

Основы статистики.

ЛАБОРАТОРНЫЙ ПРАКТИКУМ

ОГЛАВЛЕНИЕ

Лабораторная работа № 1	
Основные термины. Знакомство с программами для статистического анализа: пакет анализа MS Excel и Statistica 6	4
Лабораторная работа № 2	
Описательная статистика. Построение графиков распределения	31
Лабораторная работа № 3	
Сравнение групп. Дисперсионный анализ.....	54
Лабораторная работа № 4	
Сравнение групп. Критерий Стьюдента	72
Лабораторная работа № 5	
Анализ зависимостей. Корреляционный и регрессионный анализ. Парная корреляция.....	85
Лабораторная работа № 6	
Криволинейная корреляция и регрессия.....	101
Лабораторная работа № 7	
Сравнение групп Непараметрические критерии для анализа количественных признаков	115
Лабораторная работа № 8	
Анализ качественных признаков.....	138
Лабораторная работа № 9	
Классификация. Кластерный и дискриминантный анализы.....	154
Литература.....	176

Лабораторная работа № 1

Основные термины. Знакомство с программами для статистического анализа: пакет анализа MS Excel и Statistica 6

Краткие сведения из теории

Биомедицинская информация — это сведения о свойствах биологических объектов и явлениях, являющихся предметами медицинских исследований, а также представления и суждения об этих свойствах и явлениях.

Биомедицинская статистика — инструмент для анализа данных, полученных в ходе эксперимента и клинических наблюдений из повседневной практики, а также язык, с помощью которого исследователь сообщает читателю полученные им результаты.

Какие бывают данные?

Любое статистическое исследование в первую очередь работает с данными (показателями или признаками того или иного исследуемого объекта).

Данные, полученные в ходе эксперимента, могут быть качественными, количественными и порядковыми. Для корректного использования статистических методов **важно** представлять, какого типа данные будут обрабатываться.

Количественные данные — признаки, которые можно выразить в числовой форме: возраст, вес, количество детей в семье и т. п. В свою очередь, они делятся на непрерывные и дискретные.

Непрерывные данные (continuous data) — количественные данные, которые могут принимать любое значение на непрерывной шкале. Другое название — *признаки, измеряемые в интервальной шкале* (температура, артериальное давление, рост). Например, рост взрослого человека может принимать *любое* значение в интервале от 150 до 220 см: 178, 178,25, 182,33456 см, т. е. произвольное числовое значение на шкале в заданном интервале.

Дискретные данные (discrete data) — количественные данные, принимающие, как правило, конечное число значений, хотя иногда и очень большое: количество смертей в течение года в исследуемой группе, количество пропущенных по болезни рабочих дней.

Качественные данные (классификационные, неупорядоченные, номинальные) — это признаки, которые нельзя выразить количественно: диагноз, место проживания, пол, жив человек или мертв, есть температура или нет и т. п.

Порядковые данные — показатели, измеряемые в шкале порядка (например, стадии болезни, оценки — «плохо», «удовлетворительно», «хорошо»). При этом порядок изменить нельзя, только в обратном направлении, но смысл от этого не меняется. Такие признаки могут быть осмысленно оцифрованы, но важно понимать, что порядок состояний имеет значение. Часто к

таким показателям следует относить балльные оценки, полученные при проведении тестов или экспертиз. Особенность порядковых шкал — отсутствие количественного измерения расстояний между величинами на шкале (можно сказать, что течение болезни «хуже», чем среднетяжелое, при этом очень тяжелое «еще хуже», однако сложно сказать во сколько раз «хуже»).

Для различных типов переменных применяются разные методы статистического анализа.

Генеральная совокупность и выборка. Свойства выборки

Обычно исследователь в ходе статистического анализа стремится сделать выводы обо всей совокупности объектов (например, как некий препарат воздействует на каждого человека с конкретной болезнью). В сущности, в этом и заключается смысл анализа: иметь представление о свойствах *всех* изучаемых объектов по тому или иному признаку (например, артериальное давление — признак, люди в возрасте от 30 до 45 лет — исследуемый объект). Весь массив исследуемых объектов образует *генеральную совокупность*. Генеральная совокупность обычно представляет собой достаточно *большое число элементов*, исследователь, в силу различных факторов не может осуществить эксперимент над всеми элементами генеральной совокупности, поэтому он останавливается на достаточном количестве элементов, по возможности характеризующим всю генеральную совокупность. Это количество исследуемых объектов называются *выборкой*. Предполагается, что выборка характеризует всю генеральную совокупность, если это условие выполняется, то такую выборку называют *репрезентативной (представительной)*. Репрезентативность — очень важное свойство выборки, если выборка не является репрезентативной, то исследователь может сделать ошибочные выводы обо всех объектах исследования (всей генеральной совокупности). Стоит заметить, что в медицинских исследованиях часто бывает так, что выборки имеют очень небольшой объем (обычно в формулах число элементов выборки обозначается как n), порядка 10–20 элементов.

Обеспечение репрезентативности выборки важный аспект при планировании статистического исследования. При недостаточном качественном выполнении данного условия имеется большой шанс получить превратные представления об исследуемом объекте.

Классический пример

Классический пример **нерепрезентативной** выборки, произошедший в 1936 г. в США во время президентских выборов.

Журнал «Литэри дайджест», который до этого весьма успешно прогнозировал результаты предыдущих выборов, на этот раз ошибся в своих прогнозах, хотя разослал несколько миллионов письменных вопросов подписчикам, а также респондентам, которых они выбрали из телефонных книг и из списков регистрации автомобилей. В 1/4 бюллетеней, которые вернулись заполненные обратно, голоса распределились следующим образом: 57 % отдали первенство кандидату от республиканцев по имени Альф Лэндон, а 41 % отдали предпочтение действующему президенту — демократу Франклину Рузвельту.

В действительности, на выборах победил Ф. Рузвельт, который набрал почти 60 % голосов. Ошибка «Литэри дайджест» была в следующем. Они захотели увеличить репрезентативность выборки. А так как они знали, что большинство их подписчиков относят себя к республиканцам, то они решили расширить выборку за счёт респондентов, выбранных ими из телефонных книг и автомобильных регистрационных списков. Но они не учли существующих реалий и фактически отобрали ещё больше сторонников республиканцев, потому что во времена Великой депрессии иметь автомобили и телефоны мог позволить себе средний и высший класс. А это и были по большей части республиканцы, а не демократы.

Еще одним важным свойством выборки является ее **случайный характер (рандомизация)**. Это означает, что каждый член генеральной совокупности **равновероятно** может попасть в выборку для проведения эксперимента.

Т. е. вероятность оказаться в выборке одинакова для всех членов генеральной совокупности.

Осуществить рандомизацию выборки необходимо с целью снижения возможной подтасовки результатов. Например, если исследуется воздействие препарата на артериальное давление и в генеральную совокупность входят люди разной возрастной группы, но с одинаковыми показаниями к препарату, стоит учитывать, что исследователь может выбрать людей помладше, тем самым улучшить показатели воздействия препарата, и подобный отбор уже не является случайным. Следовательно, выводы могут оказаться завышенными или заниженными, слишком оптимистичными или наоборот.

Распределение значений признака. Полигон частот

Каждая генеральная совокупность характеризуется **распределением** значений исследуемой переменной (признака) или графическим представлением **частоты встречаемости**.

Другими словами, графическим представлением того, как часто (сколько раз) появляется в результатах эксперимента то или иное значение переменной.

Выборка также характеризуется распределением признака (*выборочное распределение*).

Пример:

В результате исследования группы людей на предмет влияния правильности метода лечения на сроки госпитализации (где переменной является число дней госпитализации) были получены следующие значения:

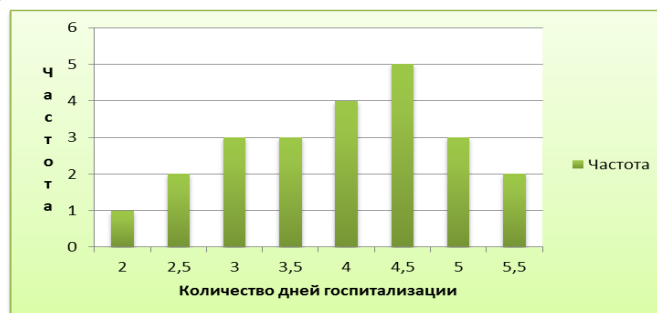
Количество дней госпитализации	2	2,5	2,5	3	3	3	3,5	3,5	3,5	4	4	4	4	4,5	4,5	4,5	4,5	4,5	5	5	5	5,5	5,5
--------------------------------	---	-----	-----	---	---	---	-----	-----	-----	---	---	---	---	-----	-----	-----	-----	-----	---	---	---	-----	-----

Запишем их в виде таблицы частот:

Количество дней	Частота
2	1
2,5	2
3	3
3,5	3
4	4
4,5	5
5	3
5,5	2

Под **частотой** подразумевается сколько раз то или иное значение появилось в выборке в ходе проведения эксперимента или сбора данных.

Для построения графика распределения на оси X (горизонтальной) отмечаются значения «Количество дней госпитализации», по оси Y (вертикальной) — отмечается *сколько раз то или иное значение появилось в ходе исследования*.



Обычно строят огибающую (линию тренда в MS Excel):



Столбчатую диаграмму чаще всего называют полигоном частот, огибающую линию — графиком распределения частот.

Довольно часто вместо *частоты встречаемости* на графике изображают **относительную частоту встречаемости**.

Относительная частота встречаемости конкретного члена выборки (или генеральной совокупности) объемом N определяется следующим образом:

$$\text{Количество членов выборки с заданным конкретным значением} / \text{Объем выборки}$$

Или

$$f = M/N,$$

где M — количество элементов выборки с заданным конкретным значением.

Из выше приведенного примера рассчитаем относительную частоту встречаемости дней госпитализации со значением 4,5:

$$f = 5/(1+2+3+3+4+5+3+2) = 5/23 = 0,2174.$$

Относительная частота встречаемости по количеству дней госпитализации со значением 4,5 дня равна 0,22, если это значение выразить в процентах, то получается 22 %.

Т. е. 22 % от всех участников эксперимента были выписаны спустя 4,5 суток после начала лечения.

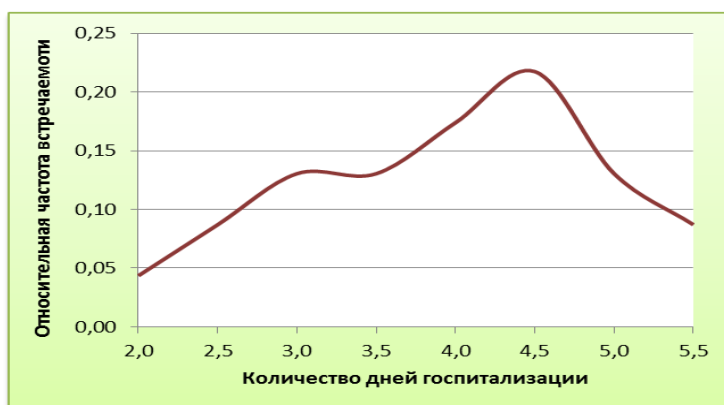
Подсчитав все относительные частоты можно получить следующую таблицу:

Количество дней	Частота встречаемости	Относительная частота встречаемости	Относительная частота встречаемости (%)
2	1	0,0435	4,35
2,5	2	0,0870	8,70
3	3	0,1304	13,04
3,5	3	0,1304	13,04
4	4	0,1739	17,39
4,5	5	0,2174	21,74
5	3	0,1304	13,04
5,5	2	0,0870	8,70
Сумма:	23	1	100

Построим гистограмму (случайная переменная — количество дней госпитализации — является дискретной, так как исследователь измеряет время пребывания в дискретных значениях, т. е. день делится не непрерывно — 2,5 дня; 2,75 дня, 3,8965 дня и т. п., а дискретно 2 дня, 2,5 дня, 3 дня и т. д.) Следовательно, для отображения распределения частот предпочтительнее выбрать график в виде **гистограммы** или **полигона частот**:



Если же представить случайную переменную как непрерывную, то можно изобразить **график распределения**:

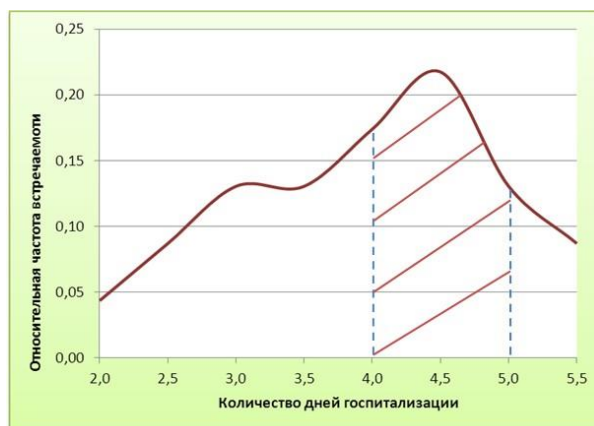


Смысл использования *относительных частот встречаемости* заключается в том, что довольно часто необходимо выразить количество членов выборки с разными значениями исследуемого признака в их **процентном соотношении**, или иными словами указать *процентное соотношение данного значения признака в выборке*. Что дает возможность предположить о частотном проявлении этого значения в генеральной совокупности, при условии, что выборка *репрезентативна*.

