVA and Median test** (Результат: ANOVA по Крускалу-Уоллису и медианный тест) (рис. 55).

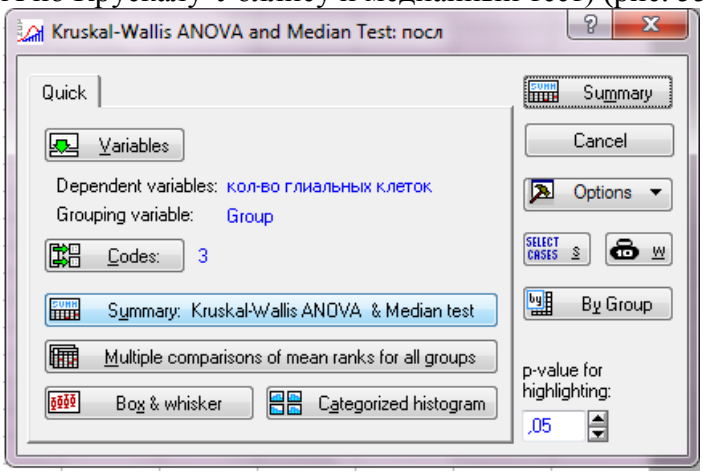


Рис. 55. Диалоговое окно для запуска дисперсионного анализа

5. В таблице с результатами (рис. 56) найти величину ошибки P для нулевой гипотезы о том, что количество глиальных клеток в разных структурах головного мозга не различается.

Kruskal-Wallis ANOVA by Ranks; кол-во глиальных клеток (посл)
Independent (grouping) variable: Group
Kruskal-Wallis test: H (2, N= 19) =14,44188 p =,0007

Depend.: кол-во глиальных клеток	Code	Valid N	Sum of Ranks	Mean Rank
черн. субстанц.	101	8	123,0000	15,37500
красн. ядра	102	6	24,0000	4,00000
интраламинар. ядра	103	5	43,0000	8,60000

Рис. 56. Результаты дисперсионного анализа

Если P<0.05 (как в нашем примере), исследуемые группы статистически достоверно отличаются друг от друга. Помимо результатов теста Крускал-Уоллиса программа предлагает результаты так называемого медианного теста. Он проверяет ту же нулевую гипотезу, но является менее мощным.

Факторный анализ (Factorial ANOVA).

Факторный анализ предназначен для выявления влияния различных факторов (условий) или их комбинации на изменение исследуемого признака. Иными словами, позволяет исследовать зависимость какого-либо количественного (зависимого) признака от одного или нескольких качественных признаков (факторов).

Рассмотрим работу факторного анализа, используя следующий пример: допустим, нам необходимо ответить на вопрос: как возраст, пол и уровень образования влияют на показатели артериального давления?

Предположим, что по возрасту всех испытуемых, принимавших участие в исследовании, можно разделить на 4 группы:

гр. 1: до 30 лет;
гр. 2: от 31 до 40 лет;
гр. 3: от 41 до 50 лет;
гр. 4: более 51 года.

По уровню образования испытуемых можно разделить на 3 группы:
гр. 1: высшее образование;
гр. 2: среднее образование;
гр. 3: без образования.

Кроме того, нам известны показатели систолического артериального давления (САД) каждого участника исследования, а также его пол.

Таким образом, САД это количественный признак, а «Возраст», «Пол» и «Образование» это факторы, влияние которых нам необходимо исследовать.

Прежде чем приступить к анализу, внесем данные в таблицу. Обратите внимание, все данные расположены в вертикальных колонках. Первые 3 колонки качественные признаки (факторы) или группирующие переменные, четвертая колонка зависимая переменная (рис. 57).

	1 пол	2 возраст	3 образование	4 САД
1	M	1	c	120
2	W	1	v	125
3	M	1	n	130
4	M	2	c	141
5	M	3	v	160
6	M	4	n	161
7	M	4	n	135
8	W	3	c	123
9	M	1	v	130
10	M	2	v	128
11	M	2	v	141
12	M	3	c	161
13	M	3	n	167
14	W	4	n	125
15	W	1	n	129
16	W	2	v	142
17	M	3	c	165
18	M	4	c	161
19	M	4	c	159

Рис. 57. Пример оформления данных для выполнения факторного анализа

1. Из меню **Statistics / ANOVA** запустить модуль **Factorial ANOVA** (Факторный дисперсионный анализ) (рис. 58).

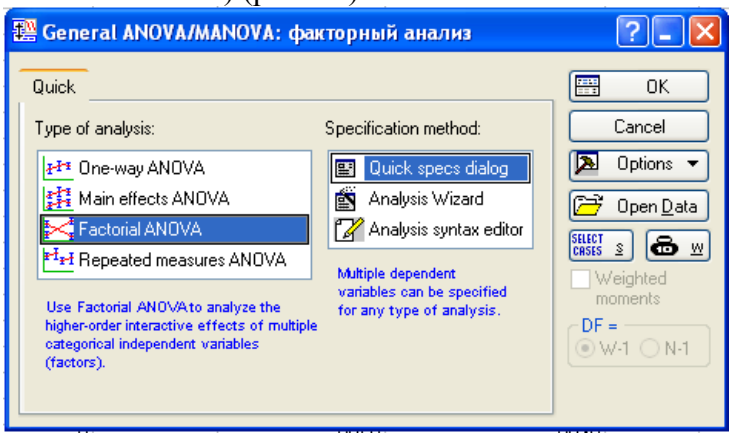


Рис. 58. Диалоговое окно для выбора факторного анализа

2. В появившемся окне нажать кнопку **Variables** и выбрать группирующие переменные (Пол, Возраст, Образование) и зависимую переменную (САД) (рис. 59).

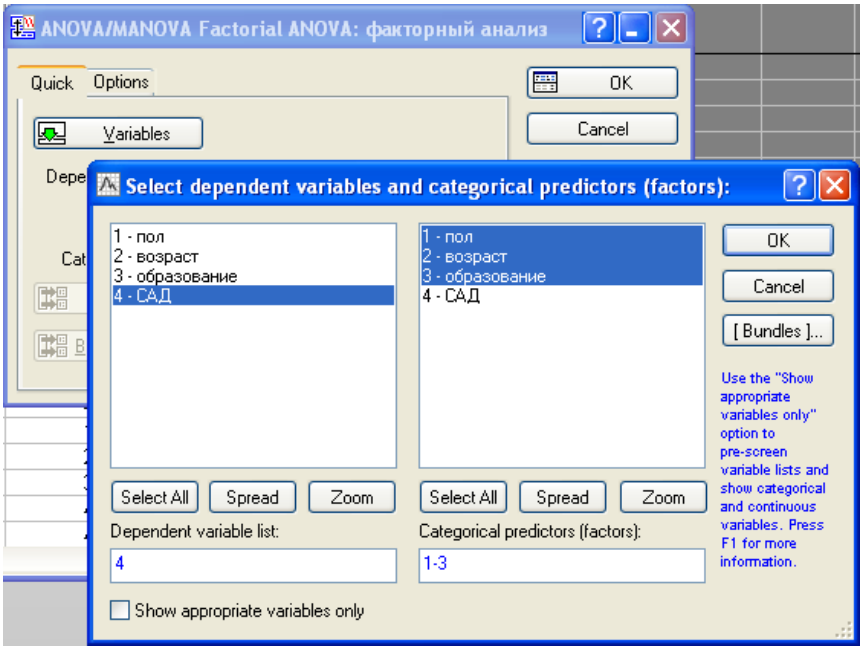


Рис. 59. Диалоговое окно для выбора группирующих и зависимой переменных

3. Далее нажать кнопку **Factor code** (Коды факторов), указать коды факторов, влияние которых необходимо оценить. Поскольку факторов не очень много, нажимаем кнопку **All** (выбрать все) (рис. 60).

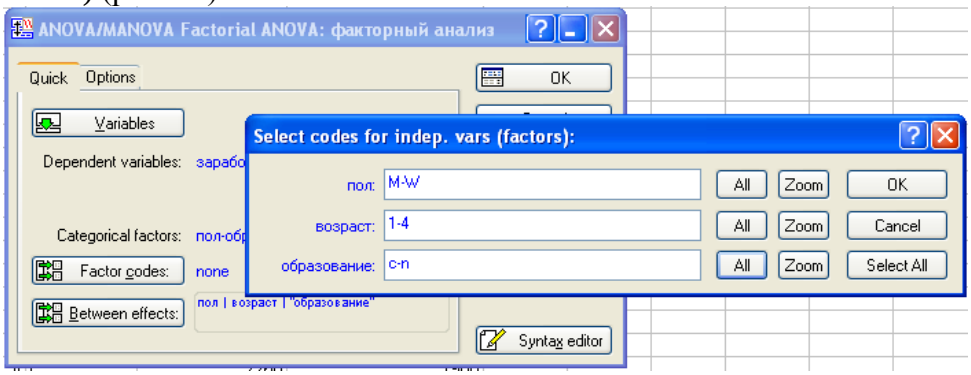


Рис. 60. Диалоговое окно для выбора кодов факторов

4. Коды факторов можно и не задавать: если нажать кнопку **ОК**, программа задаст их автоматически. В итоге появится окно с 8 закладками. Выбрав закладку **Summary** (Итоги) нажимаем кнопку **Test All effects** (проверить все эффекты) (рис. 61).

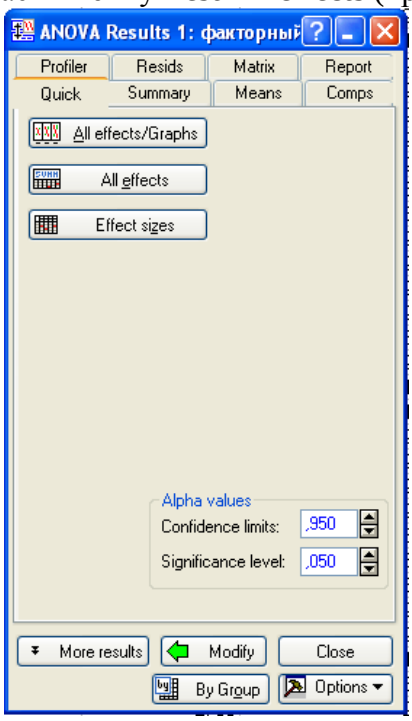


Рис. 61. Диалоговое окно для выбора кодов факторов

В итоге появится таблица с результатами дисперсионного анализа (рис. 62). В конце строк этой таблицы приведены вероятности ошибок для нулевых гипотез об отсутствии влияния фактов «Пола», «Возраста» и «Образования» на уровень САД.

Univariate Tests of Significance for САД (факторн
Sigma-restricted parameterization
Effective hypothesis decomposition

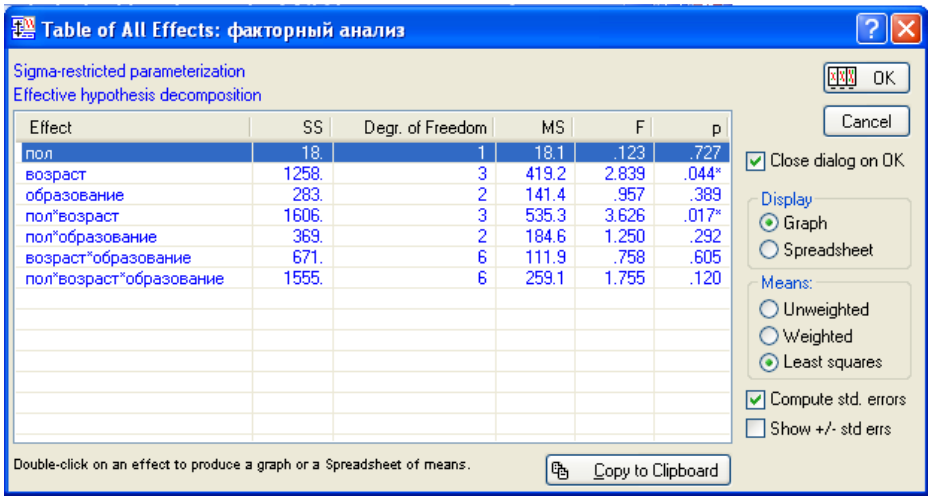
Effect	SS	Degr. of Freedom	MS	F	p
Intercept	1267884	1	1267884	8587.862	0.00000
пол	18	1	18	0.123	0.72720
возраст	1258	3	419	2.839	0.04358
образование	283	2	141	0.957	0.38852
пол*возраст	1606	3	535	3.626	0.01675
пол*образование	369	2	185	1.250	0.29237
возраст*образование	671	6	112	0.758	0.60528
пол*возраст*образование	1555	6	259	1.755	0.12008
Error	11073	75	148		

Рис. 62. Результаты факторного анализа

Из таблицы видно, что наиболее значительное влияние на САД оказывает «Возраст»: P<0.05. Доказать подобный эффект для «Пола» и «Образования» нам в данном эксперименте не удалось ($P > 0.05$). Строки «Пол \ Возраст» и т. д. касаются взаимного влияния исследуемых факторов на САД. Как видно, наибольшее взаимодействие (взаимовлияние) на величину САД оказывают факторы «Пол» и «Возраст» ($P > 0.05$).

Для оценки различий средних значений САД по категориям, можно воспользоваться графическими средствами.

1. Нажать на кнопку **All effects/Graphs** (Все эффекты/графики). В полученном окне (рис. 63) перечислены все рассматриваемые эффекты. Статистически значимые эффекты помечены *.



Effect	SS	Degr. of Freedom	MS	F	p
пол	18.	1	18.1	.123	.727
возраст	1258.	3	419.2	2.839	.044*
образование	283.	2	141.4	.957	.389
пол*возраст	1606.	3	535.3	3.626	.017*
пол*образование	369.	2	184.6	1.250	.292
возраст*образование	671.	6	111.9	.758	.605
пол*возраст*образование	1555.	6	259.1	1.755	.120

Рис. 63. Дополнительные результаты факторного анализа

2. Выбрать эффект **Возраст**, в группе **Display** (Отображать), указать **Spreadsheet** (Таблица) и нажать **ОК.**

В появившейся таблице для каждого уровня эффекта приведены средние значения зависимой переменной САД, величина стандартной ошибки и границы доверительных пределов (рис. 64).

возраст; LS Means (факторный анализ)
Current effect: F(3, 75)=2.8392, p=.04358
Effective hypothesis decomposition

Cell No.	возраст	САД Mean	САД Std.Err.	САД -95.00%	САД +95.00%	N
1	1	135.6667	3.788611	128.1194	143.2140	15
2	2	141.6071	3.061660	135.5080	147.7063	25
3	3	145.3899	3.033626	139.3466	151.4332	28
4	4	147.7734	2.227169	143.3367	152.2102	31

Рис. 64. Таблица показателей систолического артериального давления (САД) во всех возрастных группах

Эту таблицу удобно представить в графическом виде. Для этого в группе **Display** (Отображать) выбрать **Graph** (График) и нажать **ОК.** Появится соответствующий график (рис. 65).

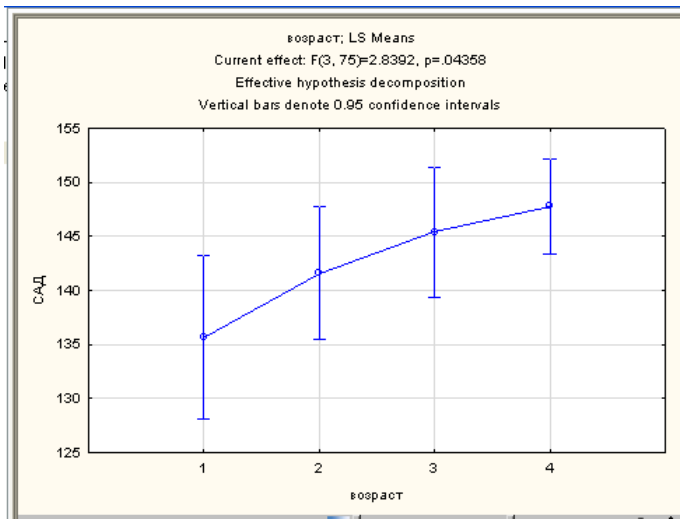


Рис. 65. График показателей среднего дохода в разных возрастных группах

На графике хорошо видно, как изменяются средние значения САД в разных возрастных группах (в зависимости от фактора «Возраст»).