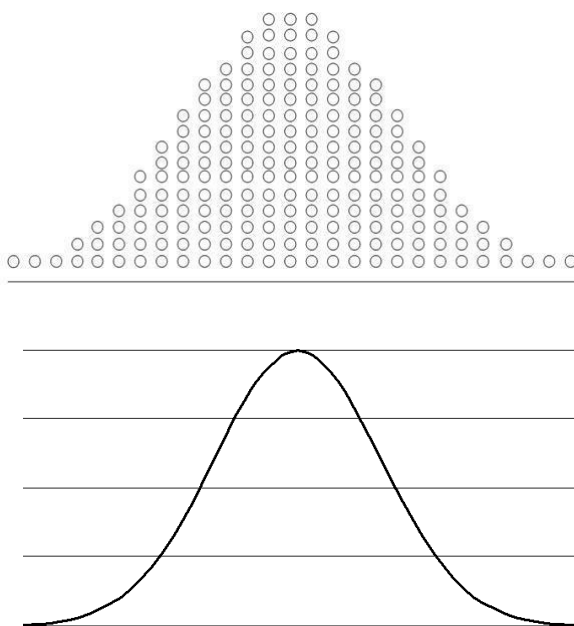
Также обратите внимание, что сумма относительных частот равна 1, а их процентного соотношения соответственно 100 %.

Забегая вперед, следует сказать, что площадь под кривой распределения **всегда** равна 1 (естественно, если при этом используется выражение *частоты встречаемости* признака в виде *относительной частоты встречаемости*).

Также: площадь ограниченной области под кривой распределения равна доле и вероятности появления признака с заданными значениями. Т. е. исходя из рисунка, доля членов выборки со значениями в интервале от 4,0 до 5,0 равна площади заштрихованной области на графике.



Далее, если принять, что рассматриваемая случайная величина (признак исследуемого объекта) *непрерывна*, то увеличивая количество измерений и при этом, уменьшая размер интервалов (карманов¹) мы получим следующие графики (графики соответствуют идеальному случаю нормального распределения):

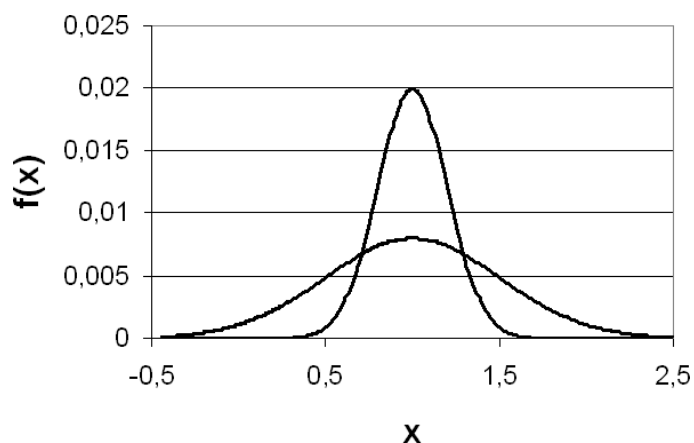
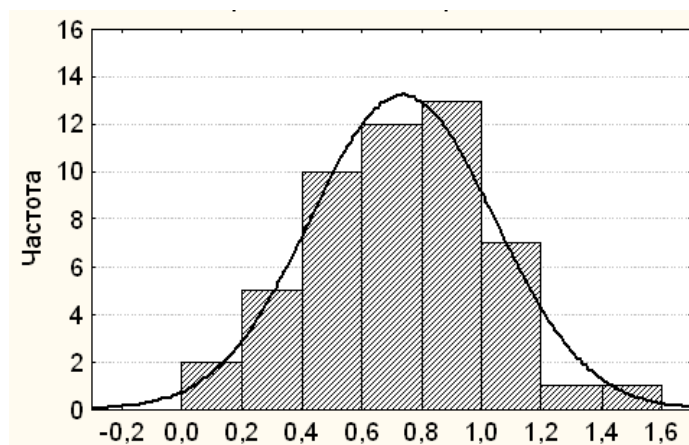


¹ **Карман** — область на оси X, которая вмещает значения переменной, соответствующие заданному интервалу. Например, если в ходе измерения веса группы людей получены непрерывные значения: 52 кг ... 57 кг, 58 кг, 58 кг, 59 кг, 60 кг, 62 кг... 69 кг... 71 кг... 76 кг и т. д. до 87 кг, то для построения гистограммы можно воспользоваться карманами: 50–65 кг, 65–70 кг, 70–75 кг и т.д. до максимального значения веса имеющегося в выборке (90 кг — граница последнего кармана). В этом случае частота встречаемости того или иного значения переменной определяется как сумма частот переменных, соответствующих интервалу кармана. При необходимости интервалы карманов можно уменьшать или увеличивать, например: 50–60 кг, 60–70 кг и т. д.

В большинстве случаев в медико-биологических исследованиях встречаются следующие виды распределения:

- Нормальное.
- Ассиметричное.
- Равномерное.
- Полимодальное.

Нормальное (колоколообразное, гауссово) распределение

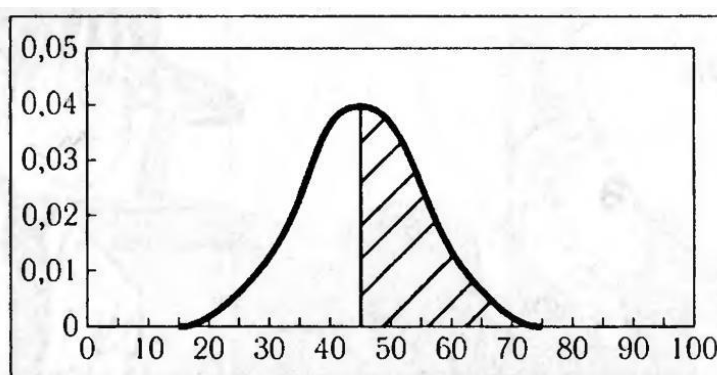


Нормальное распределение подразумевает, что большая часть значений признака находится в районе так называемого среднего значения.

При нормальном распределении наиболее часто в выборке встречаются значения близкие по величине к среднему по выборке и располагающиеся симметрично ему (значений больше среднего и значений меньше среднего приблизительно одинаковое количество). Или если выразить в процентном соотношении (используя относительные частоты встречаемости), то можно говорить, что наибольший процент значений признака находится в районе среднего значения, тогда как всего несколько процентов — по краям кривой.

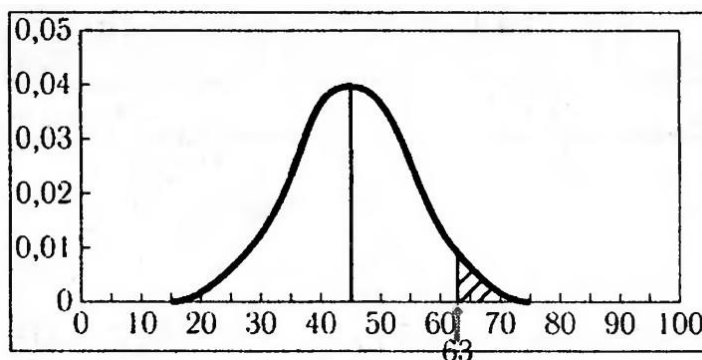
При изучении распределений как теоретической базы статистических заключений наибольший интерес представляет площадь под нормальной кривой. Эту площадь можно представить как интеграл от функции $f(x)$.

Как было сказано выше площадь под кривой распределения всегда равна 1 (при выражении частоты встречаемости в виде относительных значений), а площадь ограниченная какими-то значениями признака соответствует вероятности или доле.

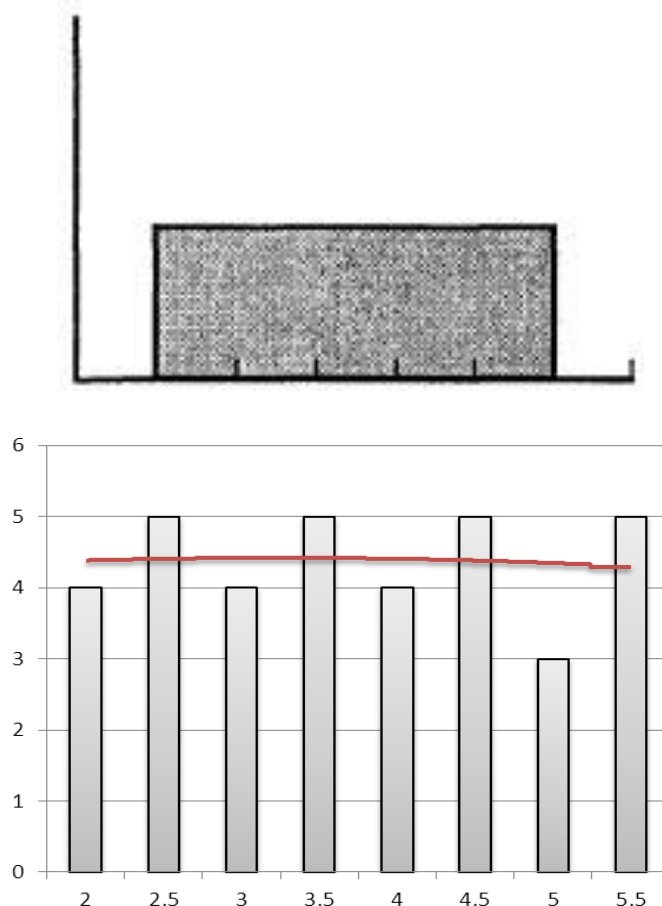


На графике изображено распределение случайной величины. Оно соответствует нормальному распределению, если разделить область под кривой пополам, то обе половины будут равной площади — 0,5 (50 %), отсюда можно говорить, что вероятность возникновения значений признака больших 45 (согласно графику) равна 0,5, следовательно доля членов выборки со значением меньше 45 также равна 0,5 (т. е. половине все членов выборки).

Если же мы захотим узнать какая вероятность возникновения признака со значениями больше 65, то изобразив это на графике: видно, что доля таких членов выборки существенно меньше и вычислив площадь под кривой получим около 3,5 %, соответственно меньше 65 равна $100\% - 3,5\% = 96,5\%$.

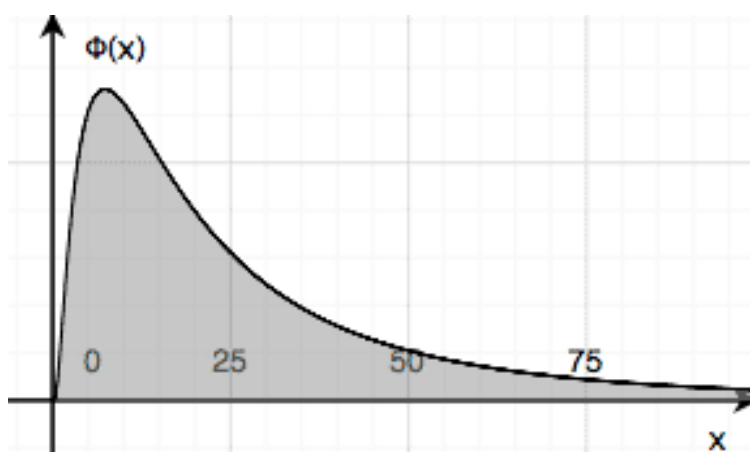


Равномерное распределение



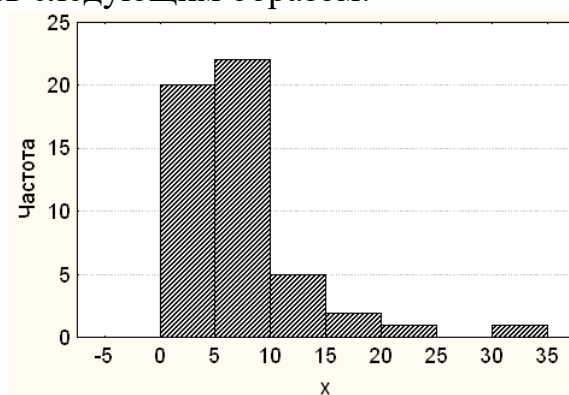
Равномерное распределение указывает на малое влияние переменной на исследуемый процесс или малое влияние исследуемого фактора на значения случайной величины (признака).

Асимметричное (если асимметрия левосторонняя — *логнормальное распределение*):

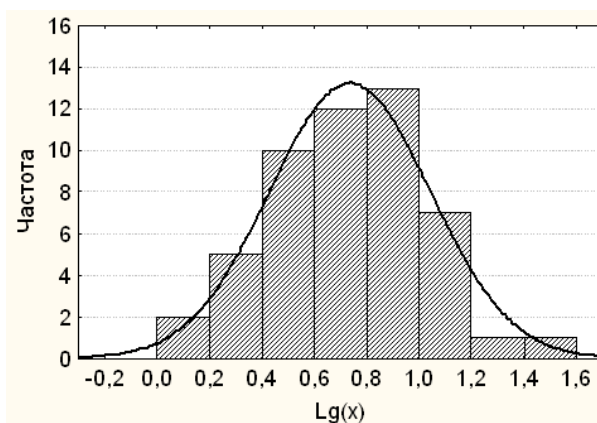


Если функцию $f(x)$ логнормального распределения преобразовать на ее логарифм $\log(f(x))$, то в этом случае полученная функция будет иметь нормальное распределение и характеризоваться теми же параметрами.

Используя графическое представление, такой случай можно продемонстрировать следующим образом:

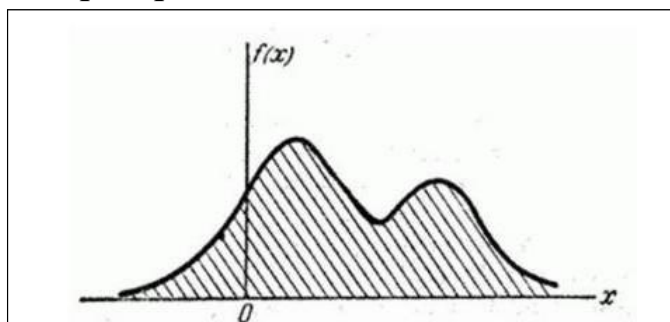


Теперь, если рассчитать логарифм десятичный от x и построить распределение получившихся значений, то мы получим следующий график:



Соответствующий нормальному распределению.

Полимодальное распределение

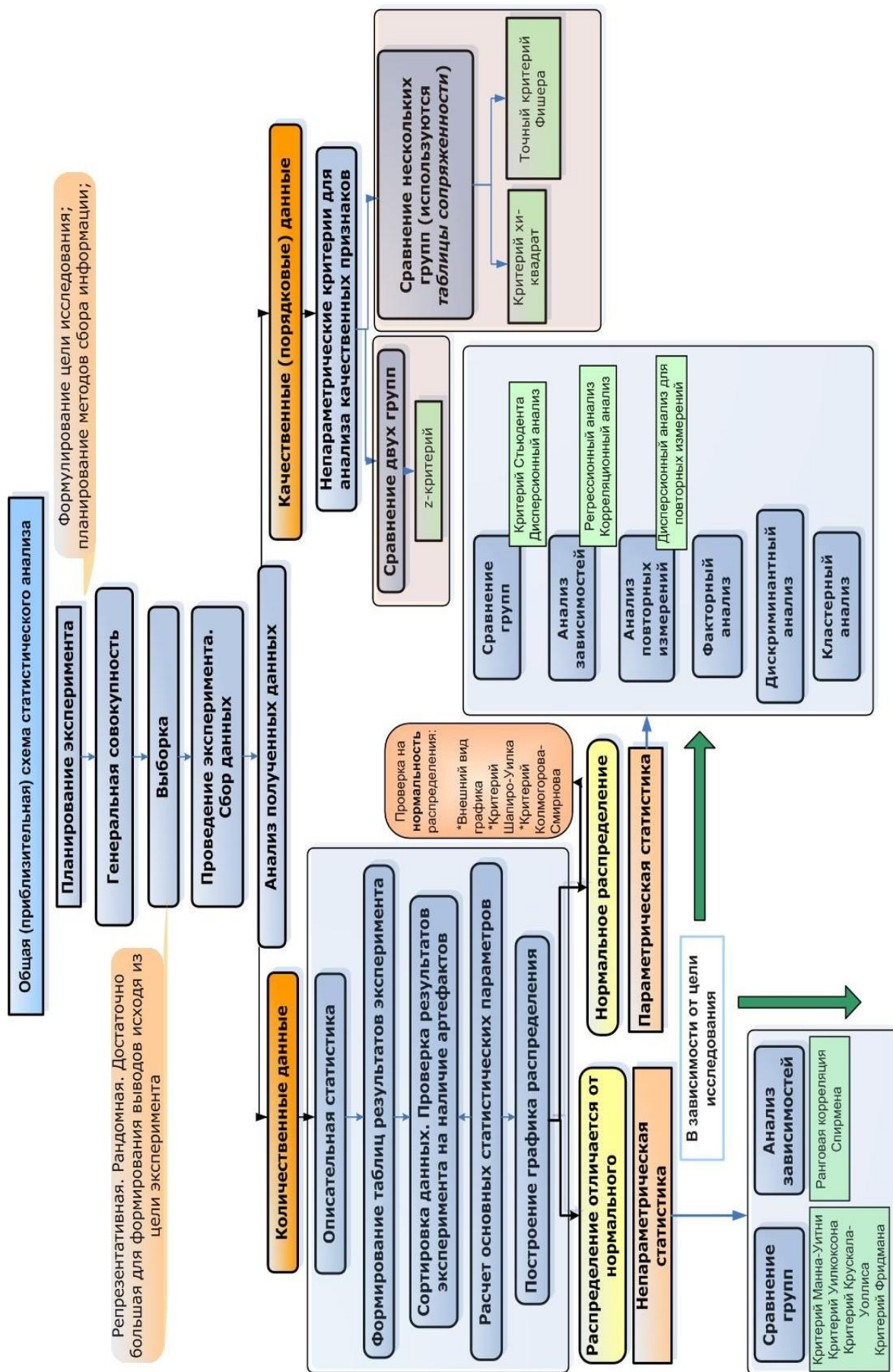


Полимодальное распределение может быть обусловлено действием нескольких скрытых факторов. Или о, возможно, неправильном построении исследования, например, выборка не является достаточно репрезентативной.

В зависимости от типа распределения выбираются методы статистического анализа.

Если распределение является нормальным или логнормальным, то применяют методы так называемой **параметрической статистики**.

Ниже приведена краткая общая схема статистических исследований.



Программы для статистического анализа Построение распределения значений признака

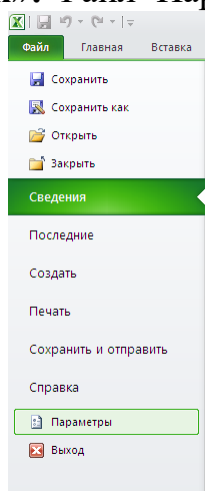
Пакет анализа MS Excel

Для статистического анализа данных в программе Excel используется модуль «Пакет анализа».

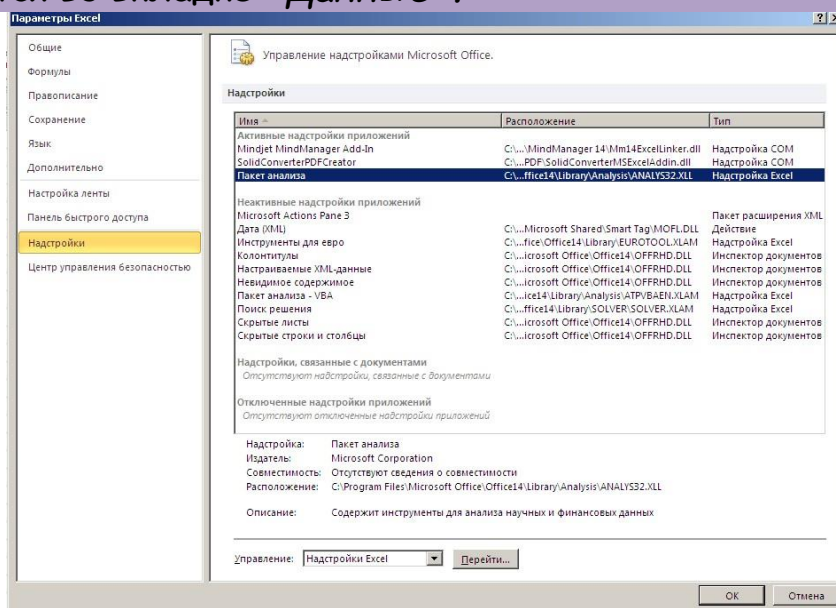
Модуль анализа находится во вкладке «Данные — Анализ данных». Если он отсутствует, то его необходимо подключить.

Для активации модуля «Анализ данных» необходимо в меню «Надстройка» выбрать «Пакет анализа». Нажать кнопку «Перейти» ... и поставить галочку напротив опции «Пакет анализа».

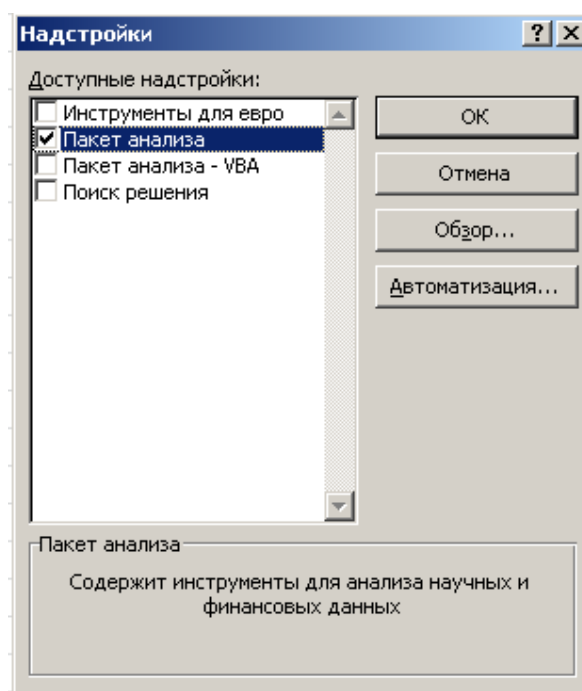
Вызов меню «Надстройки»: Файл–Параметры–Надстройки.