Помимо оценки влияния исследуемых факторов («Возраст», «Образование» и «Пол») и их взаимодействия на показатели САД, очень важно понимать, какую долю изменчивости они объясняют.

1. Во вкладке **Summary** (Итоги) нажать на кнопку **Whole model R** (Общие R модели) (рис. 66), после чего появится соответствующая таблица (рис. 67).

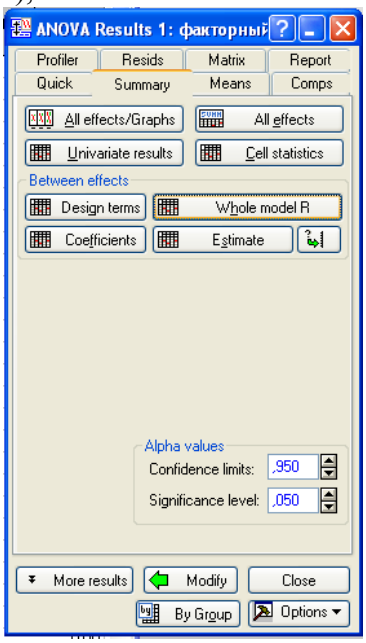


Рис 66. Диалоговое окно для выбора кодов факторов

Dependent Variable	Test of SS Whole Model vs. SS Residual (факторный анализ)										
	Multiple R	Multiple R?	Adjusted R?	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
САД	0.586255	0.343695	0.142428	5798.595	23	252.1128	11072.76	75	147.6368	1.707656	0.043807

Рис. 67. Таблица SS модели и SS остатков

Наибольший интерес представляет коэффициент **R (Multiple R)** – квадрат множественного коэффициента корреляции или коэффициент детерминации. Он показывает, какую долю изменчивости объясняет построенная модель. Чем ближе R к единице, тем лучше построена модель. В нашем случае R2 = 0.58, что говорит о не очень хорошем качестве модели. Можно сказать, что уровень САД на 58% определяется факторами, включенными в нашу модель («Полом», «Возрастом» и «Образованием»). Тем не менее, на уровень САД влияют и другие факторы, не учтенные в нашей модели.

Дисперсионный анализ с повторными измерениями (Repeated measure ANOVA).

Описанный выше вид дисперсионного анализа применяется только в том случае, если зависимая переменная только одна. Если зависимых переменных несколько и они являются результатом повторных измерений одного и того же признака, применяются методы дисперсионного анализа для повторных измерений.

Разберем данный вид анализа на аналогичном примере (см. факторный анализ). Оценим влияние факторов «Возраст», «Пол» и «Образование» на величину систолического артериального давления (САД). Допустим, что САД измерялось дважды через определенный промежуток времени. Таким образом, мы будем иметь дело с повторным измерением одного и того же признака.

Внесем данные повторных измерений в таблицу в отдельную колонку (рис. 68).

	1 пол	2 возраст	3 образование	4 САД	5 САД1
1	M	1	c	120	123
2	W	1	v	125	125
3	M	1	n	130	134
4	M	2	c	141	140
5	M	3	v	160	163
6	M	4	n	161	162
7	M	4	n	135	136
8	W	3	c	123	127
9	M	1	v	130	132
10	M	2	v	128	127
11	M	2	v	141	140
12	M	3	c	161	167
13	M	3	n	167	167
14	W	4	n	125	128
15	W	1	n	129	134
16	W	2	v	142	149
17	M	3	c	165	164
18	M	4	c	161	163
19	M	4	c	159	161
20	M	3	v	133	131
21	M	2	v	121	124
22	M	1	n	137	138
23	W	4	v	147	150
24	W	3	n	131	130
25	W	2	n	135	136

Рис. 68. Пример оформления данных для выполнения факторного анализа с повторными измерениями

1. Из меню **Statistics / ANOVA** запустить модуль **Repeated measure ANOVA** (Дисперсионный анализ повторные измерения) (рис. 69).

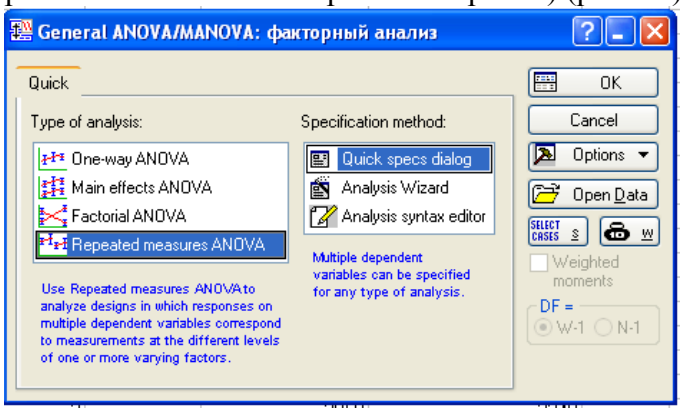


Рис. 69. Диалоговое окно для выбора факторного анализа с повторными измерениями

2. В появившемся окне нажать кнопку **Variables** и выбрать группирующие переменные («Пол», «Возраст», «Образование») и зависимые переменные (САД и САД1) (рис. 70).

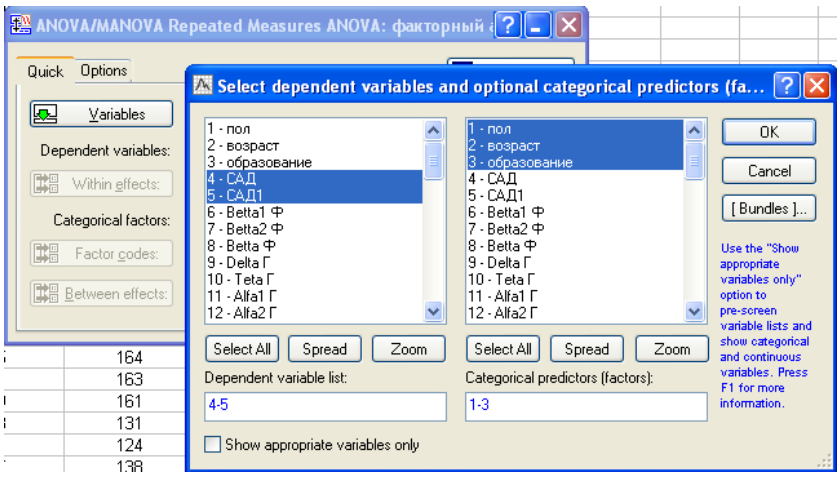


Рис. 70. Диалоговое окно для выбора группирующих и зависимых переменных

3. Нажать кнопу **Within effects** (внутригрупповые эффекты). В открывшемся окне в поле **Factor Name** задать имя фактора повторных измерений, например «Образование» (по умолчанию программа предложит выбрать один фактор для повторных измерений с именем R). В поле **№ of levels** (число уровней) можно задать число повторных измерений, в нашем случае их 2 (рис. 71).

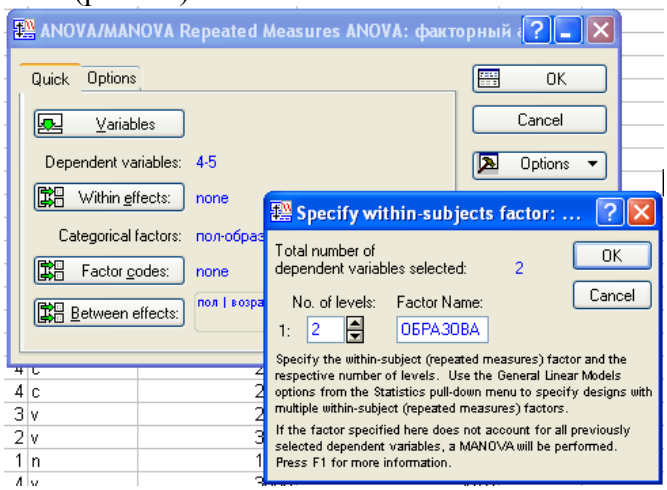


Рис. 71. Диалоговое окно для выбора фактора повторных измерений

4. Нажать кнопку **Factor code,** указать коды факторов, влияние которых необходимо оценить (рис. 72). Поскольку факторов не очень много, нажимаем кнопку **All** (выбрать все). Коды факторов можно и не задавать: если нажать кнопку **ОК**, программа задаст их автоматически.

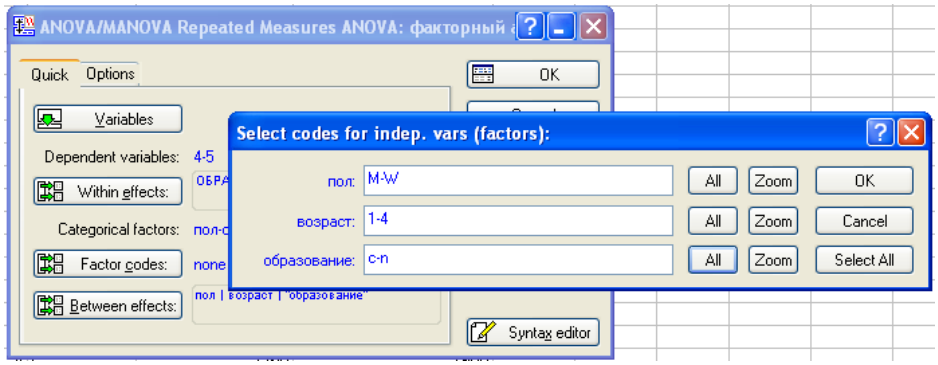


Рис. 72. Диалоговое окно для выбора кодов факторов

5. Нажать **OK**. В итоге появится уже знакомое окно с 8 закладками. Выбрав закладку **Summary** (Итоги), нажать кнопку **Test All effects** (проверить все эффекты). Из таблицы

видно, что гипотеза о неравенстве средних верна для фактора «Возраст» и сочетания факторов «Пол \ Возраст». Такие факторы как «Образование» и «Пол» в отдельности не оказывают влияния на величину САД (рис. 73).

Effect	Repeated Measures Analysis of Variance (факторны Sigma-restricted parameterization Effective hypothesis decomposition SS	Degr. of Freedom	MS	F	p
Intercept	2574179	1	2574179	9215.447	0.000000
{1}пол	11	1	11	0.040	0.841285
{2}возраст	2471	3	824	2.949	0.038139
{3}образование	552	2	276	0.989	0.376776
пол*возраст	3005	3	1002	3.586	0.017583
пол*образование	642	2	321	1.150	0.322268
возраст*образование	1065	6	177	0.635	0.701564
пол*возраст*образование	2953	6	492	1.762	0.118506
Error	20950	75	279		
{4}ОБРАЗОВА	144	1	144	38.502	0.000000
ОБРАЗОВА*пол	7	1	7	1.885	0.173884
ОБРАЗОВА*возраст	2	3	1	0.152	0.927896
ОБРАЗОВА*образование	4	2	2	0.506	0.604664
ОБРАЗОВА*пол*возраст	10	3	3	0.852	0.469963
ОБРАЗОВА*пол*образование	11	2	5	1.405	0.251860
ОБРАЗОВА*возраст*образование	22	6	4	0.985	0.441355
4*1*2*3	7	6	1	0.329	0.919601
Error	281	75	4		

Рис. 73. Результаты факторного анализа

Для визуализации полученных результатов, как в случае с факторным анализом, можно воспользоваться различными графическими эффектами, нажав на кнопку **All effects/Graphs (**Все эффекты/графики) (рис. 61).

Раздел 4. Корреляционный анализ.

В научных исследованиях часто возникает необходимость поиска взаимосвязи между различными признаками исследуемых групп (количеством осадков и урожайностью, ростом и весом человека, температурой тела и частотой пульса и т.д.). В приведенных примерах признаки связаны между собой, изменение одной переменной приводит к изменению другой.

Для решения задач данного типа используют различные разновидности корреляционного анализа. Корреляционный анализ позволяет оценить направление взаимосвязи между двумя признаками (прямое или обратное), а так же выразить ее количественно при помощи коэффициента корреляции. Чем ближе коэффициент к 1 (по модулю), тем сильнее связь между признаками. Знак коэффициента (+ или -) указывает на направление зависимости.

Коэффициент корреляции Пирсона.

Коэффициент корреляции Пирсона относится к группе параметрических методов статистического анализа и требует выполнения следующих обязательных условий:

1. Распределение данных должно подчиняться закону нормального распределения;
2. Взаимосвязь между признаками должна иметь линейный характер.

Проверить данные на «нормальность» распределения можно с использованием модуля **Distribution fitting** (Настройка распределения). Данный вид анализа подробно рассмотрен выше. Оценить линейность зависимости между признаками можно при помощи модуля **Scatterplots** (диаграммы рассеяния). Подробное описание этой процедуры представлено ниже в разделе Регрессионный анализ.

Предположим, необходимо выяснить, имеется ли связь между диаметром тела и ядра нервной клетки. Для этого были выполнены соответствующие измерения у 12 клеток. Полученные данные приведены на рисунке 74.

	1 диаметр нейрона	**2 диаметр ядра нейрона**
1	18	3
2	14	2
3	16	2,5
4	15	2,5
5	19	4
6	21	4
7	22	5
8	19	3
9	23	5
10	20	5

Рис. 74. Пример оформления данных при проведении корреляционного анализа

Для расчета коэффициента корреляции Пирсона необходимо:

1. Запустить из меню **Statistics / Basic Statistics/Tables / Correlation Matrices** (Корреляционные матрицы) (рис. 75).

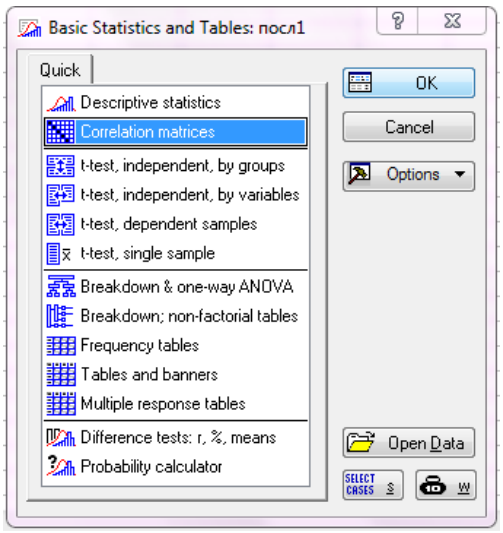


Рис. 75. Диалоговое окно для выбора корреляционной матрицы

2. Выбрать переменные, которые необходимо проанализировать. Для этого нажать кнопку **One variable list** (Один список переменных) или **Two lists (rect. matrix)** (Два списка (прямоугольная матрица)). В первом случае переменные выбираются из одного списка, а во втором – из двух (рис. 76).

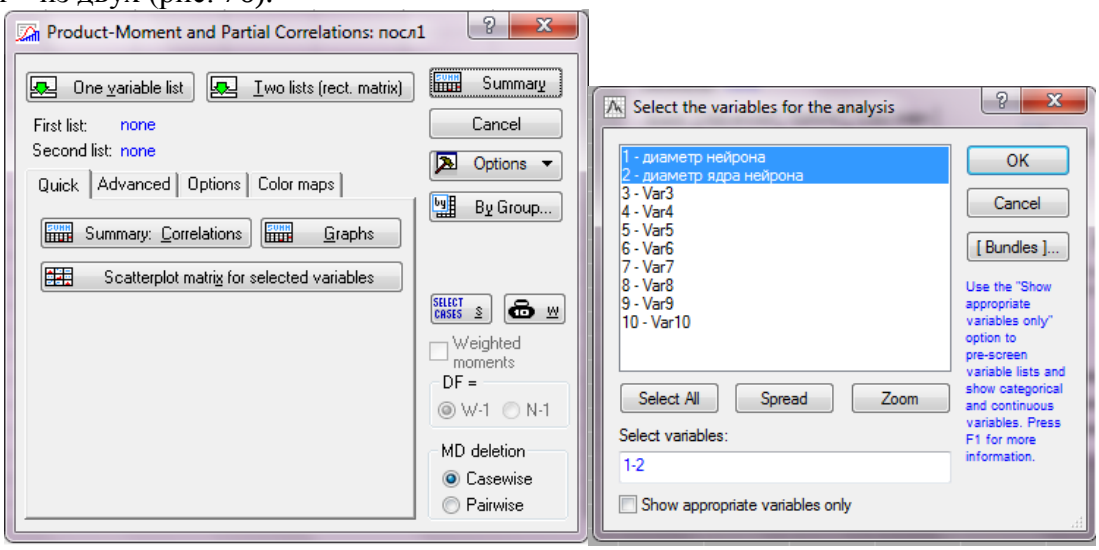


Рис. 76. Диалоговые окна для выбора переменных, связь между которыми необходимо проверить

3. Нажать на кнопку **Summary: Correlation matrix** (Результат: Корреляционная матрица). Появится таблица, в которой содержатся рассчитанные коэффициенты корреляции (рис. 77).