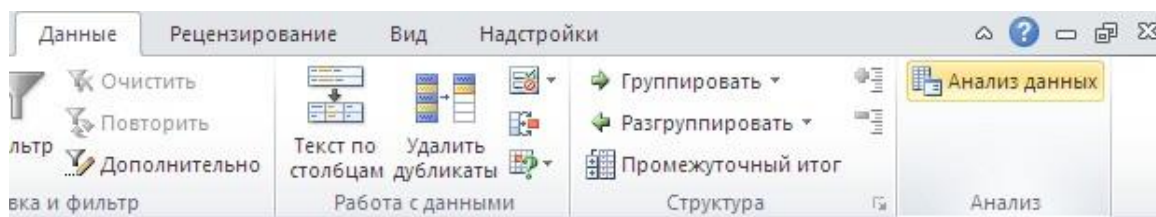
Прежде чем перейти в меню «Надстройка» для активации модуля, убедитесь в необходимости этого действия, возможно, что модуль «Анализ данных» уже активирован. В этом случае он уже находится во вкладке «Данные».



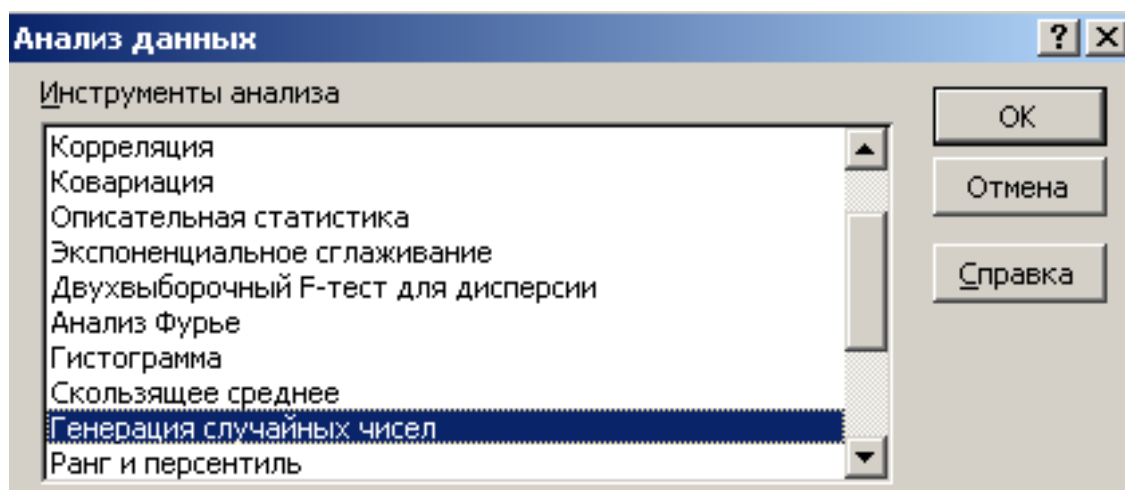
Перейти... «Ок».



Модуль анализа располагается во вкладке «Данные–Анализ данных».



Внешний вид модуля:



Генерация случайных чисел и построение графика распределения

Рассмотрим работу модуля на простейшем примере *генерации случайных чисел* и построении *графика их распределения*.

1. В системе **Excel** в меню откройте модуль «Анализ данных».

2. В модуле «Анализ данных» выберите «Генерация случайных чисел», после чего нажмите «ОК».
3. В появившемся окне выполните установки, как показано на рисунке:

Генерация случайных чисел

Число переменных: 1

Число случайных чисел: 50

Распределение: Нормальное

Параметры

Среднее = 170

Стандартное отклонение = 10

Случайное рассеивание:

Параметры вывода

☒ Выходной интервал: \$A\$1

☐ Новый рабочий лист:

☐ Новая рабочая книга

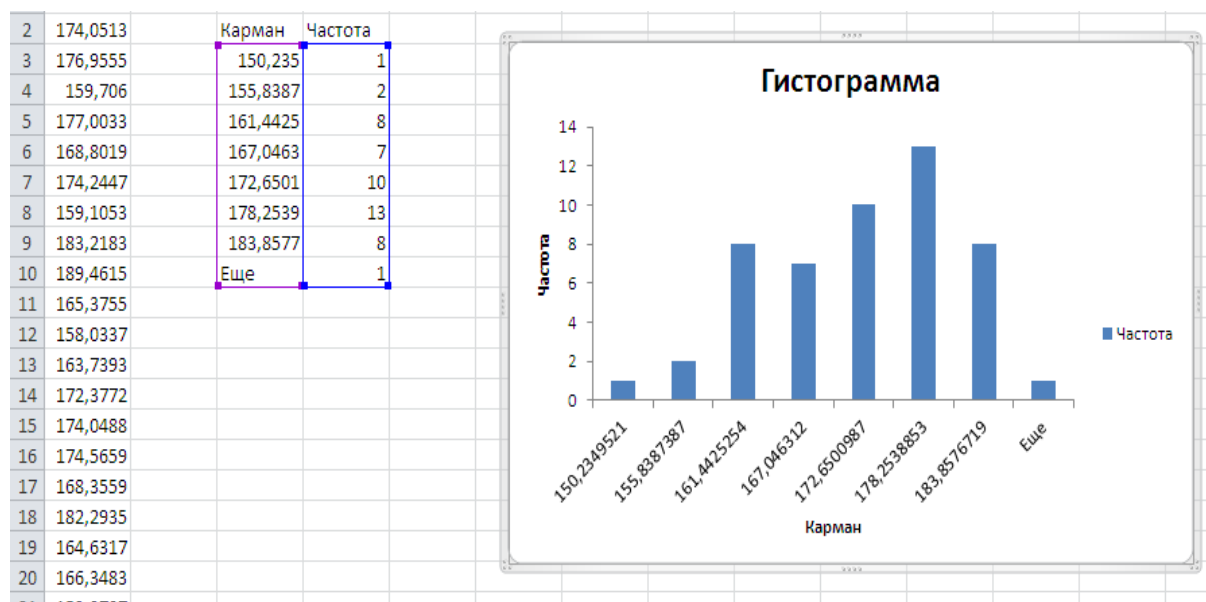
Результаты операции представлены в виде таблицы (ваши значения могут отличаться).

	A	B
1	182,3869	
2	174,0513	
3	176,9555	
4	159,706	
5	177,0033	
6	168,8019	
7	174,2447	
8	159,1053	
9	183,2183	
10	189,4615	
11	165,3755	
12	158,0337	
13	163,7393	
14	172,3772	
15	174,0488	
16	174,5659	
17	168,3559	
18	182,2935	
19	164,6317	
20	166,3483	
21	153,9707	
22	174,4963	
23	158,029	
24	157,4045	
25	154,2571	
26	166,2024	

4. Построение графика распределения. В меню «Анализ данных» выберите «Гистограмма». В появившемся окне в окошке «Входной интервал»

выберите все данные, полученные в ходе выполнения процедуры генерации случайных чисел. Остальные поля заполните, как указано на рисунке:

Результат представлен в виде таблицы карманов и графика распределения (полигон частот):

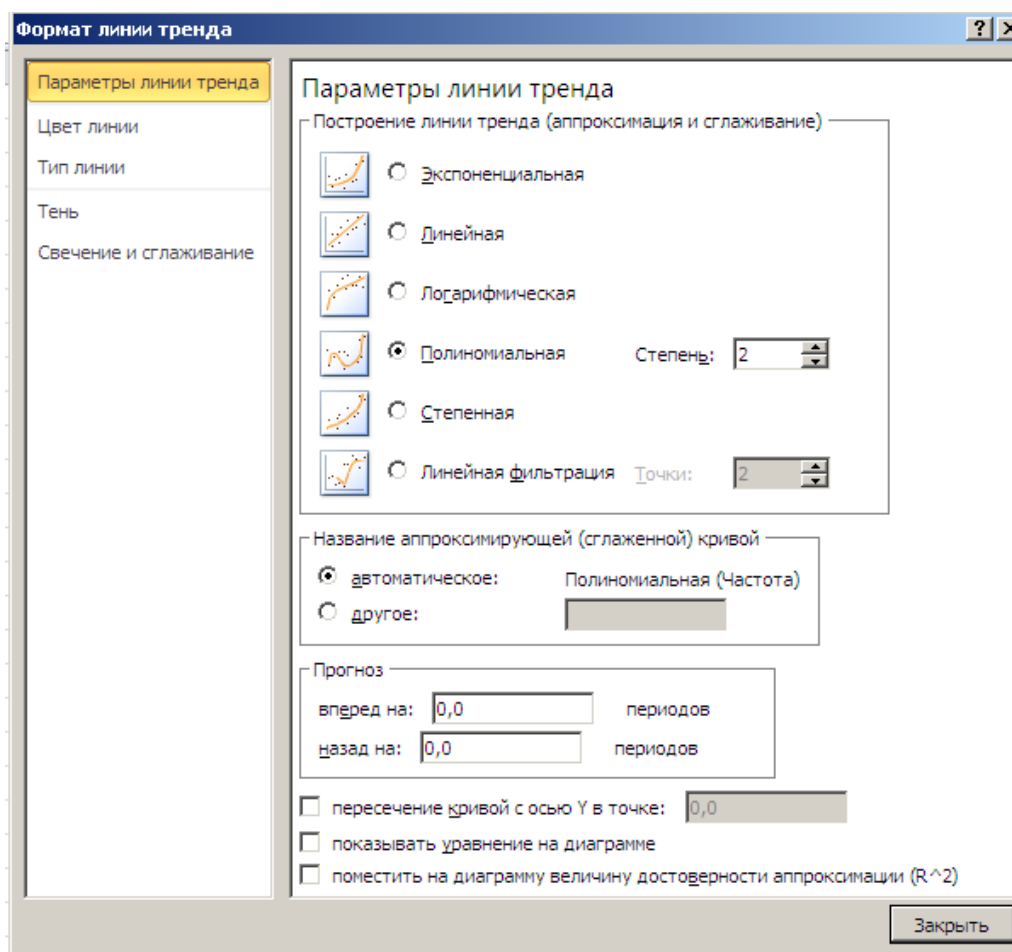


Где **карманы** — это интервалы, в которые попадают соответствующие значения в выборке. Другими словами, карман 150,235–155,8387 «вмещает» все числа, сгенерированные ранее, попадающие по своей величине в заданный интервал: 152,97; 154.25 и т. д. Карманы используются для более удобного графического представления результата. Их можно сформировать самостоятельно или предоставить эту процедуру программе.

5. Добавление линии тренда. Щелкните правой кнопкой мыши по любому столбцу на диаграмме и в появившемся меню выберите **«Добавить линию тренда»**.



В появившемся окне выбрать «Полиномиальная 2 степени». Закрыть окно.