Correlations (Spreadsheet5)
Marked correlations are significant at p < ,05000
N=10 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	диаметр нейрона	диаметр ядра нейрона
диаметр нейрона	18,70000	2,983287	1,000000	0,916636
диаметр ядра нейрона	3,60000	1,149879	0,916636	1,000000

Рис. 77. Результат расчета коэффициента Пирсона

В нашем случае коэффициент корреляции является положительным и очень высоким (r = 0.92). Это указывает на прямую и очень высокую степень взаимосвязи между диаметром тела клетки и диаметром ее ядра. Помимо расчета коэффициента корреляции программа

оценивает и его статистическую значимость. Статистически значимые коэффициенты корреляции выделяются красным цветом ($P < 0.05$).

Коэффициент корреляции Спирмена.

Допустим, при расчете коэффициента корреляции Пирсона для диаметра тела нейрона и его ядра оказалось, что значения этих признаков не подчиняются закону нормального распределения. Применение коэффициента Пирсона в подобной ситуации приведет к выводам, не соответствующим действительности. В этом случае необходимо использовать один из непараметрических коэффициентов корреляции. К числу таковых относится ранговый коэффициент корреляции Спирмена.

Рассмотрим его применение:

1. Из меню **Statistics / Nonparametrics** (Непараметрические корреляции) запустить модуль **Correlations (Spearman, Kendall tau, gamma)** Корреляции (Спирмена, тау Кендалла, гамма) (рис. 78).

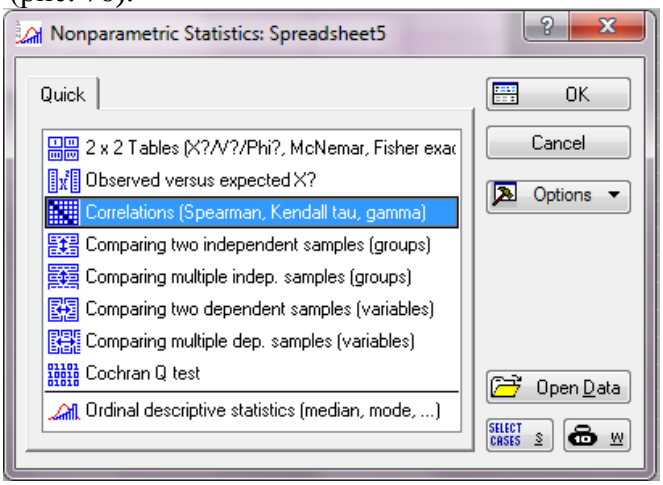


Рис. 78. Диалоговое окно для выбора корреляционной матрицы

2. Нажать на кнопку **Variables** и выбрать столбцы, содержащие необходимые данные (рис. 79).

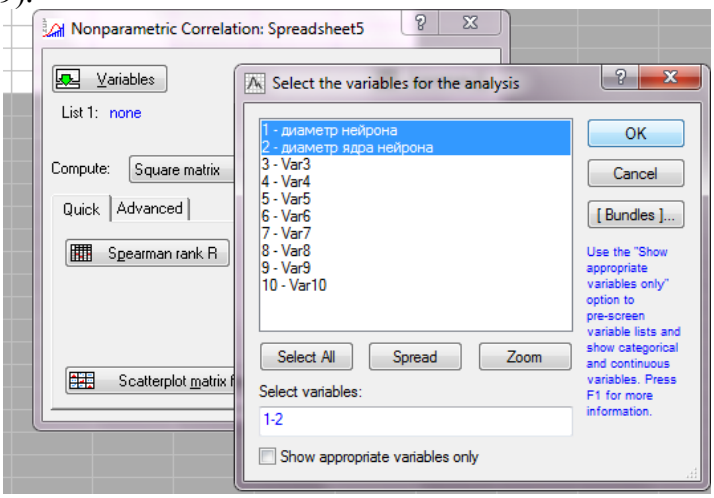


Рис. 79. Диалоговое окно для выбора переменных связь, между которыми необходимо проверить

3. Нажать кнопку **Spearman R** или **Spearman rank R**. Появится таблица с результатами анализа (рис. 80).

Variable	Spearman Rank Order Correlations (Spreadsheet5) MD pairwise deleted Marked correlations are significant at p <,05000 **диаметр нейрона**	диаметр ядра нейрона
диаметр нейрона	1,000000	0,938051
диаметр ядра нейрона	0,938051	1,000000

Рис. 80. Результат расчета коэффициента корреляции Спирмена

Как можно видеть, коэффициент Спирмена оказался даже выше рассчитанного ранее коэффициента Пирсона.

Коэффициент ассоциации (связанности).

Применение коэффициентов корреляции Пирсона и Спирмена возможно только в том случае, если изучаемые признаки имеют количественный характер. Однако в биологии признаки очень часто имеют качественную природу (пол, окраска, форма, состояние и т.д.). Классический корреляционный анализ в подобных случаях не возможен. Однако для таких признаков также можно рассчитать степень связанности. Это позволяет сделать коэффициент ассоциации или связанности φ (фи). Чем ближе данный коэффициент 1, тем сильнее взаимосвязь.

Рассмотрим применение данного анализа на иммунологическом примере. Группу из 111 крыс разбили на группы по 57 и 54 животных. Первую группу инфицировали патогенными бактериями, с последующим введением антител. Животных второй группы также инфицировали, но антитела не вводили (данная группа является контрольной).

После истечения инкубационного периода подсчитали погибших и выживших животных в обеих группах. Всего погибло 38 животных, 73 выжило. В первой группе погибло 13 животных, во второй – 25. Необходимо понять, имеется ли связь между введением антител и выживаемостью животных? Полученные данные необходимо представить в виде *таблицы сопряженности* размером 2x2 (четырехпольная таблица, или таблица с двумя входами) Таблица 1.

Таблица 1. Таблица сопряженности

	погибло	выжило
Бактерии + сыворотка	13	44
Бактерии	25	29

1. Из меню **Statistics / Nonparametrics** запустить модуль анализа четырехпольных таблиц **2X2 Tables** (рис. 81).

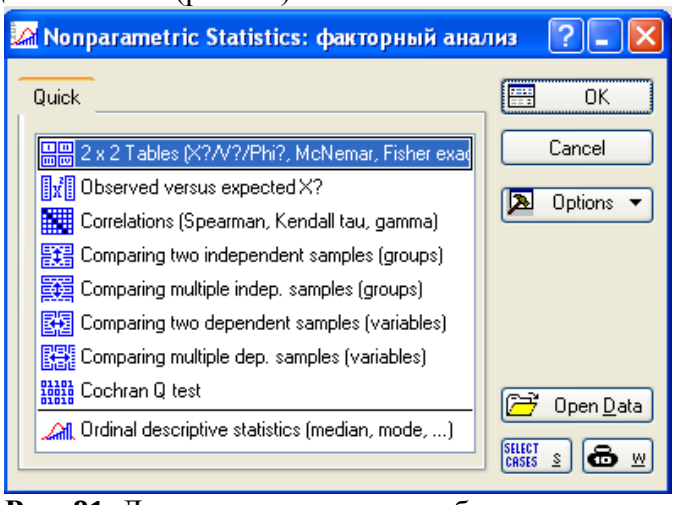


Рис. 81. Диалоговое окно для выбора расчета коэффициента ассоциации

2. В соответствии с приведенной выше таблицей ввести данные о численности животных в каждой из экспериментальных групп (рис. 82).

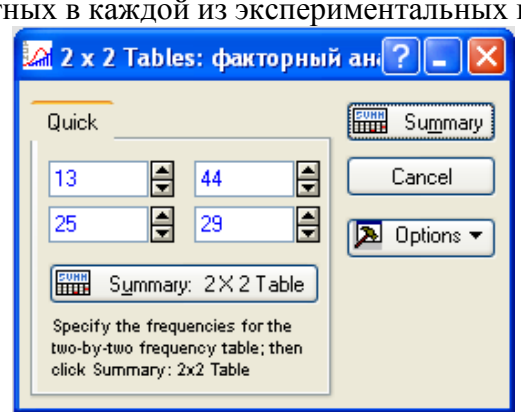


Рис. 82. Модуль анализа таблиц сопряженности

3. Нажать кнопку **Summary**. В результате этого появится таблица, содержащая набор статистических показателей (рис. 83).

2 x 2 Table (факторный анализ)

	Column 1	Column 2	Row Totals
Frequencies, row 1	13	44	57
Percent of total	11,712%	39,640%	51,351%
Frequencies, row 2	25	29	54
Percent of total	22,523%	26,126%	48,649%
Column totals	38	73	111
Percent of total	34,234%	65,766%	
Chi-square (df=1)	6,80	p= ,0091	
V-square (df=1)	6,73	p= ,0095	
Yates corrected Chi-square	5,79	p= ,0161	
Phi-square	,06122		
Fisher exact p, one-tailed		p= ,0078	
two-tailed		p= ,0102	
McNemar Chi-square (A/D)	5,36	p= ,0206	
Chi-square (B/C)	4,70	p= ,0302	

Рис. 83. Результаты анализа таблицы сопряженности

Нас интересует строка **Phi-square** (фи-квадрат). Для придания данному коэффициенту положительного значения он возведен в квадрат. В нашем случае фи-квадрат равен 0.061. После извлечения корня получаем φ = 0.247 (эту операцию необходимо проделать самостоятельно, в данном модуле она не предусмотрена). Таким образом, связь между введением антител и выживаемостью инфицированных животных является достаточно слабой.

Помимо коэффициента ассоциации, в данной таблице приведены значения: критерия *Chi-square* (критерий $\chi 2$). Данный тест проверяет нулевую гипотезу о случайном характере взаимосвязи между введением антител и выживаемостью животных. Поскольку вероятность ошибиться, отклонив данное предположение, менее 0.05 (P = 0.0091), можно сделать заключение о том, что несмотря на слабый эффект от введения антител, выживаемость животных в контрольной и экспериментальной группах все же статистически значимо различается.

Раздел 5. Регрессионный анализ.

Регрессионный анализ, наряду с корреляционным, является одним из наиболее распространенных методов обработки экспериментальных данных при изучении зависимостей. Суть этого анализа заключается в определении того, в какой степени изменение одной величины (зависимого признака) обусловлено влиянием одной или нескольких независимых величин (факторов).

Так как регрессионный анализ относится к группе параметрических методов статистического анализа, его применение требует выполнения ряда обязательных условий:

1. Линейный характер зависимости;
2. «Нормальное» распределение данных.

Выполним регрессионный анализ на примере показателей систолического артериального давления (САД) у людей разного возраста.

Для начала внесем данные, полученные в ходе исследования, в таблицу (рис. 84)

	1 возраст	2 давление мм.рт.ст.	3 Var3
1	30	108	
2	30	110	
3	40	125	
4	40	120	
5	40	118	
6	50	132	
7	50	137	
8	50	134	
9	60	148	
10	60	151	
11	60	146	
12	60	147	
13	70	162	
14	70	156	
15	70	164	
16	70	159	

Рис. 84. Пример оформления данных для выполнения регрессионного анализа

Выполнить регрессионный анализ можно в нескольких модулях программы STATISTICA. Воспользуемся модулем **Multiple Regression Analysis** (Анализ множественной регрессии).

1. Запустить соответствующий модуль из меню: **Statistics / Multiple Regression Analysis** (Анализ множественной регрессии) (рис. 85).

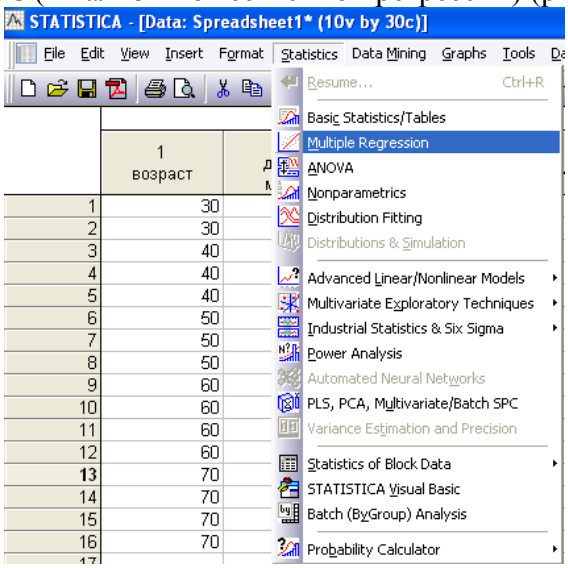


Рис. 85. Диалоговое окно для выбора регрессионного анализа

2. Нажать на кнопку **Variables** и указать зависимую (Dependent variable) и независимую переменные (Independent variable) (в нашем случае САД зависит от возраста) (рис. 86).

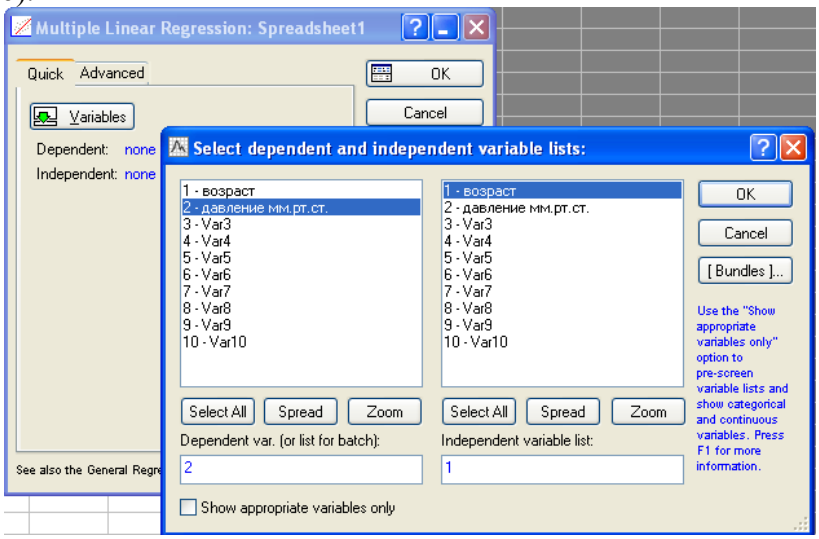


Рис. 86. Диалоговое окно для выбора переменных

3. Нажать кнопку **ОК**. В итоге появится окно с предварительными результатами анализа (рис. 87).