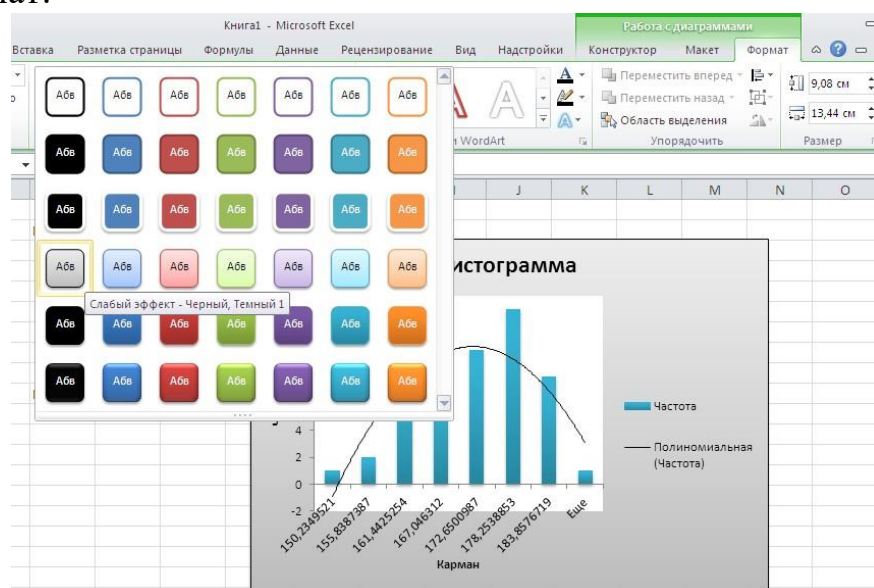


6. С помощью меню «**Работа с диаграммами**» измените внешний вид полученной диаграммы и линии тренда.

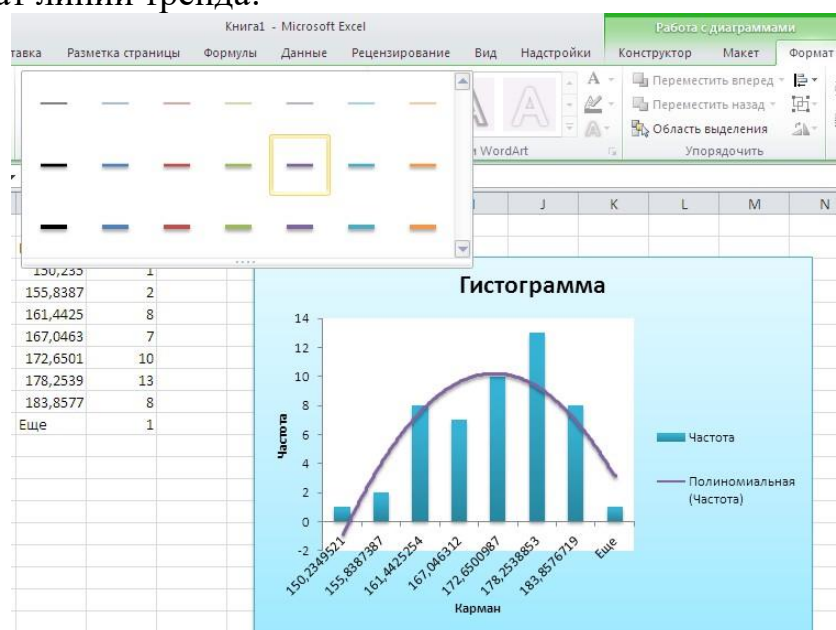
Конструктор:



Формат:

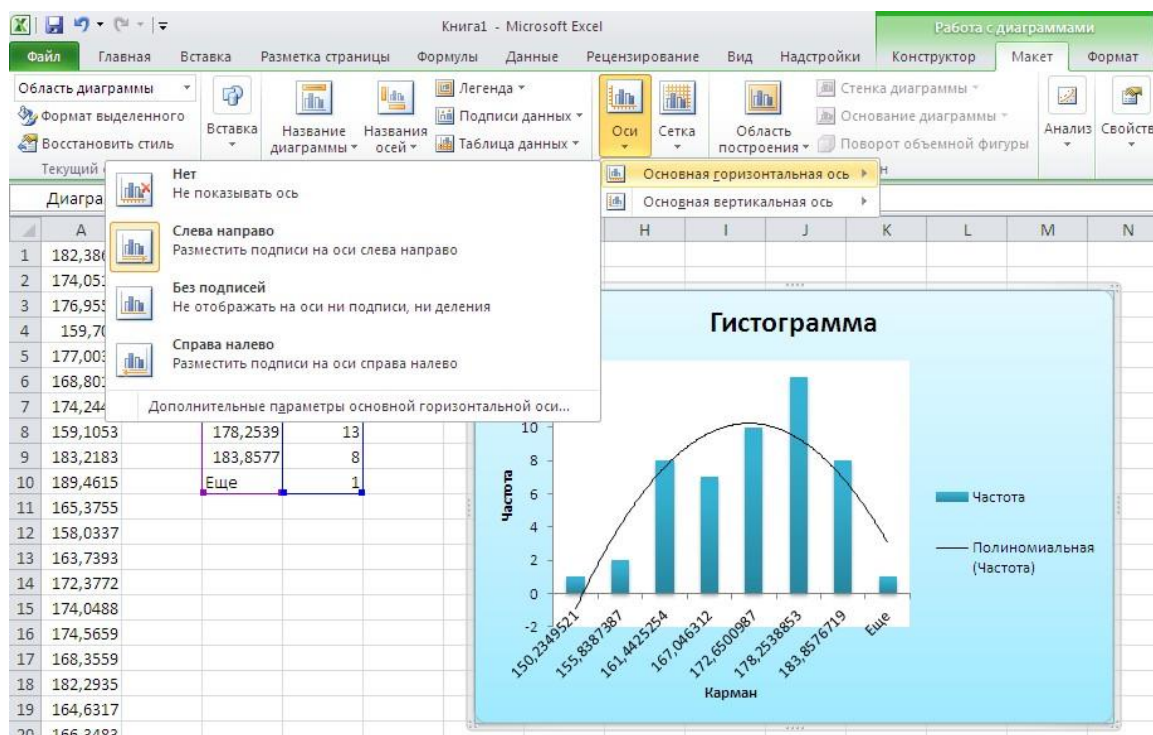


Формат линии тренда:

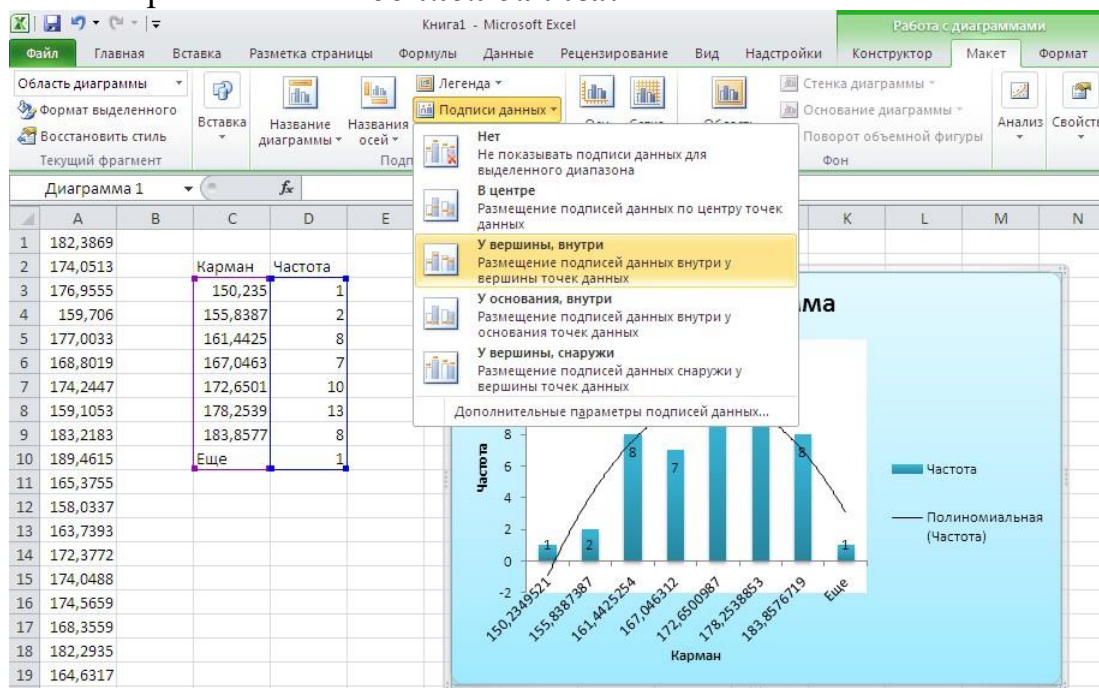


- ✓ Почему линия тренда «ушла» в отрицательную область?
- ✓ Возможна ли такая ситуация? Соответствует ли ее поведение действительности?

7. Ознакомьтесь с вкладкой «**Макет**». Измените названия осей, название диаграммы: горизонтальная ось — «*Рост*», вертикальная — «*Частота встречаемости*»; название гистограммы — «*Распределение по росту*».



Посмотрите меню «**Подписи данных**»



Удалите значение границы кармана обозначенной «Еще», замените его на подходящее по смыслу значение.

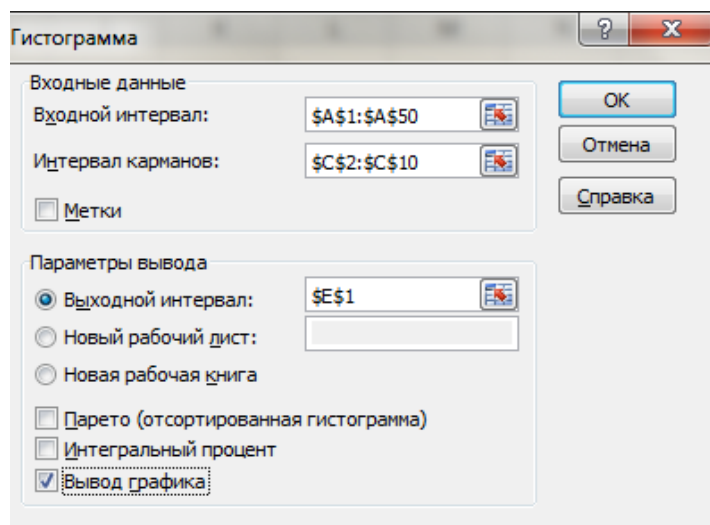
Исправьте значения интервалов карманов на более приемлемые по внешнему виду: вместо 150,2349521 на 150, 23 или 150,2.

Другими словами, доработайте внешнее представление графика до приемлемого отчетного вида.

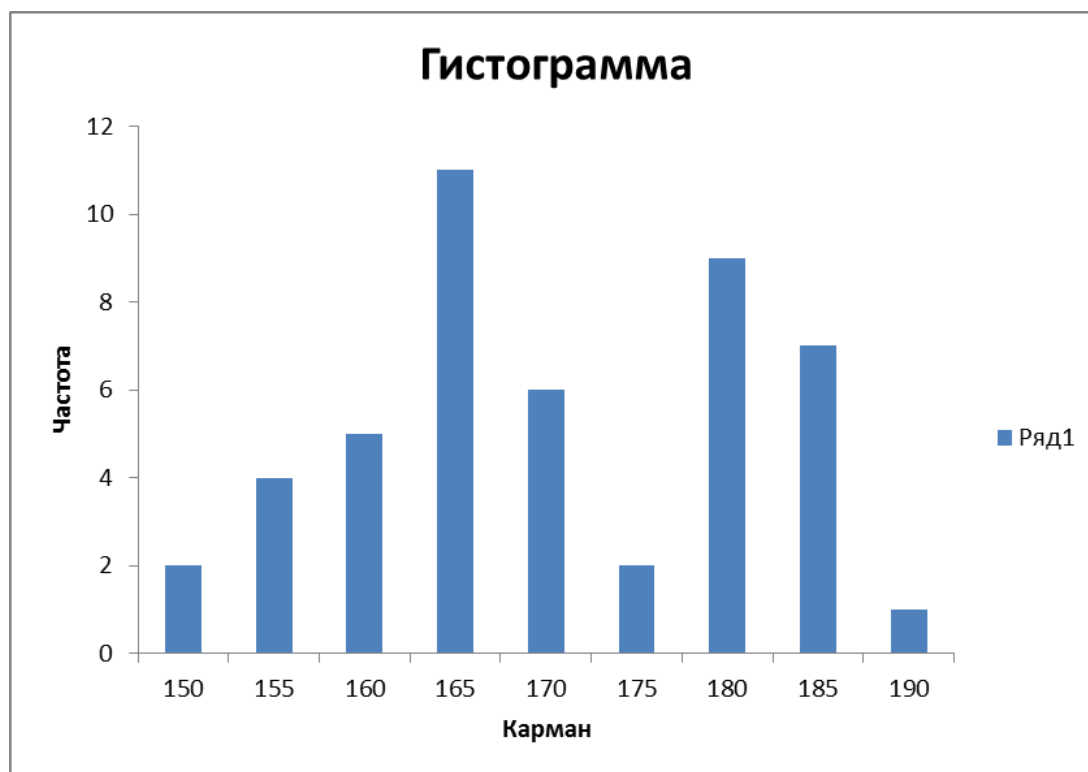
8. Вторым вариантом оформления графика может быть выполнен посредством задания интервалов карманов самостоятельно. Для этого удалите получившийся график и таблицу частот с сформированными программой интервалами *карманов*. После этого создайте таблицу карманов самостоятельно. Осуществите сортировку сгенерированных данных от минимума к максимуму (выделите таблицу с данными и выполните команду *вкладка Данные – Сортировка от А до Я*), после этого посмотрите свое минимальное значение и округлите его в меньшую сторону до круглого числа (например, 150) — это будет минимальное значение первого кармана. Далее на свое усмотрение самостоятельно задайте интервалы карманов до максимального значения в таблице сгенерированных данных, округленного в большую сторону до круглого числа (например, 190). Вы можете получить таблицу вида:

Карманы
150
155
160
165
170
175
180
185
190

Выполните пункт 4, но в графе «Интервал карманов» укажите ячейки с полученной таблицей.



Результат:



Если на графике появилось значение «Еще», то доработайте график до приемлемого вида. Линию тренда строить не обязательно.

9. Полученную таблицу случайных чисел и график распределения скопируйте в документ Word. Сохраните результат работы в отдельную папку.

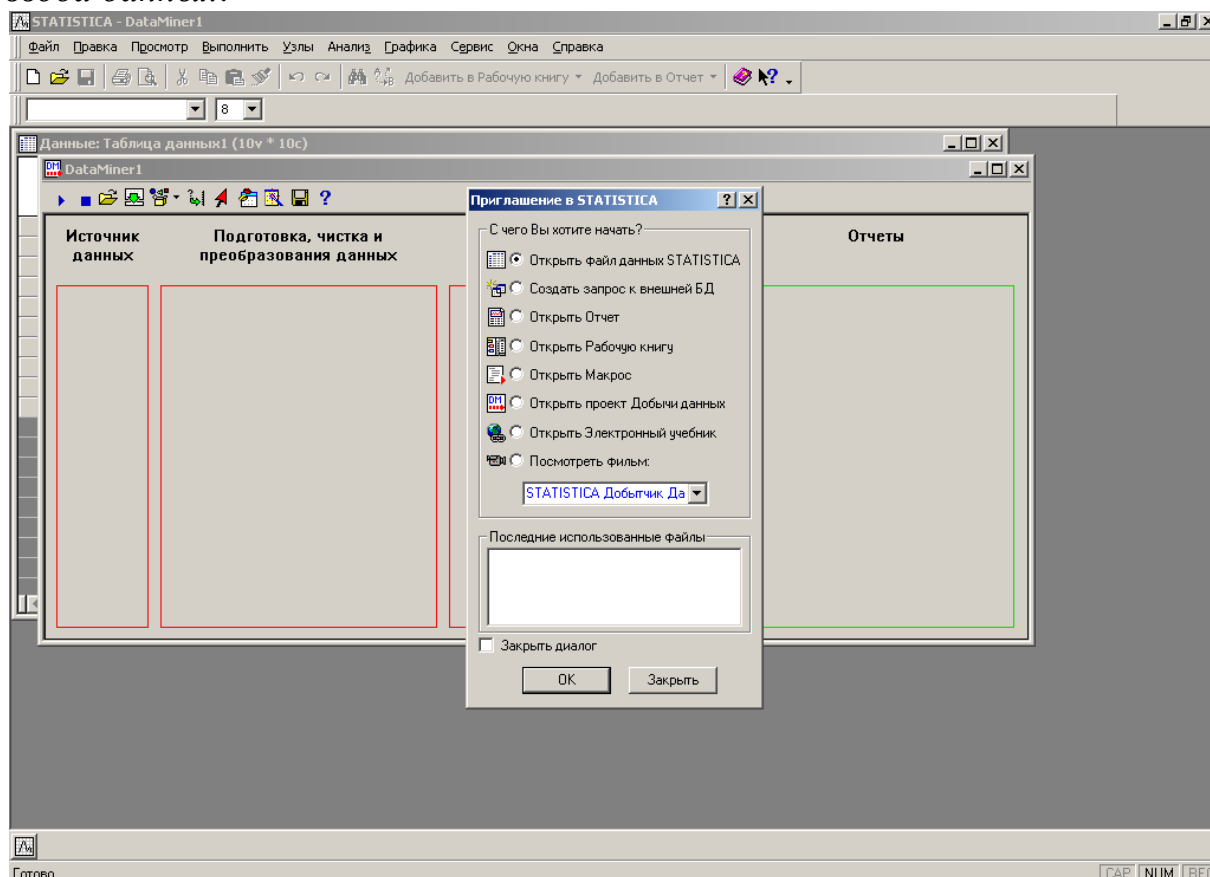
Знакомство с интерфейсом программы «Statistica» 6

Программа «Statistica» 6 является мощной специализированной платформой для статистической обработки данных на ПК.



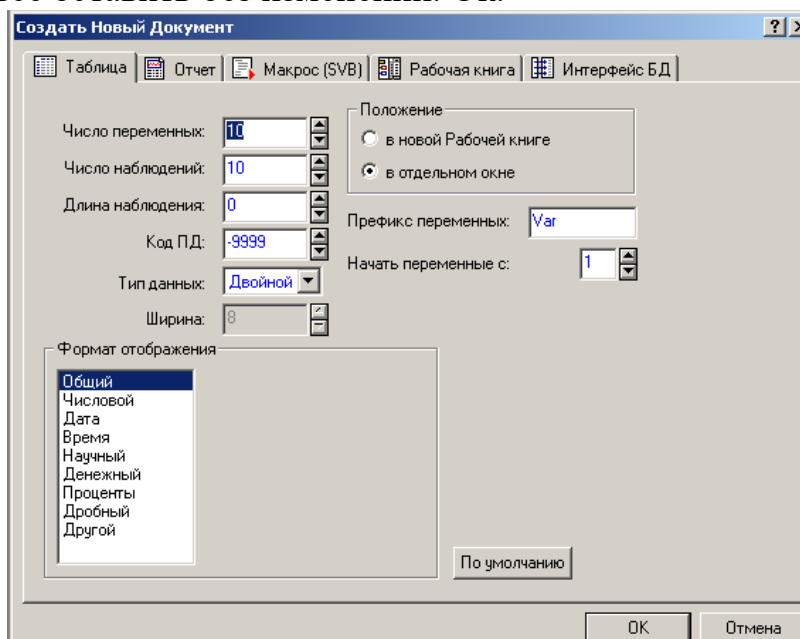
Первый запуск

При первом запуске программы появляются окно *помощника* и окна *ввода данных*.

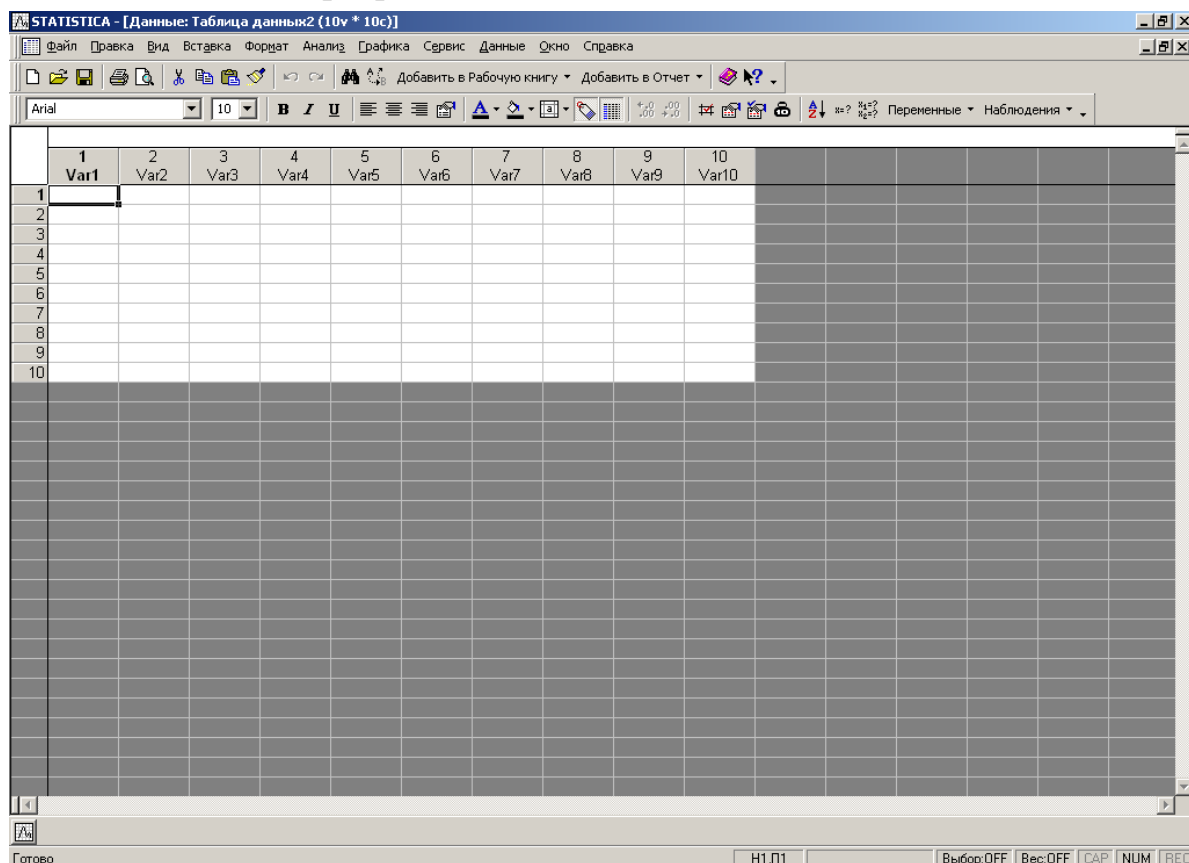


Закрыв все начальные окна необходимо создать новый файл.

В появившемся окне необходимо ввести *число переменных*, *количество наблюдений*, *формат отображения данных*. При первом знакомстве все оставить без изменений. Ок.



Рабочее поле программы:

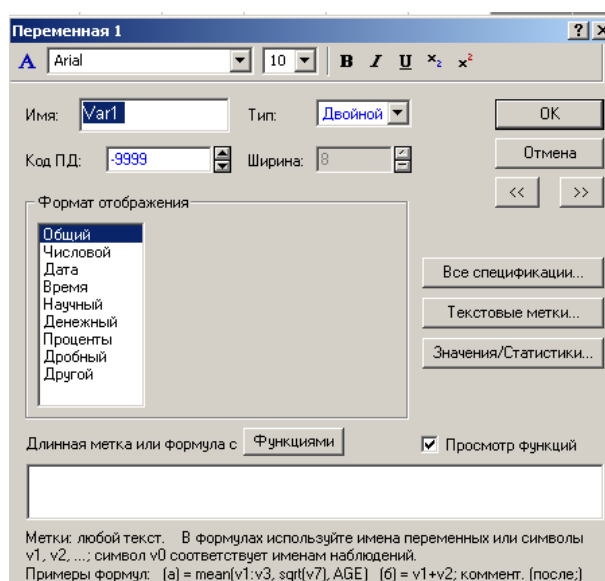


Var1 ...Var10 — переменные (названия переменных).