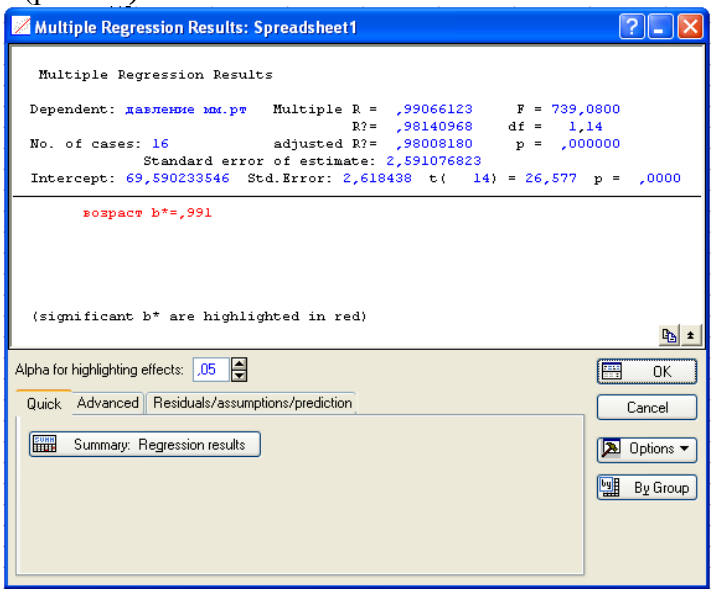


Рис. 87. Предварительные результаты регрессионного анализа

Наиболее важным показателем, приведенным в таблице, является коэффициент *R*2 или коэффициент детерминации. Он отражает «качество» рассчитанной регрессии. В нашем случае *R2* = 0.98, следовательно, изменения зависимой переменной (САД) на 98% объясняются изменением независимого фактора или переменной (возраст). Можно сказать, что построенная регрессионная модель отлично описывает связь между возрастом и артериальным давлением;

4. Нажать на кнопку **Summary: Regression results** (Результаты регрессионного анализа). Появится таблица со следующими результатами (рис. 88):

Regression Summary for Dependent Variable: давление мм.рт.ст. (Spreads
R= ,99066123 R?= ,98140968 Adjusted R?= ,98008180
F(1,14)=739,08 p<,00000 Std.Error of estimate: 2,5911

N=16	b*	Std.Err. of b*	b	Std.Err. of b	t(14)	p-value
Intercept			69,59023	2,618438	26,57700	0,000000
возраст	0,990661	0,036440	1,29830	0,047756	27,18603	0,000000

Рис. 88. Результаты регрессионного анализа

Из таблицы видно, что оба коэффициента регрессии достоверно отличаются от 0 (P << 0.001).

Однако в биологии наибольший интерес представляет не сама регрессия, а влияние, которое оказывает одна переменная на другую. В нашем случае важно знать, как будет меняться величина САД при изменении возраста.

Для этого необходимо:

1. В окне **Multiple Regression Result** (Результат множественной регрессии) выбрать вкладку **Residuals / assumptions / prediction** (Остаточные / Допущения / Предсказания) и нажатьм кнопку **Predict dependent variable** (Предсказать зависимую переменную) (рис. 89).

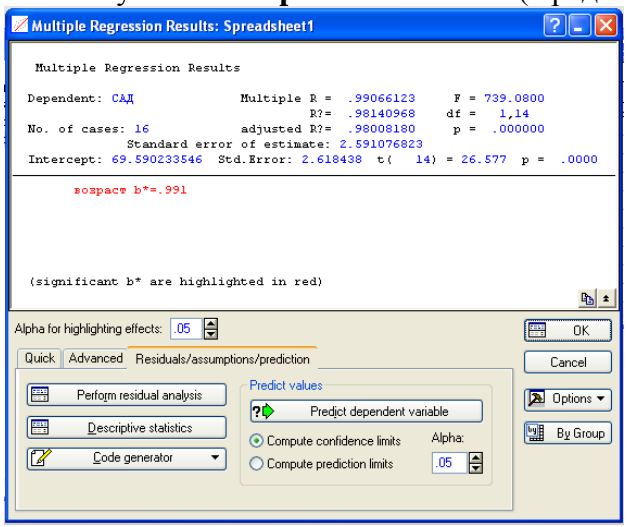


Рис. 89. Предварительные результаты регрессионного анализа

2. В появившемся окне, ввести данные для прогноза. Допустим, нам необходимо знать, какие показатели САД можно ожидать у лиц в возрасте 80 лет (рис. 90).

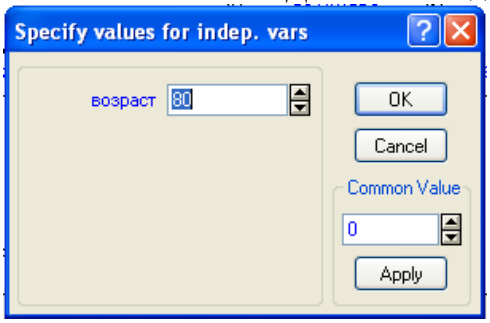


Рис. 90. Диалоговое окно для ввода данных для прогноза

3. Нажать кнопку **ОК**. В итоге появится окно, в котором представлены результаты предсказания (рис 91.).

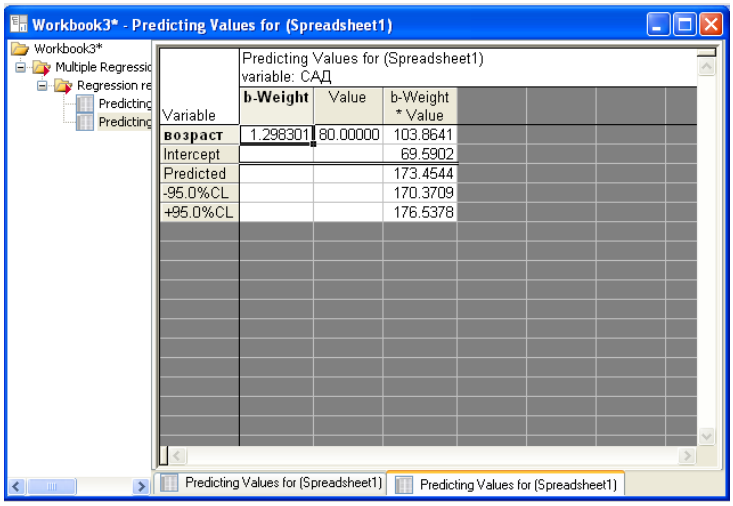


Рис. 91. Результаты регрессионного анализа

В таблице нас интересует строка **Predicted** (Предсказание). Как можно видеть, к 80 годам величина САД достигнет 173 мм. Рт. Ст.

Значимой частью регрессионного анализа является *анализ остатков* (разность между наблюдаемыми значениями зависимой переменной и теми ее значениями, которые предсказывает регрессионная модель).

Для выполнения данного анализа необходимо:

1. Выбрать закладку **Residuals / Assumptions / Predictions** (Остатки / Условия / Предсказания). Нажать кнопку **Perform residual analysis** (Выполнить анализ остатков) (рис. 92).

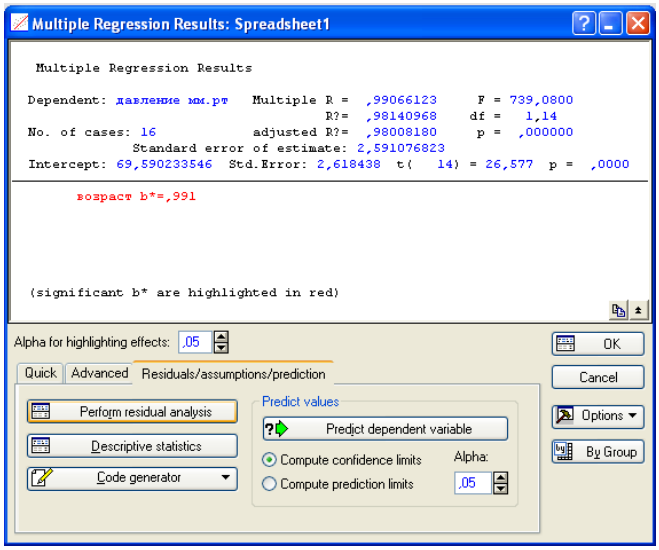


Рис. 92. Диалоговое окно анализа остатков

2. Во-первых: проверить «нормальность» распределения остатков. На закладке **Quick** (Быстро) нажать кнопку **Normal plot of residuals** (нормальный график остатков) (рис. 93).

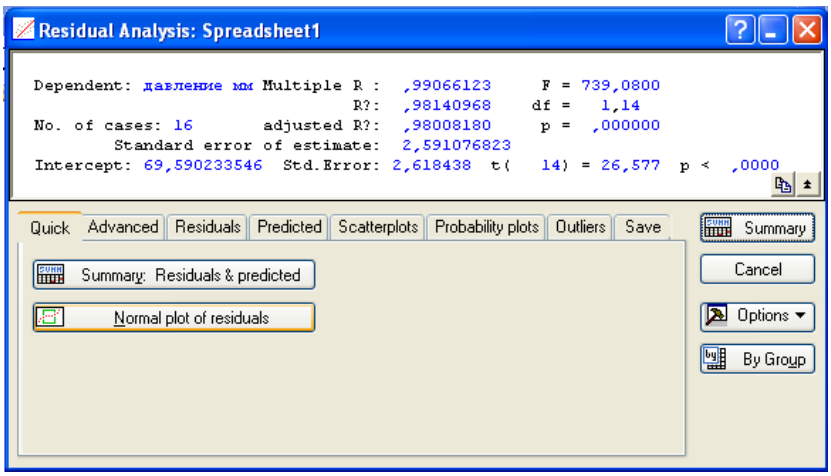


Рис. 93. Диалоговое окно для оценки остатков на соответствие закону нормального распределения

3. В результате будет построен график нормальных вероятностей (рис. 94). В том случае, если точки на этом графике компактно располагаются вдоль теоретически ожидаемой прямой, остатки распределены «нормально», применение линейной регрессионной модели корректно. Если данное условие не выполняется, применение линейной регрессии невозможно. Выходом в таком случае может стать трансформация данных (методы трансформации описаны ниже).

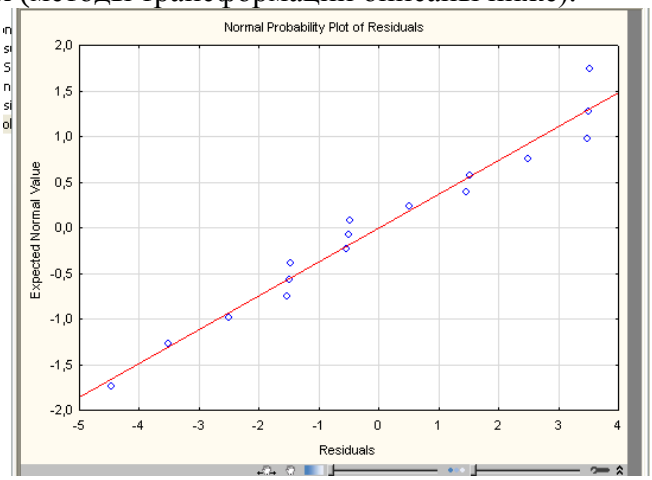


Рис. 94. График нормальных вероятностей остатков

4. Во-вторых: проверить дисперсию остатков. Дисперсия должна оставаться неизменной во всем диапазоне значений анализируемых переменных. На закладке **Scatterplots** (Диаграммы рассеяния) нажать кнопку **Predicted vs. Residuals** (прогнозируемые остатки) (рис. 95).

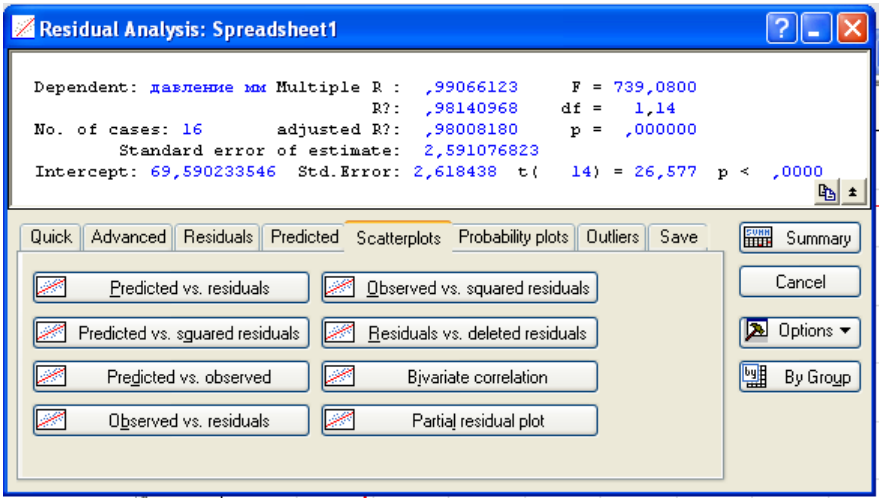


Рис. 95. Диалоговое окно для оценки дисперсии остатков

В результате будет построен график зависимости значений остатков от предсказываемых моделью значений зависимой переменной (рис. 96).

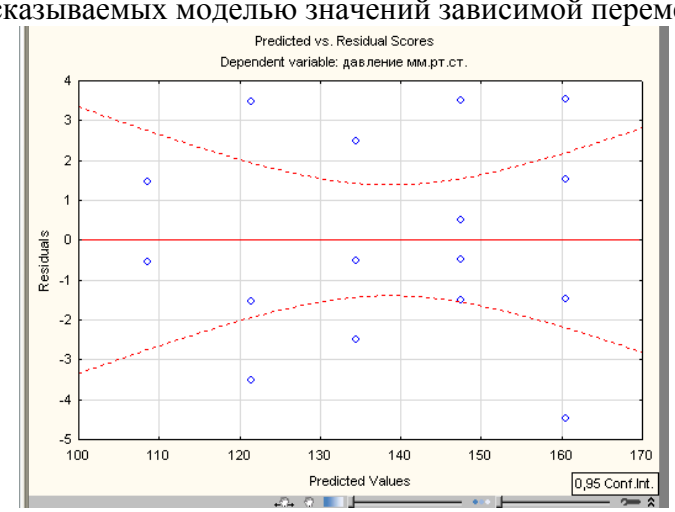


Рис. 96. График зависимости значений остатков от предсказываемых моделью значений зависимой переменной

Если дисперсия неизменна (условие выполняется), то точки на этом графике будут располагаться хаотично. Если же в расположении точек наблюдается какая-либо закономерность (точки группируются слева или справа, укладываются вдоль прямой и т.п.), применение линейного регрессионного анализа невозможно. Выходом также может стать трансформация данных.

В нашем случае оба условия выполняются, что подтверждает корректность рассчитанной регрессионной модели.

Трансформация нелинейно-связанных признаков

Очень серьезным ограничением для применения регрессионного анализа в биологии является нелинейный характер взаимосвязей между многими биологическими признаками. Например, зависимость между размерами тела и интенсивностью обменных процессов имеет нелинейный характер (степенной или экспоненциальный). В такой ситуации может помочь определенная трансформация исходных данных. Это позволяет перевести их в другую шкалу измерения и тем самым «выровнять» нелинейность.

На рис. 97 представлены данные об интенсивности процессов дыхания и размерах тела у рачков *Daphnia magna*.

	1 длинна тела мм.	2 интенсивн. дыхания
1	0,556	0,018
2	0,6	0,021
3	0,694	0,035
4	0,779	0,13
5	0,849	0,055
6	0,954	0,28
7	1,099	0,48
8	1,102	0,54
9	1,204	0,72
10	1,205	0,731
11		

Рис. 97. Пример оформления данных для выполнения регрессионного анализа

Характер связи между двумя переменными необходимо проверить еще до проведения регрессионного анализа. Для этого необходимо:

1. На закладке **Graphs** (графики) выбрать раздел **Scatterplots** (диаграммы рассеяния). В результате будет построена диаграмма рассеяния (рис. 98, 99).