1...10 — значения переменных.

По мере необходимости как сами переменные, так и их значения можно удалять. Программа позволяет добавлять пустые строки или переменные, для этого используется меню вызываемое правой кнопкой мыши.

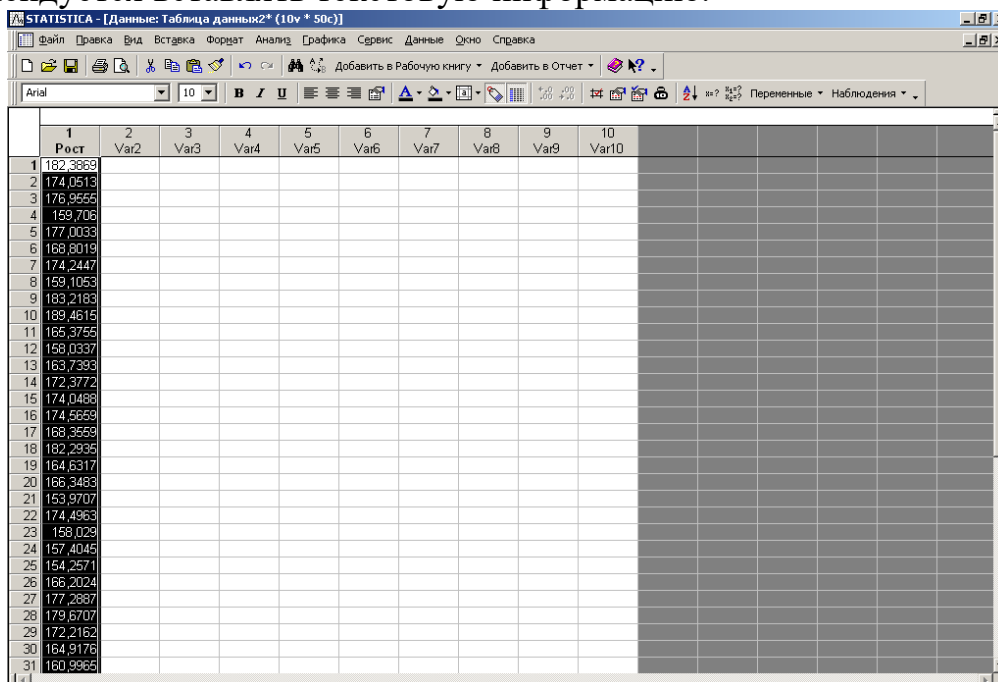
Двойным щелчком мыши по ячейке «Var1» вызывается окно редактирования свойств переменной, где можно поменять ее имя, тип, формат и т. д.



✓Задание

1. Измените название переменной «Var1» на слово «Рост» и выберите формат *числовой*. Ок.

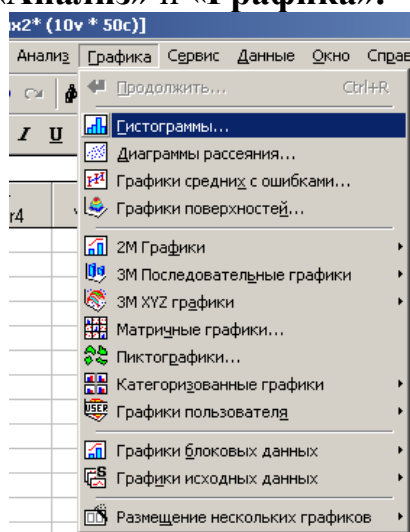
2. Скопируйте данные из предыдущей части работы, полученные при помощи генерации случайных чисел. И вставьте в столбец «Рост» программы «Statistica». Важно помнить, что программа «Statistica» работает с цифрами, поэтому в поле значений переменной не рекомендуется вставлять текстовую информацию.



	1 Рост	2 Var2	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8	9 Var9	10 Var10
1	182,3869									
2	174,0513									
3	176,9555									
4	159,706									
5	177,0033									
6	168,8019									
7	174,2447									
8	159,1053									
9	183,2183									
10	169,4615									
11	165,3755									
12	158,0337									
13	163,7393									
14	172,3772									
15	174,0488									
16	174,5659									
17	168,3559									
18	182,2935									
19	164,6317									
20	166,3483									
21	153,9707									
22	174,4863									
23	158,029									
24	157,4045									
25	154,2571									
26	166,2024									
27	177,2887									
28	179,6707									
29	172,2162									
30	164,9176									
31	180,9365									

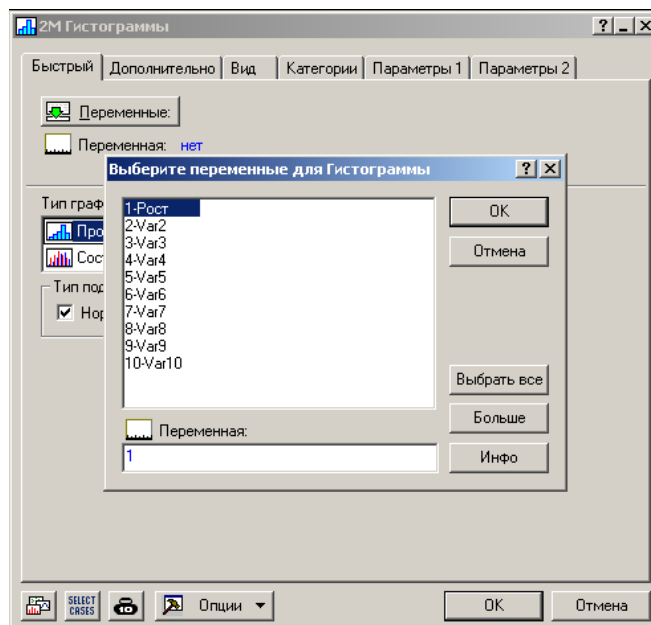
- ✓ Сколько переменных в данных эксперимента, которые вы скопировали в программу «Statistica»?
- ✓ Сколько значений этих переменных вы скопировали?

Методы анализа и инструменты для построения графиков находятся в отделах верхнего меню «Анализ» и «Графика».



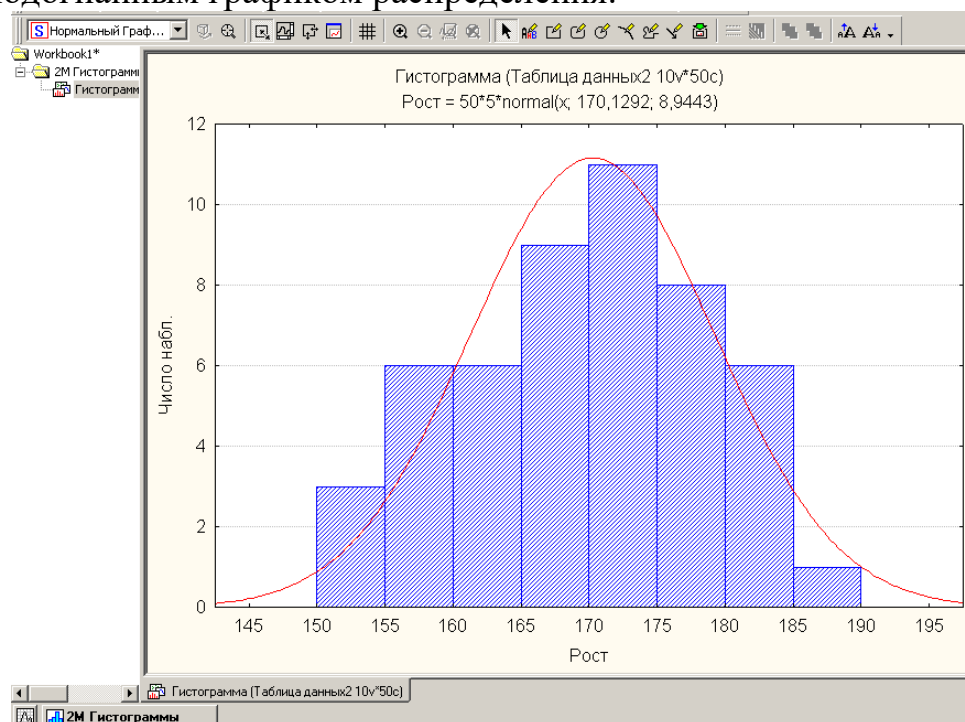
3. Для построения графика (гистограммы) необходимо вызвать меню «Графика–Гистограммы».

В появившемся окне во вкладке «Быстрый» выбрать переменную, нажав кнопку «Переменные»: и указав переменную «Рост» (Var1), нажмите «ОК».



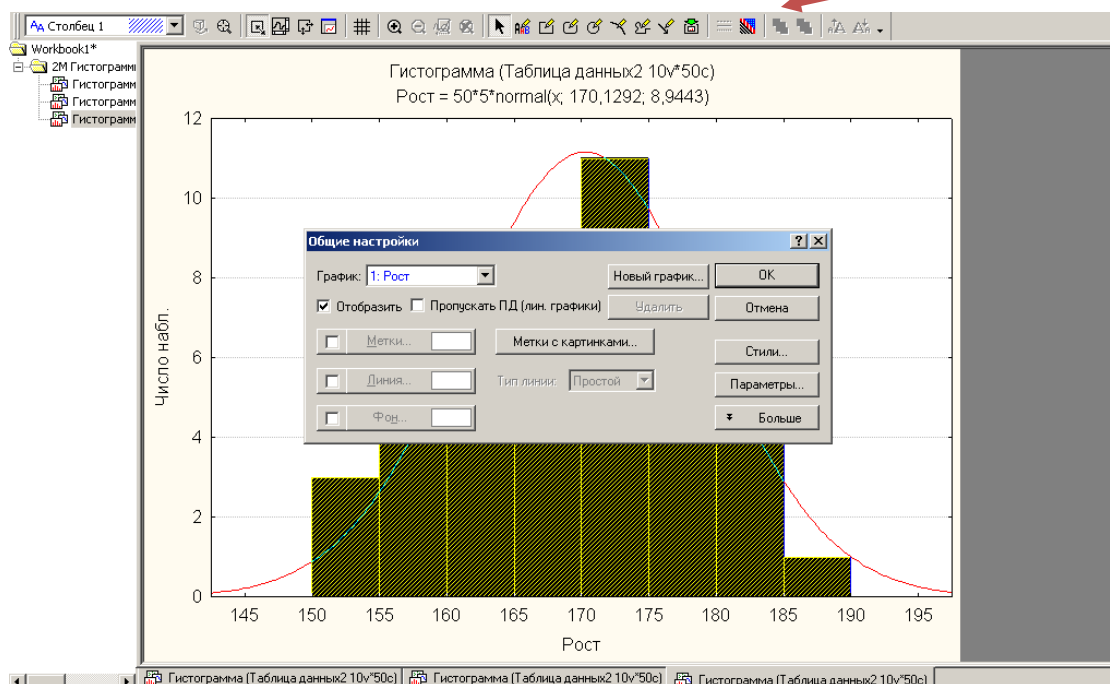
Остальные вкладки служат для более детальной настройки вывода графика.

4. При нажатии «Ок» появляется окно с изображением полигона частот и подогнанным графиком распределения.