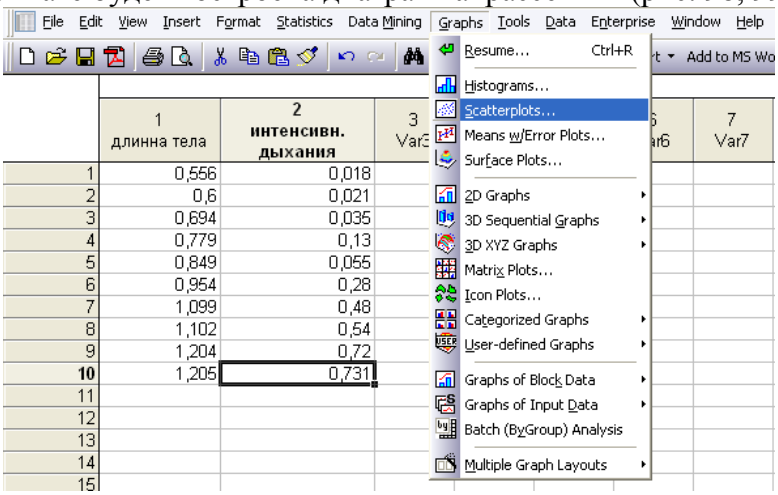


Рис. 98. Диалоговое окно для построения диаграммы рассеяния

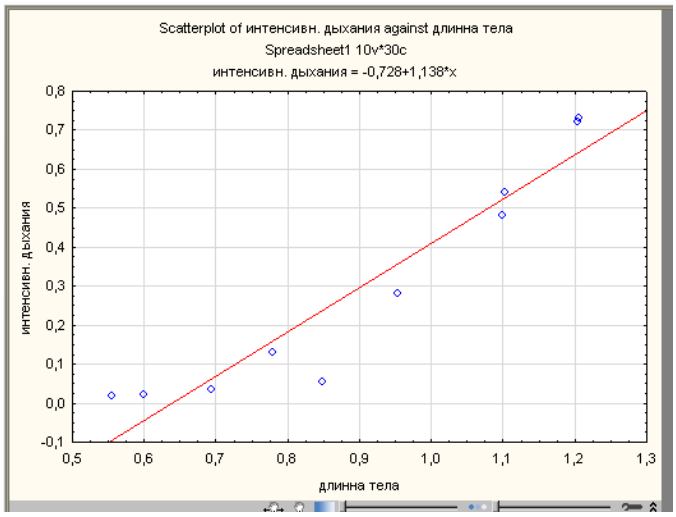


Рис. 99. Диаграмма рассеяния для данных о размере тела и интенсивности дыхания

Расположение точек на диаграмме показывает, что связь между размером тела дафний и интенсивностью обмена веществ не является линейной. Линейный регрессионный анализ неприменим. Однако решить данную проблему можно путем логарифмирования значений одного или (чаще) обоих анализируемых признаков.

Такую трансформацию можно выполнить, присвоив переменным так называемые *длинные имена* (*Long name*) в виде формул.

1. Прологарифмируем столбец 1, содержащий значения размера тела дафний. Для этого кликнуть два раза по заголовку любого свободного столбца (например, столбца Var 3). Это приведет к появлению окна настроек свойств переменной (рис.100).

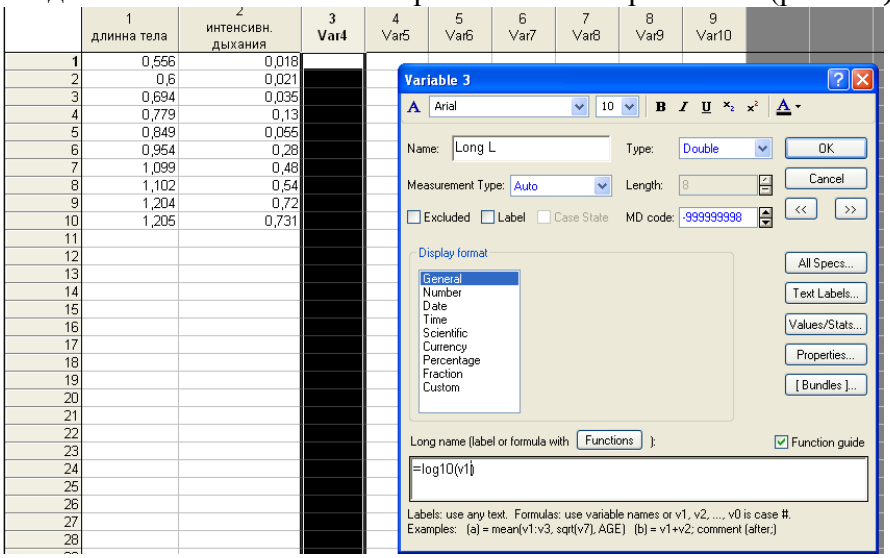


Рис. 100. Диалоговое окно для присвоения переменным длинных имен

2. В поле *Long name* ввести формулу =*log*10(*v*1), где *v*1 – это столбец с данными о длине рачков. В поле *Name* ввести короткое имя переменной, например, «*Log L*».

3. Нажимаем кнопку **ОК. П**оявится панель **«Expression OK. Recalculate the variable now?»** (Формула введена правильно. Пересчитать значения переменной?). Нажать **Yes**. Программа трансформирует значения первого столбца и они появятся в столбце 3 (рис. 101).

	1 длинна тела	2 интенсивн. дыхания	3 Long L
1	0,556	0,018	-0,25493
2	0,6	0,021	-0,22185
3	0,694	0,035	-0,15864
4	0,779	0,13	-0,10846
5	0,849	0,055	-0,07109
6	0,954	0,28	-0,02045
7	1,099	0,48	0,040998
8	1,102	0,54	0,042182
9	1,204	0,72	0,080626
10	1,205	0,731	0,080987
11			

Рис. 101. Результаты логарифмического преобразования экспериментальных данных

4. Аналогичную операцию необходимо выполнить для данных по интенсивности обменных процессов. Обратите внимание, в качестве его длинного имени необходимо ввести формулу =*log*10(*v*2) (рис. 102).

	1 длинна тела	2 интенсивн. дыхания	3 Long L	Intensity of breathin
1	0,556	0,018	-0,25493	-1,74473
2	0,6	0,021	-0,22185	-1,67778
3	0,694	0,035	-0,15864	-1,45593
4	0,779	0,13	-0,10846	-0,88606
5	0,849	0,055	-0,07109	-1,25964
6	0,954	0,28	-0,02045	-0,55284
7	1,099	0,48	0,040998	-0,31876
8	1,102	0,54	0,042182	-0,26761
9	1,204	0,72	0,080626	-0,14267
10	1,205	0,731	0,080987	-0,13608

Рис. 102. Результаты логарифмического преобразования экспериментальных данных

5. Если построить диаграмму рассеяния для трансформированных данных, то можно увидеть, что точки укладываются вдоль прямой линии значительно компактнее (рис. 103), и возможно применение обычной линейной регрессии.

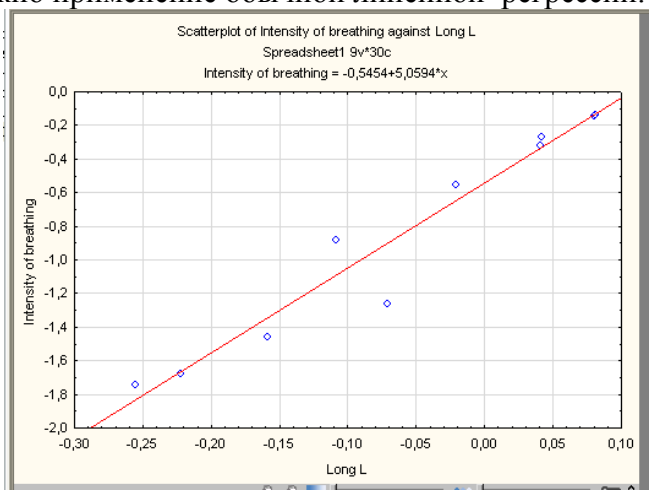


Рис. 103. Диаграмма рассеяния для данных о размере тела и интенсивности дыхания после процедуры логарифмирования

Важно отметить, что процедура трансформации данных применима не только для «выравнивания» нелинейных связей между признаками. Логарифмирование позволяет «приблизить» распределение данных к «нормальному», а также добиться однородности дисперсии в группах. Все это позволяет использовать более мощные параметрические методы анализа данных.

Раздел 6. Кластерный анализ.

Главная задача, решаемая кластерным анализом – разделение исходного множества исследуемых объектов и признаков на однородные, в некотором смысле, группы или кластеры.

Методы кластерного анализа позволяют решать следующие задачи:

1. Классификация объектов с учетом различных характеристик или признаков, отражающих их природу;

2. Проверка предположений или поиск некоторой внутренней структуры в совокупности изучаемых объектов.

Достоинством кластерного анализа является то, что он относится к группе непаметрических методов статистического анализа. Его применение возможно в малых группах или в том случае, когда не выполняются требования «нормальности» распределения данных.

В пакете STATISTICA реализуются следующие разновидности кластерного анализа:

1. Иерархические алгоритмы или древовидная кластеризация;
2. Метод К- средних;
3. Двухвходовое объединение.

Иерархические алгоритмы или древовидная кластеризация.

Назначение этого вида кластерного анализа заключается в объединении объектов в достаточно большие группы (кластеры) на основании сходства или «расстояния» между объектами. Результатом такой кластеризации является построение иерархического дерева. Алгоритм работы данного вида кластерного анализа заключается в последовательном объединении в группы сначала самых близких, а затем все более отдаленных друг от друга объектов.

Рассмотрим данный вид анализа на примере экологического исследования различных растительных сообществ. В пределах разных по видовому составу растительных сообществ проведено измерение девяти различных фитоценотических параметров (рис. 104). Необходимо определить, какие сообщества являются наиболее сходными по структурным характеристикам, а какие обладают наибольшими отличиями. Данные, полученные в ходе исследования, внесены в таблицу (рис. 104).

	2 диаметр\древостой	3 сомкнутость крон\древостой	4 высота\подрост	5 диаметр\подрост	6 высота\подлесок	7 диаметр\подлесок	8 высота\травостой	9 покрытие\травостой
дубрава остепненная	63.9	11.9	1.9	2.6	0.7	0.6	0.32	25.8
дубрава мятликовая	60.8	29.8	1.3	1.1	1.1	0.8	0.37	30
дубрава разнотравная	64.3	23.6	2.95	2.2	1.9	1.6	0.47	45.9
дубрава ландышева	62.3	25.5	2.1	1.9	1.5	1.2	0.62	75.3
липо-дубрава ландышевая	56.1	58.4	2.1	2.6	0.9	0.7	0.6	63.4
липо-дубрава крапивная	62.4	36.4	2.2	2.4	0.9	1	0.3	32.7
липняк снытевый	54.9	62.4	3.1	2.59	1.2	1.5	0.58	66.4
липо-кленовник	48.9	70	3.5	2.9	1.9	1.1	0.25	32.5
березняк	50.1	62.5	3	1.9	1.6	1.2	0.24	70.8
осинник	49.3	30.2	1.5	1.2	1.1	0.3	0.36	60.2
сосняк	49.3	30	0.5	0.3	1.1	0.3	0.25	5.01

Рис. 104. Пример оформления данных для выполнения древовидной кластеризации

1. Запустить модуль кластерного анализа **Statistics / Multivariate Exploratory / Cluster Analysis** (Статистика / Многомерные исследовательские методы / Анализ кластеров) (рис. 105).

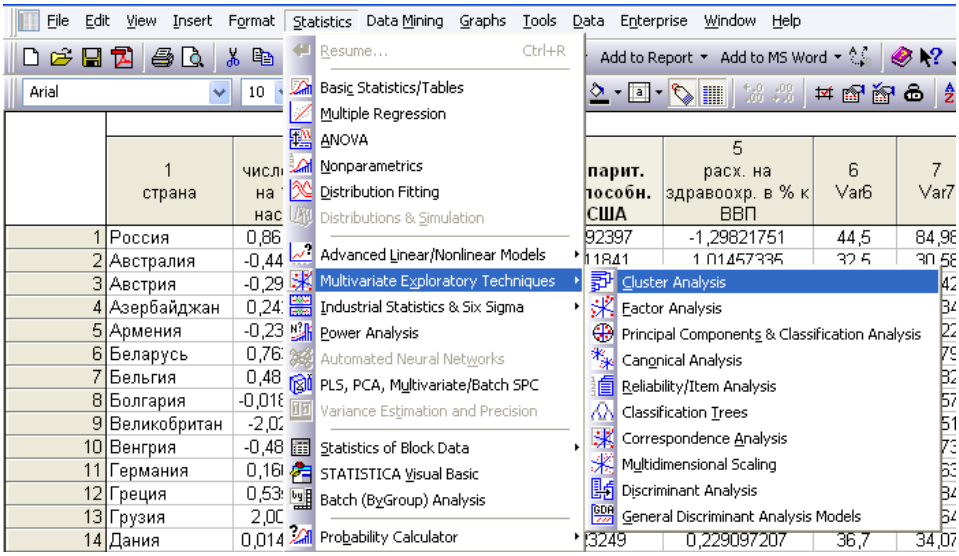


Рис. 105. Диалоговое окно для выбора метода проведения кластерного анализа

2. В списке методов выбрать **Joining (tree clustering)** (Древовидная кластеризация) и нажать кнопку **ОК** (рис. 106).

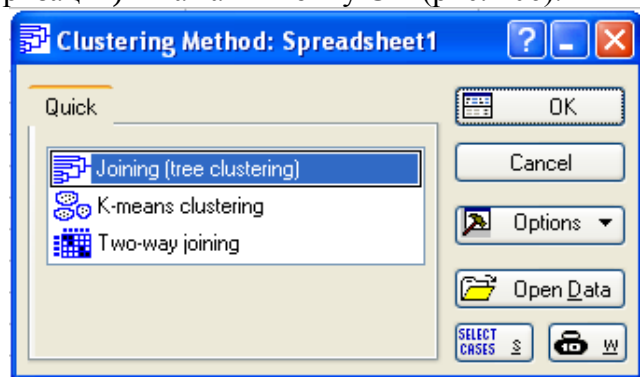


Рис. 106. Диалоговое окно для выбора метода кластерного анализа

3. Нажать кнопку **Variables** и выбрать переменные для анализа (для кластеризации). Нажать **ОК** (рис. 107).

4. Поскольку показатели, характеризующие растительные сообщества, расположены в строках и не подвергались трансформации, в строке **Cluster** (Кластер) выбрать пункт меню **Cases (rows)** (Случаи), а в строке **Input file** (Исходные данные) выбрать **Raw data** (Необработанные данные) (рис. 107).

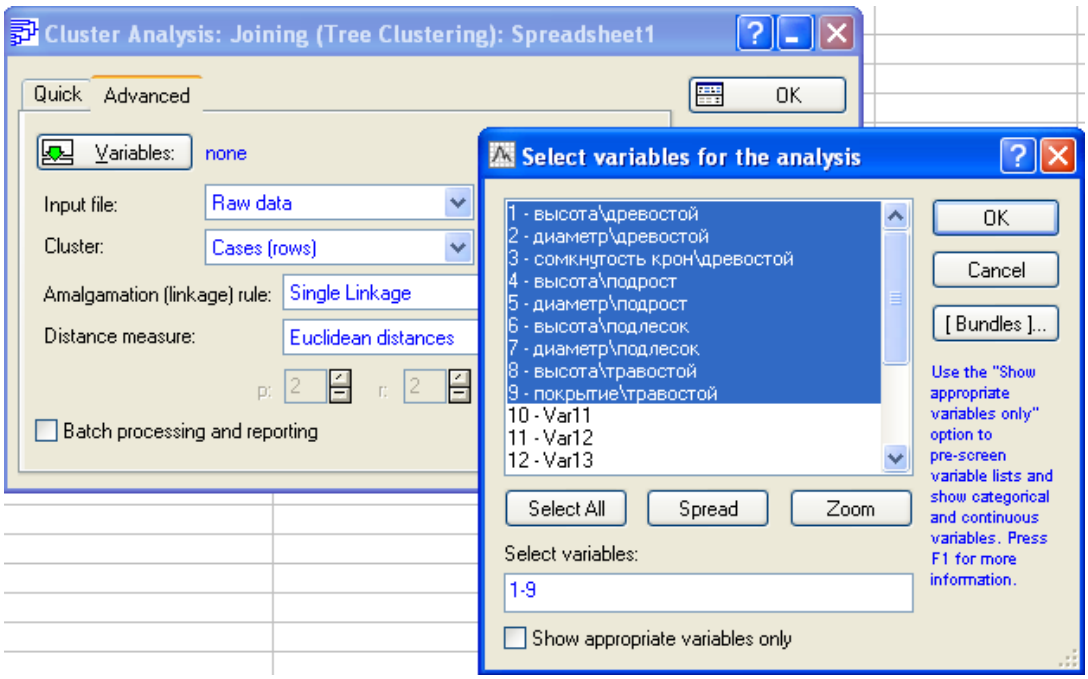


Рис. 107. Диалоговое окно для выбора настроек кластерного анализа

5. После нажатия **ОК,** появляется диалоговое окно, в котором необходимо выбрать тип древовидной диаграммы (рис. 108). Выбрать удобный для представления результатов тип диаграммы, получаем результат (рис. 109).

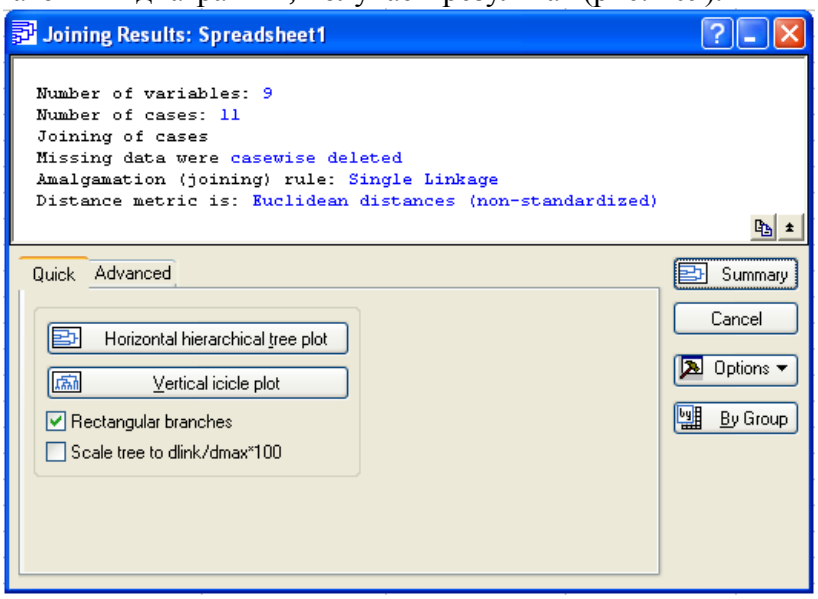


Рис. 108. Диалоговое окно для выбора настроек кластерного анализа

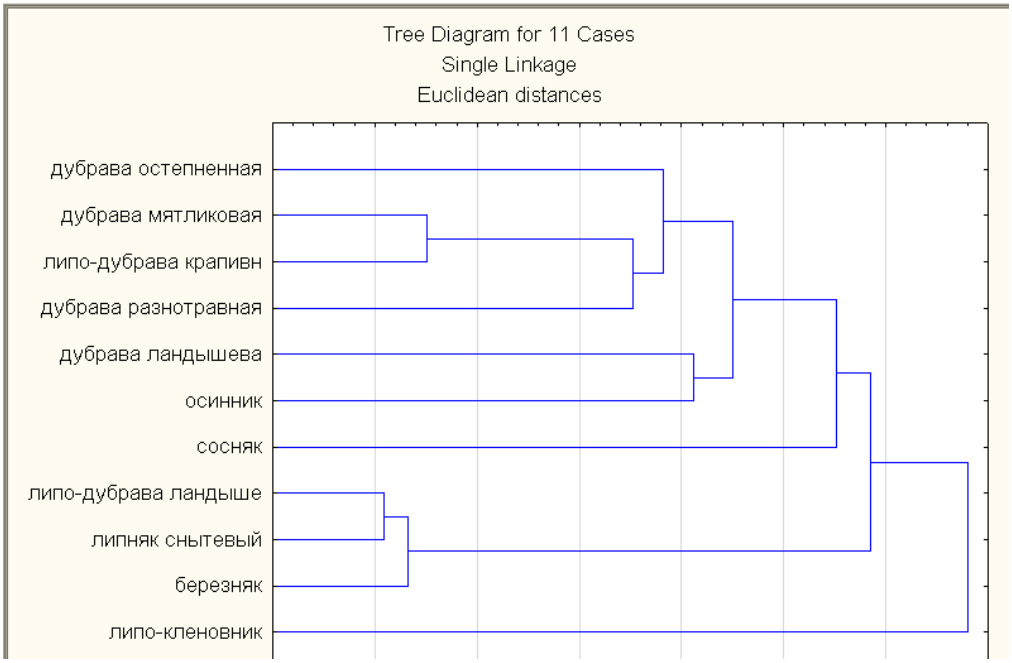


Рис. 109. Отображение результатов кластерного анализа

На полученной кластерной диаграмме отражено объединение в группы растительных сообществ по сходству фитоценотических параметров. На диаграмме хорошо видны 2 кластера. Первый образуют дубравы остепненные, мятликовые, липо-крапивные. Во второй кластер вошли липо-дубрава ландышевая, липняк снытевый, липо-кленовник и березняк. Осинник и дубрава ландышевая схожи с сообществами первой и второй группы. Наиболее отличается от всех остальных сообществ сосняк.

Метод К-средних.

Является одним из наиболее часто используемых методов кластерного анализа. Данный метод позволяет разбить множество объектов на заданное число кластеров **К.** Данные кластеры расположены на максимальных расстояниях друг от друга. После получения результатов кластеризации, рассчитывают средние для каждого кластера по каждому измерению, и оценивают, насколько кластеры отличаются друг от друга. В идеале, для большинства измерений, должны получиться сильно различающиеся средние. Значения F-статистики, полученные для каждого измерения, являются еще одним индикатором качества дискриминации кластеров.

Очень важно помнить, что применяя данный метод анализа, исследователь должен иметь гипотезу (предположение) относительно числа возможных кластеров. Данный метод строит ровно столько кластеров, сколько указано исследователем. Сформированные кластеры будут расположены на максимально больших расстояниях друг от друга.

Для примера рассмотрим результаты исследования стран по некоторым показателям уровня жизни населения. На основании этих данных, необходимо разделить страны на группы. Различия между группами должны быть максимальными, а внутри групп минимальными (страны должны быть максимально похожи). Результаты исследования и пример их оформления приведены на рисунке 110.

	1 число врачей на 10 тыс. населения	2 смертность на 100 тыс. населения	3 ВВП по парит. покуп. способн. в % к США	4 расх. на здравоохр. в % к ВВП
Россия	44.5	84.98	20.4	3.2
Австралия	32.5	30.58	71.4	8.5
Австрия	33.9	38.42	78.7	9.2
Азербайджан	38.8	60.34	12.1	3.3
Армения	34.4	60.22	10.9	3.2
Беларусь	43.6	60.79	*20.4*	5.4
Бельгия	41	29.82	79.7	8.3
Болгария	36.4	70.57	17.3	5.4
Великобритания	17.9	34.51	69.7	7.1
Венгрия	32.1	64.73	24.5	6
Германия	38.1	36.63	76.2	8.6
Греция	41.5	32.84	44.4	*5.7*
Грузия	*55*	62.64	11.3	3.5
Дания	36.7	34.07	79.2	6.7
Ирландия	15.8	39.27	57	6.7
Испания	40.9	28.46	54.8	7.3
Италия	49.4	30.27	72.1	8.5
Казахстан	38.1	69.04	13.4	3.3
Канада	27.6	25.42	79.9	10.2
Киргизия	33.2	53.13	11.2	3.4

Рис. 110. Пример оформления данных для проведения кластерного анализа

Поскольку переменные, используемые для анализа, имеют разные единицы измерения (или если резко не совпадают масштабы измерений), необходима предварительная **нормировка.** Нормировка это перевод (преобразование) исходных данных в безразмерные величины. Для этого необходимо:

1. Щелкнуть правой кнопкой мыши по имени переменной. В открывшемся окне выбрать последовательность команд: **Fill / Standardize Block / Standardize Columns** (Заполнение / стандартизация блоков / стандартизация столбцов) (рис. 111).

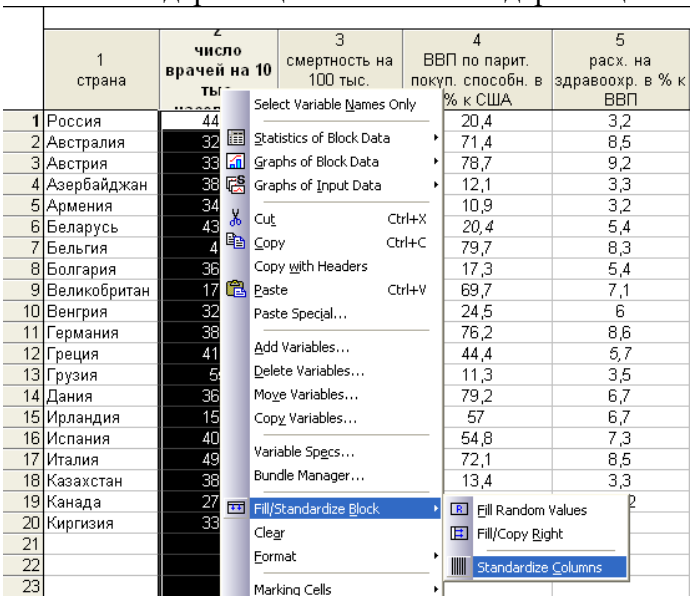


Рис. 111. Диалоговое окно для нормирования переменных

Значения нормированной переменной станут равными нулю, а дисперсии – единице. Подобную операцию необходимо проделать со всеми переменными. После этого можно приступить к выполнению кластерного анализа.

2. Запустить модуль **Statistics / Multivariate Exploratory / Cluster Analysis** (Статистика / Многомерные исследовательские методы / Анализ кластеров) (рис. 112).

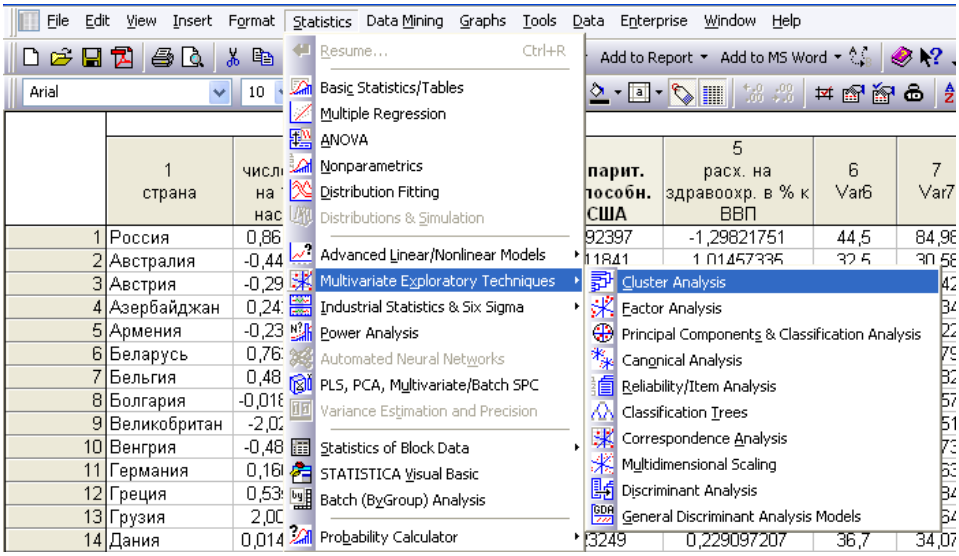


Рис. 112. Диалоговое окно для выбора метода проведения кластерного анализа

2. В списке методов выбрать **k-means clustering** (метод k-средних) и нажать кнопку **ОК** (рис. 113).

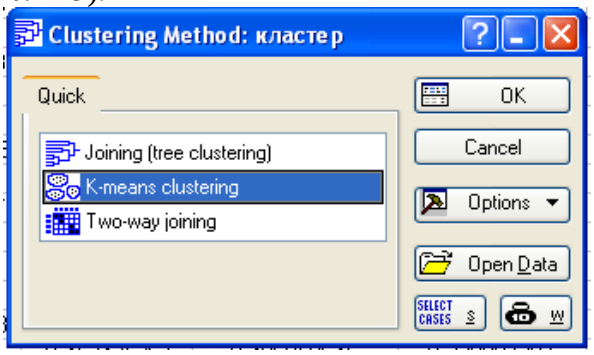


Рис. 113. Диалоговое окно для выбора метода кластерного анализа

3. Нажать кнопку **Variables** и выбрать переменные для анализа (для кластеризации). Нажать **ОК** (рис. 114).

4. Поскольку показатели уровня жизни расположены в строках, в поле **Cluster** (Кластер), выбираем пункт **Cases (rows)** (Случаи). В поле **Number of clusters** (Число кластеров) задаем количество кластеров, на которые необходимо разбить выборку, например, 3. В строке **Number of (iterations)** (Число итераций) задаем максимальное число итераций, используемых при построении классов, например, 10 (рис. 114).

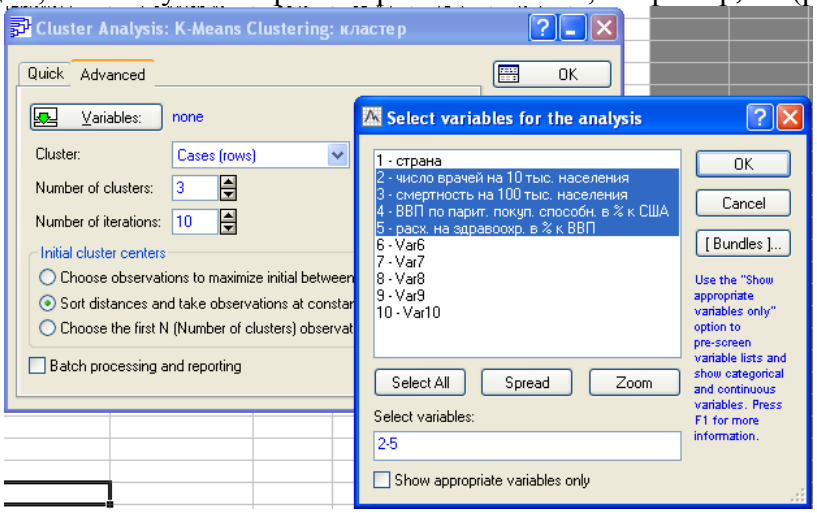


Рис. 114. Диалоговое окно для выбора настроек кластерного анализа

После нажатия **ОК**, появится диалоговое окно с результатами кластеризации (рис. 115). В окне результатов в верхней части приведена следующая информация:

Количество переменных (Number of variables) – 4;

Число регистров (Number of cases) – 20;

K-means clustering of cases – Метод кластеризации k-means clustering;

Количество групп (Number of cluster) – 3;

Solution was obtained after 2 iterations – Решение найдено после 2 итераций.

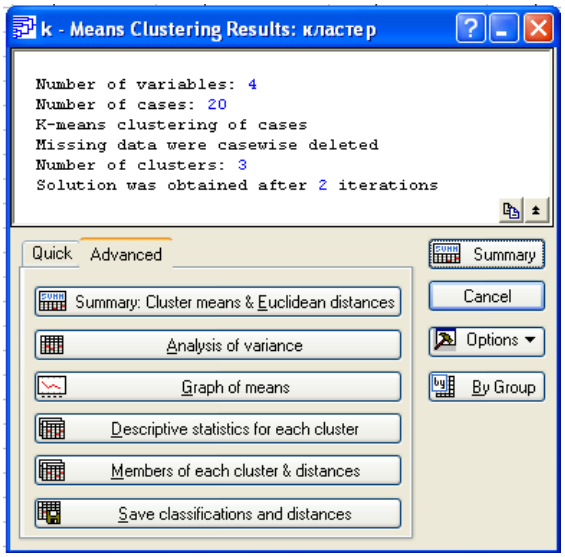
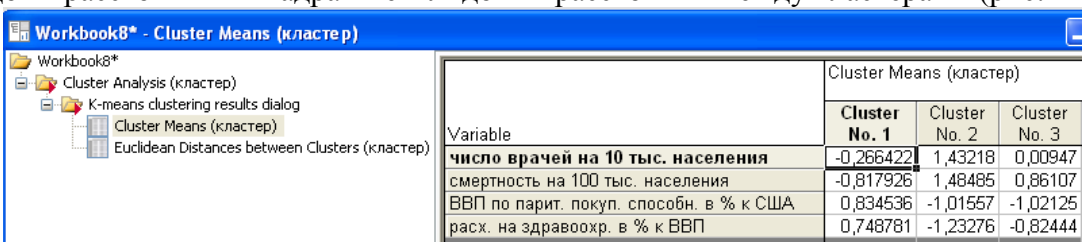


Рис. 115. Диалоговое окно с результатами кластеризации

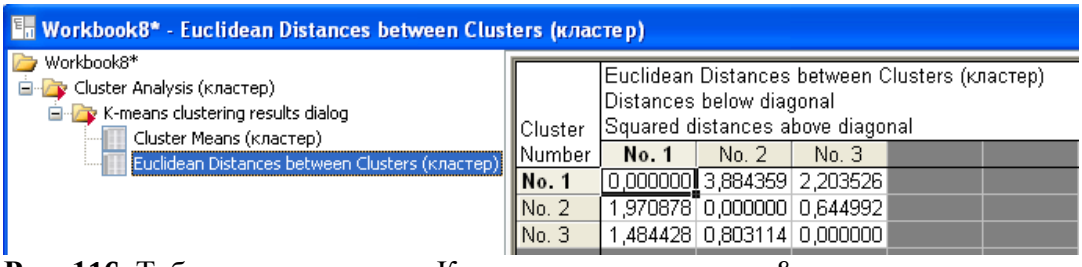
6. Выбираем закладку **Advanced** (Расширенный). При помощи кнопок данного окна можно просмотреть результаты анализа.

Функциональная кнопка Cluster Means&Euclidean Distances (Кластерные усреднения & евклидовы расстояния) выводит 2 таблицы. В первой указаны средние значения для каждого кластера (усреднение производится внутри кластера). Во второй евклидовы расстояния и квадраты евклидовых расстояний между кластерами (рис. 116).



Cluster Means (кластер)

Variable	Cluster No. 1	Cluster No. 2	Cluster No. 3
число врачей на 10 тыс. населения	-0,266422	1,43218	0,00947
смертность на 100 тыс. населения	-0,817926	1,48485	0,86107
ВВП по парит. покуп. способн. в % к США	0,834536	-1,01557	-1,02125
расх. на здравоохр. в % к ВВП	0,748781	-1,23276	-0,82444

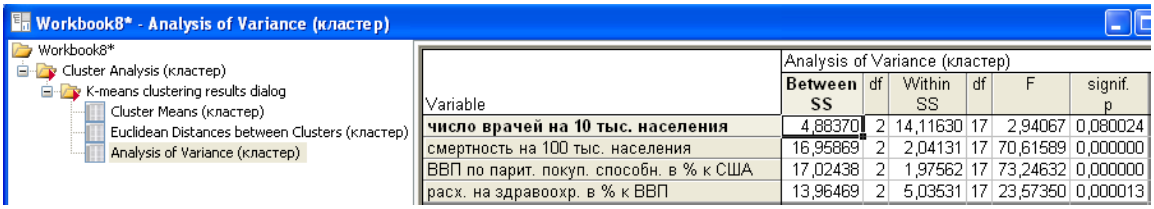


Euclidean Distances between Clusters (кластер)
Distances below diagonal
Squared distances above diagonal

Cluster Number	No. 1	No. 2	No. 3
No. 1	0,000000	3,884359	2,203526
No. 2	1,970878	0,000000	0,644992
No. 3	1,484428	0,803114	0,000000

Рис. 116. Таблицы результатов Кластерные усреднения & евклидовы расстояния

Функциональная кнопка Analysis of variance (Анализ дисперсии) позволяет просмотреть таблицу дисперсионного анализа, где например, выводятся суммы квадратов отклонения объектов от центров кластеров (SS Within) и суммы квадратов отклонений между центрами кластеров (SS Between), значения F-статистики, уровни значимости p (рис. 117).

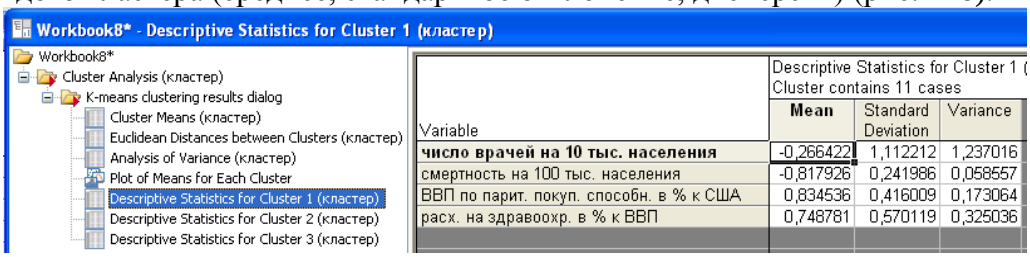


Analysis of Variance (кластер)

Variable	Between SS	df	Within SS	df	F	signif. p
число врачей на 10 тыс. населения	4,88370	2	14,11630	17	2,94067	0,080024
смертность на 100 тыс. населения	16,95869	2	2,04131	17	70,61589	0,000000
ВВП по парит. покуп. способн. в % к США	17,02438	2	1,97562	17	73,24632	0,000000
расх. на здравоохр. в % к ВВП	13,96469	2	5,03531	17	23,57350	0,000013

Рис. 117. Таблица результатов дисперсионного анализа

Функциональная кнопка Descriptive Statistics for each clusters (Описательная статистика для каждого кластера). Выводит таблицы с показателями описательной статистик для каждого кластера (среднее, стандартное отклонение, дисперсия) (рис. 118).



Descriptive Statistics for Cluster 1 (
Cluster contains 11 cases

Variable	Mean	Standard Deviation	Variance
число врачей на 10 тыс. населения	-0,266422	1,112212	1,237016
смертность на 100 тыс. населения	-0,817926	0,241986	0,058557
ВВП по парит. покуп. способн. в % к США	0,834536	0,416009	0,173064
расх. на здравоохр. в % к ВВП	0,748781	0,570119	0,325036

Рис. 118. Таблица с описательной статистикой для каждого кластера

Функциональная кнопка Graph of means (График усреднений). Отображает средние значения для каждого кластера на линейном графике. Кривые на этом графике соответствует выделенным кластерам. По горизонтальной оси отложены переменные, включенные в анализ. По вертикальной оси средние значения для стран, входящих в каждый из кластеров (рис. 119).

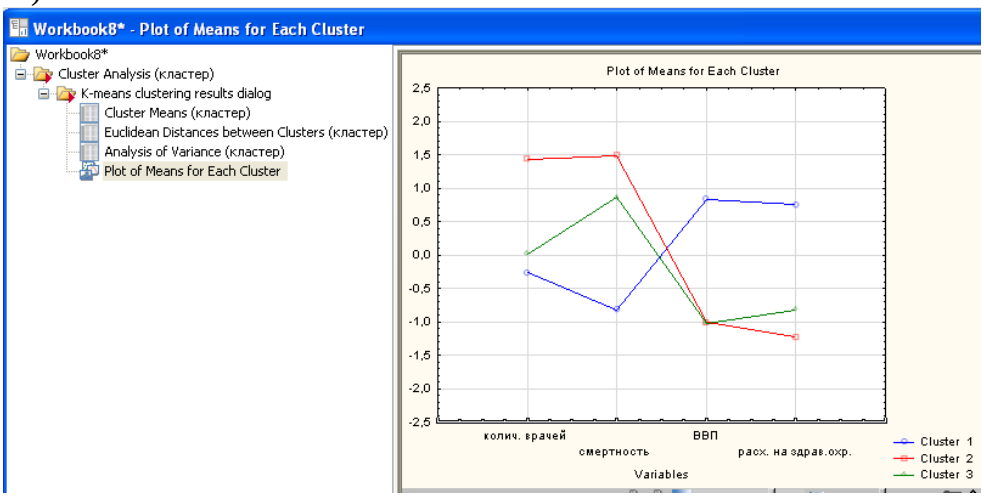


Рис. 119. График усреднений для каждого кластера

Функциональная кнопка Member of each cluster & distances (Элементы каждого кластера & расстояния). Показывает распределение стран по соответствующим кластерам. В выведенных таблицах представлены номера стран, отнесенных к тому или иному кластеру. В заголовке таблицы можно видеть, что в первый кластер попало 11 стран, их номера перечислены в столбце **Case №** (2, 3, 7, 9, 11, 12 и т.д.). В строках указаны расстояния от каждой страны до центра кластера (рис. 120).

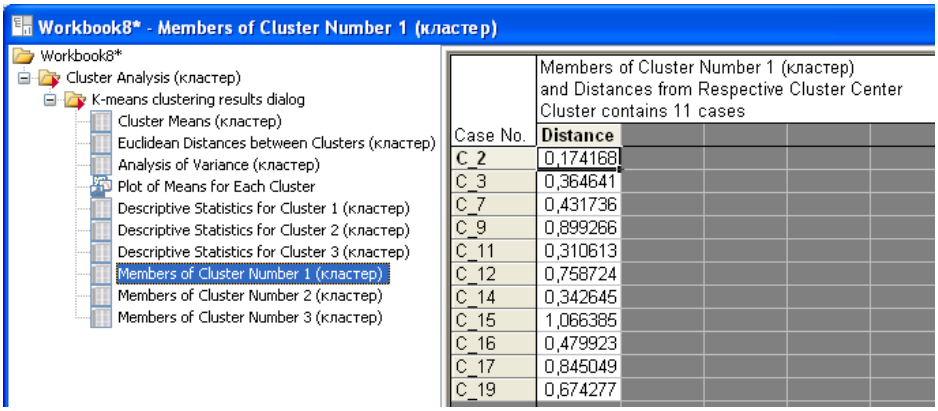


Рис. 120. Таблица распределения стран по кластерам

Функциональная кнопка Save classifications and distances (Сохранить классификации и расстояния) выводит номера объектов, входящих в каждый кластер и расстояния объектов до центра каждого кластера. Информация о принадлежности объектов к кластерам может быть записана в файл и использована в дальнейшем анализе (рис. 121).

	кластер						
	1 число врачей на 10 тыс. населения	2 смертность на 100 тыс. населения	3 ВВП по парит. покуп. способн. в % к США	4 расх. на здравоохр. в % к ВВП	5 CASE_NO	6 CLUSTER	7 DISTANCE
C_1	0,861698827	2,1113557	-0,858292397	-1,29821751	1	2	0,43
C_2	-0,442259045	-0,93984172	0,904611841	1,01457335	2	1	0,17
C_3	-0,290130626	-0,500110327	1,15694911	1,32003629	3	1	0,36
C_4	0,242318838	0,729342751	-1,14519642	-1,25457994	4	3	0,26
C_5	-0,235799048	0,722612168	-1,18667652	-1,29821751	5	3	0,29
C_6	0,763901986	0,754582435	-0,858292397	-0,338191115	6	3	0,46
C_7	0,481377781	-0,982468743	1,19151586	0,927298219	7	1	0,43
C_8	-0,0184727365	1,30312491	-0,965449321	-0,338191115	8	3	0,33
C_9	-2,02874112	-0,719415142	0,845848367	0,40364746	9	1	0,90
C_10	-0,485724307	0,975569896	-0,716568723	-0,0763657357	10	3	0,48
C_11	0,166254629	-0,600508183	1,07053224	1,05821091	11	1	0,31
C_12	0,535709359	-0,813082415	-0,0286904023	-0,207278425	12	1	0,76
C_13	2,00266196	0,858345583	-1,17284982	-1,16730482	13	2	0,43
C_14	0,0141262103	-0,744093944	1,17423249	0,229097207	14	1	0,34
C_15	-2,25693375	-0,452435367	0,406850645	0,229097207	15	1	1,07
C_16	0,470511465	-1,05874868	0,330803795	0,490922587	16	1	0,48
C_17	1,39414829	-0,957229058	0,928808566	1,01457335	17	1	0,85
C_18	0,166254629	1,21730999	-1,10025965	-1,25457994	18	3	0,29
C_19	-0,974708509	-1,22925677	1,19842921	1,75641192	19	1	0,67
C_20	-0,366194835	0,324946916	-1,17630649	-1,21094238	20	3	0,39

Рис. 121. Таблица распределения стран по кластерам

Раздел 7. Расчет размера (объема) выборки или анализ мощности.

Расчет объема выборки – это один из существенных этапов планирования эксперимента. Решение вопроса о размере групп (мощности исследования) необходимо для того, чтобы при анализе полученных данных избежать ошибки второго рода. Напомню, что ошибка 1-го рода – это вероятность ложно отклонить нулевую гипотезу, т.е. найти различия там, где их нет. Максимально допустимая вероятность этой ошибки равна 5% и называется уровнем значимости. Ошибка 2-го рода – это вероятность ложно принять нулевую гипотезу т.е. не найти различий там, где они есть.

Для того, чтобы получить представление о размере выборки, необходимо знать несколько показателей:

1. Размер ожидаемого эффекта;
2. Средние значения признаков или переменных;
3. Стандартное отклонение средних значений исследуемых признаков или переменных (величину дисперсии).

Возникает вопрос, где брать эти показатели, если исследование еще только планируется и они просто не известны? В этом случае информацию, необходимую для оценки объема выборки, получают либо из результатов собственных предыдущих исследований, либо из аналогичных исследований, описанных в литературных источниках. Кроме того, придется сделать некоторые допущения.

Разберем один пример: предположим, что мы проводим клиническое исследование эффектов двух лекарств (A и B), которые воздействуют на систолическое артериальное давление. У нас имеется достаточно ресурсов для того, чтобы привлечь к исследованию 25 пациентов для тестирования каждого из этих лекарственных средств. Будет ли этого достаточно для того, чтобы обнаружить значимые результаты? Иными словами, будет ли наше исследование иметь достаточную мощность?

Первый вопрос, на который нам необходимо ответить – насколько большим является размер эффекта, который необходимо обнаружить? Иными словами, насколько должно измениться систолическое давление у пациентов, использующих тот или иной препарат? Конечно, мы этого не знаем, именно поэтому мы проводим исследование! Но можно сделать некие предположения. Например, у нас есть результаты от предыдущих исследований, которые включали в себя лекарство A, и мы считаем, что среднее артериальное давление для лекарства B будет отличаться примерно на 10% от среднего для лекарства A. Если среднее систолическое артериальное давление для лекарства A составляет 120 мм. рт. ст., то размер эффекта составит 12 мм. рт. ст.

Второй вопрос – какова вариабельность измерения систолического артериального давления? Предыдущее исследование лекарства A продемонстрировало, что стандартное отклонение систолического артериального давления равняется 10 мм. рт. ст. Предположим, что стандартное отклонение будет примерно одинаково в группах, получающих любое из этих лекарственных средств.

Опираясь на эти положения, можно рассчитать мощность исследования.

1. Из меню запустить соответствующий модуль: **Statistics / Power Analysis** (Статистика / Анализ мощности) (рис. 122).

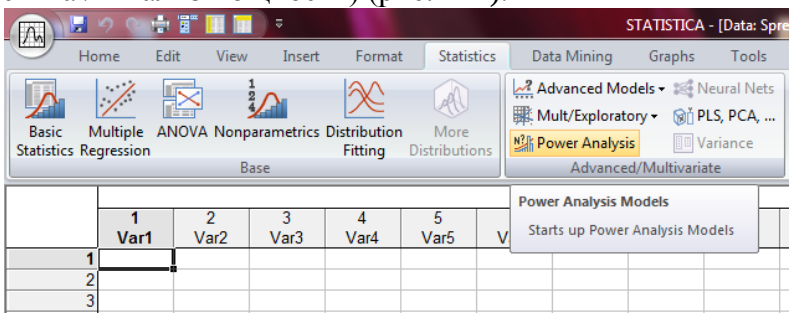


Рис. 122. Окно модуля анализа мощности

2. В появившемся окне выбрать критерий **Two Means, t-Test, Independent Samples** (критерий Т-Стьюдента для независимых выборок) (рис. 123).

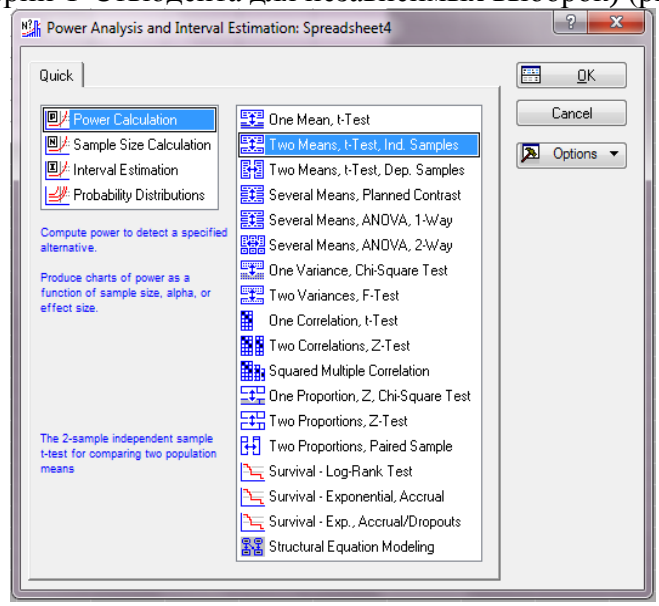


Рис. 123. Диалоговое окно для выбора статистического критерия

3. Нажать **ОК** и в следующем диалоговом окне задать известные параметры. *Mu1* и *Mu2* – это известные и ожидаемые средние значения показателей артериального давления в исследуемых группах; *N1* и *N2* – количество больных, которое планируется привлечь к исследованию; *Sigma* – стандартное отклонение в исследуемых группах: *Alpha* – уровень ошибки первого рода (α) = 0.05 (рис. 124).

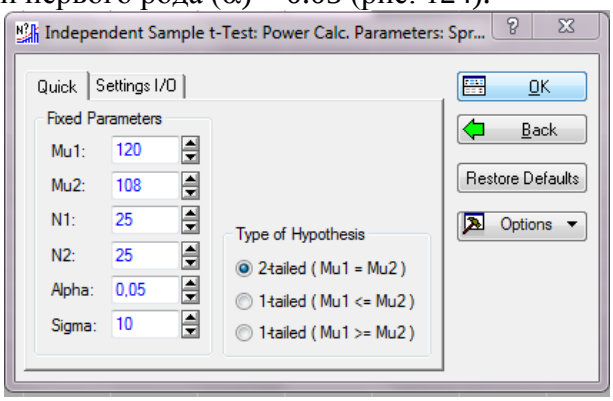


Рис. 124. Диалоговое окно для ввода известных параметров

4. Нажать **ОК**. Появляется диалоговое окно, в котором отображаются параметры, на основе которых проводится анализ (рис. 125).

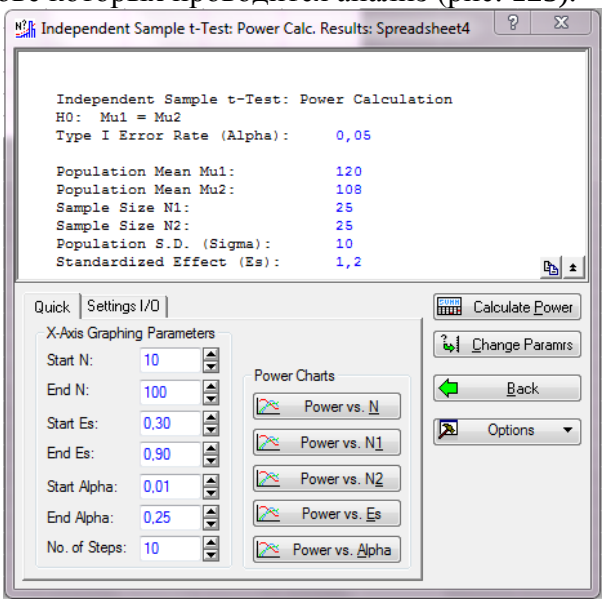


Рис. 125. Окно отображения основных параметров расчета

5. Для вычисления мощности с учетом заданных параметров нажать кнопку **Calculate power** (Рассчитать мощность). Итоговая таблица будет содержать результаты оценки мощности, как показано на рисунке (рис. 126).

	Power Calculation (Spreadsheet4) Two Means, t-Test, Ind. Samples H0: Mu1 = Mu2
	Value
Population Mean Mu1	120,0000
Population Mean Mu2	108,0000
Population S.D. (Sigma)	10,0000
Standardized Effect (Es)	1,2000
Sample Size N1	25,0000
Sample Size N2	25,0000
Type I Error Rate (Alpha)	0,0500
Critical Value of t	2,0106
Power	0,9860

Рис. 126. Результаты оценки мощности исследования

В таблице видно, что для такой комбинации параметров мощность равна 0.98. Минимально допустимый уровень мощности для биологических исследований не должен быть меньше 0.8, то есть планируемый объем выборки является более чем достаточным.

Если мощность мала (Power < 0.8), то надо понять, при каком значении N мы получим нормальную мощность.

6. Для этого необходимо нажать кнопку **Power vs. N** (рис. 127).

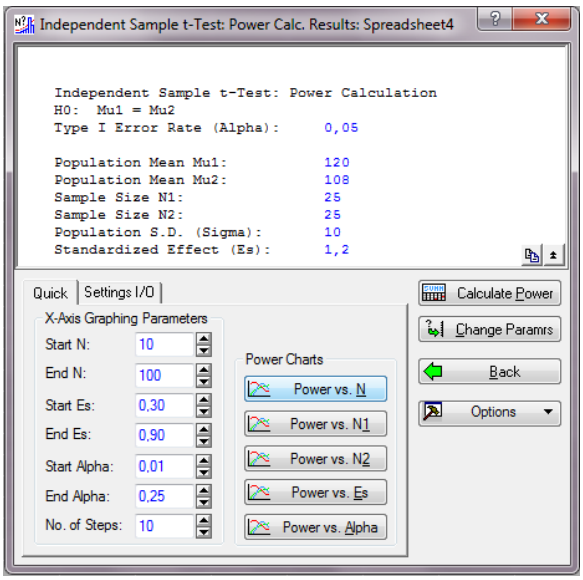


Рис. 127. Диалоговое окно расчета необходимой мощности исследования

7. В следующем окне появится график соотношения объема выборки и мощности (рис. 128). На графике видно, что мощность, равную 0.8, можно достичь при выборке, равной примерно 13 человек.

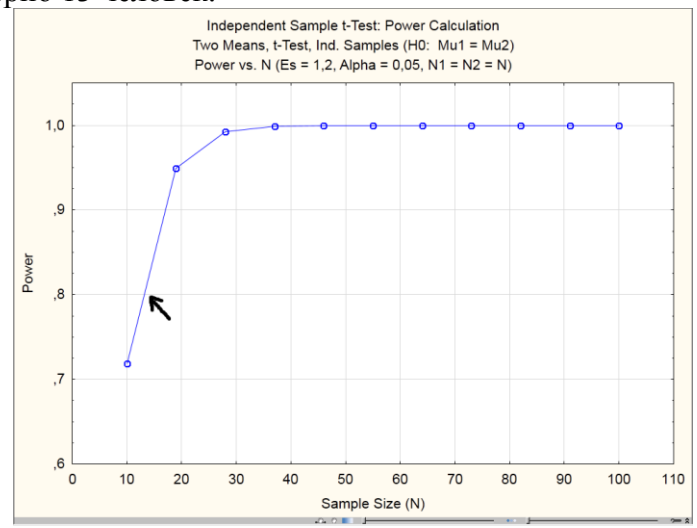


Рис. 128. График соотношения объема выборки и мощности

Аналогичным образом производится анализ мощности для зависимых выборок.

Обратимся к данным, полученным в ходе наблюдений за изменениями количества иммуноглобулинов у мышей «до» и «после» физической нагрузки (стр. 20). Напомню, что в исследовании было задействовано 19 животных, средние значения количества иммуноглобулинов «до» и «после» нагрузки составили 6,8 мм. и 7,3 мм. (диаметр кольца преципитации), стандартное отклонение 1,2 мм. и 1,4 мм.

1. Из меню запустить модуль **Power analysis / Two Means, t-Test, Dependent Samples** (рис. 129).

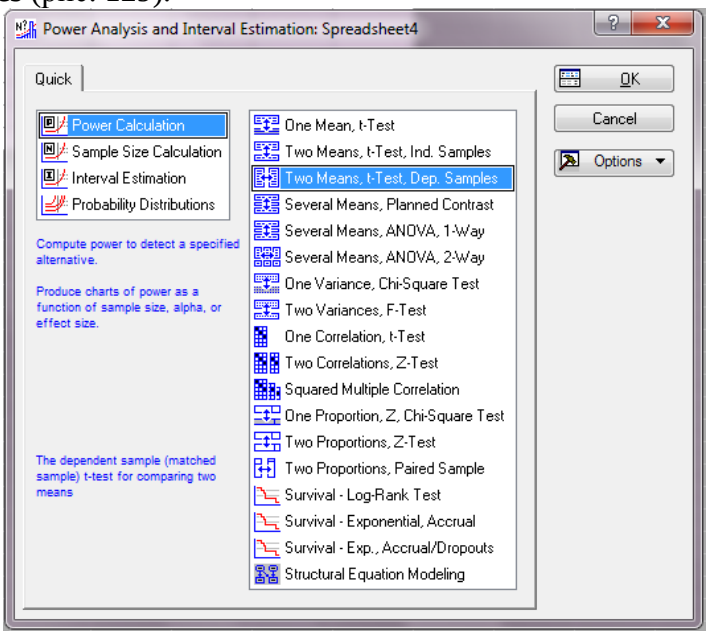


Рис. 129. Диалоговое окно для выбора статистического критерия

2. Нажать **ОК** и в следующем диалоговом окне (рис. 130) задать уже известные параметры (*N, Mu1 и Mu2, Sigma 1 и 2* и т.д.). Кроме этого необходимо ввести показатель *Rho*.

Rho – это коэффициент корреляции между двумя измерениями группами. То есть корреляция между тем, что было «до» и стало «после». Предполагается, что зависимые выборки сильно связаны, поэтому коэффициент будет высоким. Обычно его выставляют в диапазоне 0.50-0.55.

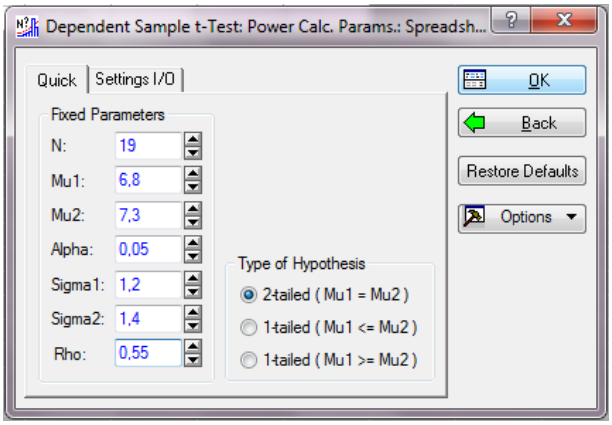


Рис. 130. Диалоговое окно для ввода известных параметров

3. Нажать **ОК**. В результате появится окно, аналогичное предыдущему, где нажимаем кнопку **Calculate power**. В таблице видно, что для такой комбинации параметров мощность равна 0.38, что намного меньше 0.8, то есть, выборка из 19 животных является недостаточной (рис. 131).

	Power Calculation (Spreadsheet4) Dependent Sample t-Test H0: Mu1 = Mu2
	Value
Population Mean Mu1	6,8000
Population Mean Mu2	7,3000
Group 1 S.D. (Sigma1)	1,2000
Group 2 S.D. (Sigma2)	1,4000
Between-group Correlation	0,5500
Stand. Error of Mean Diff.	1,2458
Standardized Effect (Es)	-0,4014
Group Sample Size (N)	19,0000
Type I Error Rate (Alpha)	0,0500
Critical Value of t	2,1009
Power	0,3808

Рис. 131. Результаты оценки мощности исследования

4. Для понимания, при каком значении N мы получим нормальную мощность, нажать кнопку **Power vs. N** (рис. 132).

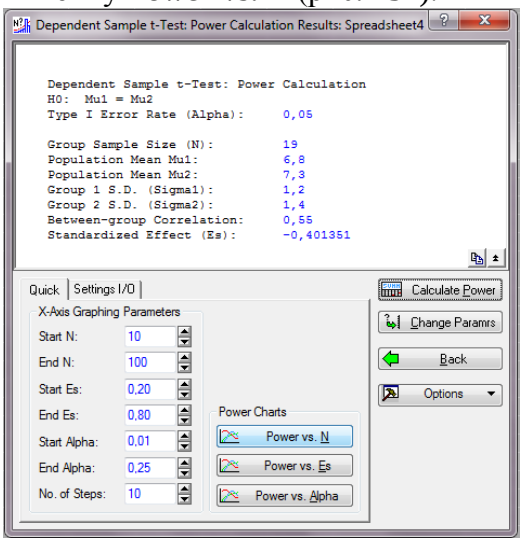


Рис. 132. Диалоговое окно расчета необходимой мощности исследования

Из полученного графика соотношения объема выборки и мощности видно, что мощность, равную 0.8 можно достичь при выборке равной примерно 50 животным (рис. 133).

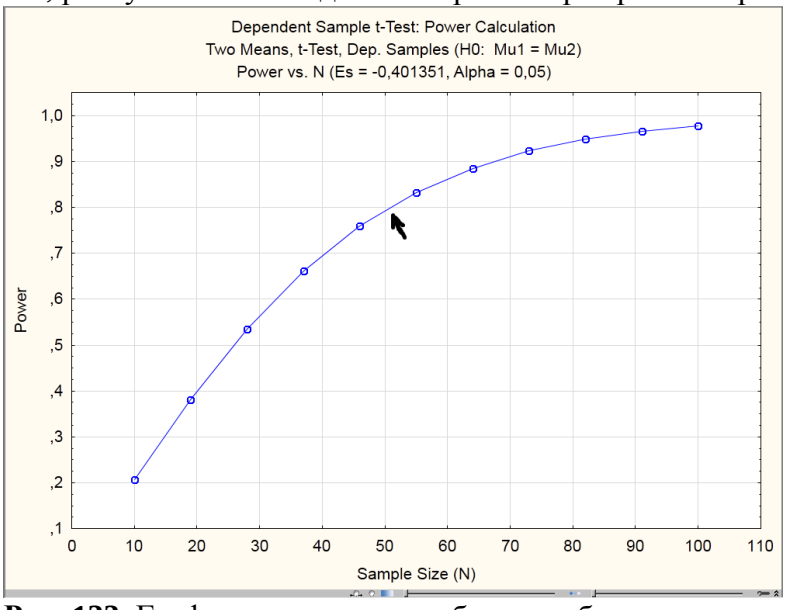


Рис. 133. График соотношения объема выборки и мощности

Список литературы:

1. Барковский С.С. Многомерный анализ данных методами прикладной статистики: Учебное пособие / С.С. Барковский, В.М. Захаров, А.М. Лукашов, А.Р. Нурутдинова, С.В. Шалагин. – Казань: Изд. КГТУ, 2010. – 126 с.

2. Гланц С. Медико-биологическая статистика / С. Гланц Пер. с англ. – М., Практика, 1998 – 459 с.

3. Давиденко Т.Н. Многомерные методы статистического анализа данных в экологии / Т.Н. Давиденко, О.Н. Давиденко, В.В. Пискунов, В.А. Болдырев. – Саратов: Изд-во Сарат. ун-та, 2006.– 56 с.: ил.

4. Койчубеков Б.К. Определение размера выборки при планировании научного исследования / Б.К. Койчубеков, М.А. Сорокина, К.Э. Мхитарян // Международный журнал прикладных и фундаментальных исследований. 2014. №4. С. 71-74.

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

6. Мухаматзанова М.Ш. О выборе метода статистической обработки данных для медико-социологических исследований / М.Ш. Мухаматзанова, М.А. Захарова, В.А. Вельш // Бюллетень Волгоградского научного центра РАМН. 2009. № 2. С. 51-53.

7. Платонов А.Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы / А.Е. Платонов. – М.: Издательство РАМН, 2000. – 52 с.

8. Стукач О.В. Программный комплекс Statistica в решении задач управлением качеством: учебное пособие / О.В. Стукач. Томский политехнический университет. – Томск – Изд-во Томского политехнического университета, 2011. 163 с.

9. Salkind N.J. Statistics for People Who (Think They) Hate Statistics. / N.J. Salkind. Fifth Edition. - Publisher: SAGE Publications, Inc [Paperback] 2014.

ISBN 978-5-4312-0652-8

Отпечатано в авторской редакции с оригинал-макета заказчика

Подписано в печать Формат….
Печать офсетная. Усл. печ. л. Уч.-изд. л.
Тираж экз. Заказ №

Издательский центр «Удмуртский университет»
426034, Ижевск, Университетская, д. 1, корп. 4