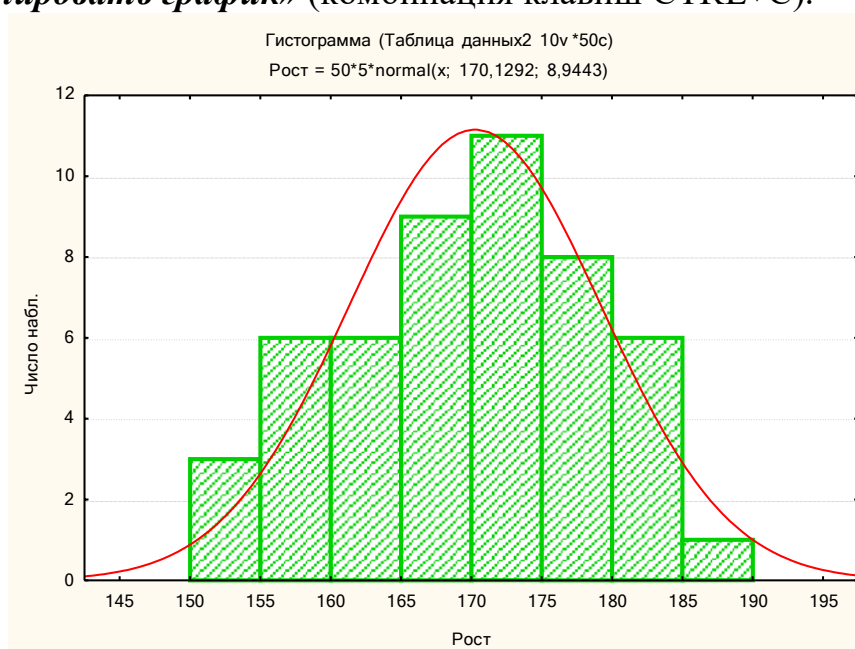
На вертикальной оси нанесены значения числа наблюдений, по горизонтальной — значение переменной. Сравните графики, полученные в MS Excel и «Statistica» 6.

5. Щелкнув правой кнопкой мыши по столбцам диаграммы можно вызвать меню свойств графика и его параметров. Где по желанию можно изменить цвет линии, фон и другие параметры внешнего вида диаграммы. Те же операции можно осуществить, нажав на иконку «**Шаблон/Цвет фона**» или «**Шаблон/Цвет линии**», предварительно выделив соответствующий объект.



6. Скопируйте полученный график в созданный вами документ Word.

Что бы скопировать график в другую программу необходимо щелкнуть правой кнопкой мыши по полю графика в появившемся меню выбрать «**Копировать график**» (комбинация клавиш CTRL+C).



- ✓График какого распределения вы получили?
- ✓Укажите интервалы карманов на графике.
- ✓Где на графике отмечены значения переменных?
- ✓Где отмечена частота их встречаемости?

Контрольные вопросы

1. Какие типы данных вы знаете?
2. Приведите примеры количественных данных.
3. Приведите примеры качественных данных.
4. Приведите примеры порядковых данных. Какая особенность данного типа данных и отличие от качественных данных.
5. Что такое генеральная совокупность?
6. Что такое выборка?
7. Почему исследователь чаще всего вынужден использовать выборку?
8. Что означает репрезентативность выборки?
9. Что означает случайный характер выборки?
10. Почему важна рандомизация выборки?
11. Что такое частота встречаемости?
12. Что показывает относительная частота встречаемости?
13. Что такое распределение значений признака?
14. Виды графического представления распределения значений признака.
15. Какие типы распределения встречаются наиболее часто?
16. На что указывает равномерное распределение?
17. На что влияет тип распределения?
18. Сколько переменных использовалось в ходе работы?
19. Сколько значений переменной было сгенерировано генератором случайных чисел?
20. Что такое карманы при построении полигона частот?
21. Что означает линия тренда полигона частот в программе MS Excel?
22. Изобразите график нормального распределения, используя столбчатую диаграмму и с использованием линейного отображения.

Лабораторная работа № 2

Описательная статистика. Построение графиков распределения Краткие сведения из теории

В результате проведения эксперимента исследователь получает данные, которые необходимо обработать для дальнейшего формирования выводов и заключений.

Данные, которые необходимо подвергнуть статистическому анализу, чаще всего представлены большим массивом чисел, показателей или другими возможными значениями проявления признака. Например, при исследовании влияния анестетика на падение артериального давления при операции на открытом сердце исследователь получает таблицу результатов, в которой перечислены значения давления у каждого пациента выборки (например, до и во время проведения операции), выживаемость после операции (умер пациент или нет) и т. п. Так как работать с большим массивом данных сложно и неудобно, его стремятся представить в более приемлемом и наглядном виде для дальнейшего анализа.

Одним из первых этапов статистического анализа является **краткое описание данных**, или **описательная статистика**.

Описательная статистика включает в себя:

- Формирование таблиц результатов анализа (строго говоря, это предварительный этап).
- Проверка данных на возможное наличие артефактов (выбросов).
- Построение графика распределения (полигона частот) значений признака.
- Расчет основных параметров распределения.
- Формирование выводов относительно полученных данных эксперимента о принадлежности их к тому или иному типу распределения и как следствие — выбор метода дальнейшего анализа.

Вариационный ряд. Типы распределения значений признака

Повторим существенные моменты из предыдущей лабораторной работы. Изобразить распределение признака можно различными способами: *вариационным рядом, гистограммой, вариационной кривой*.

Вариационный ряд — это упорядоченное отражение реально существующего распределения значений признака по отдельным особям изученной группы.

Другими словами, при измерении значений какого-то признака (например, температуры тела, роста или артериального давления), у разных членов исследуемой группы значения этого признака будут различными. Упорядоченная запись этих значений и называется **вариационным рядом**.

Например: в результате исследования группы людей на предмет влияния правильности метода лечения на сроки госпитализации (где перемен-

ной является *число дней госпитализации*) был получен следующий вариационный ряд:

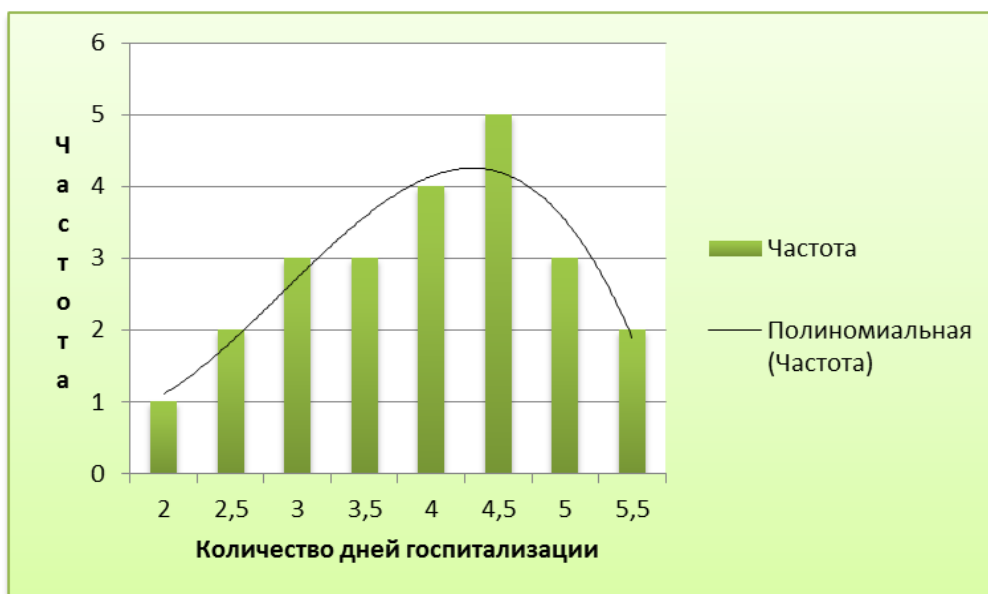
Количество дней госпитализации	2	2,5	2,5	3	3	3	3,5	3,5	3,5	4	4	4	4	4,5	4,5	4,5	4,5	4,5	5	5	5	5,5	5,5
--------------------------------------	---	-----	-----	---	---	---	-----	-----	-----	---	---	---	---	-----	-----	-----	-----	-----	---	---	---	-----	-----

Вариационный ряд можно изобразить и в виде таблицы с указанием значения переменной и частотой возникновения этого значения (или нескольких значений попадающих в определенный интервал (карман)) в ходе эксперимента:

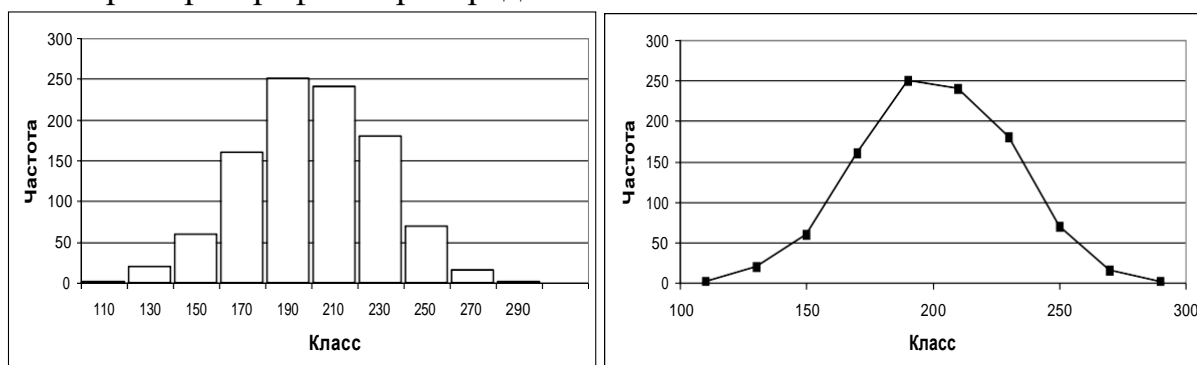
Количество дней	Частота
2	1
2,5	2
3	3
3,5	3
4	4
4,5	5
5	3
5,5	2

Каждая *генеральная совокупность* (или *выборка*) характеризуется **распределением** значений исследуемой переменной (признака) или графическим представлением **частоты встречаемости**. Т. е. графическим представлением того, с какой частотой встречается в результатах эксперимента то или иное значение переменной.

Для построения графика распределения для приведенного выше вариационного ряда на горизонтальной оси отмечаются значения «Количество дней госпитализации», на вертикальной оси — отчается сколько раз то или иное значение (дней госпитализации) появилось в ходе исследования.



Примеры графиков распределения частот:



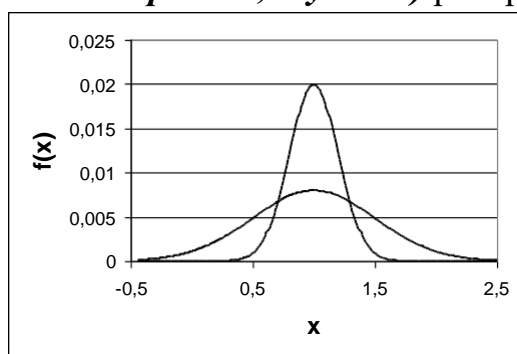
Столбчатую диаграмму чаще всего называют полигоном частот, или гистограммой, огибающую линию — графиком распределения частот, или вариационной кривой.

✓Что показывает относительная частота встречаемости?

✓В чем она может выражаться?

Наиболее часто встречаются следующие **виды распределения**:

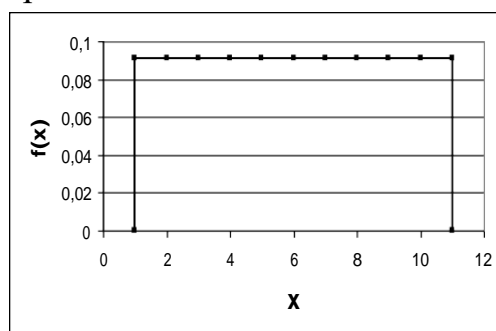
Нормальное (колоколообразное, гауссово) распределение.



Нормальное распределение подразумевает, что большая часть значений признака находится в районе так называемого среднего значения (на графике это значение часто обозначается греческой буквой мю (μ)).

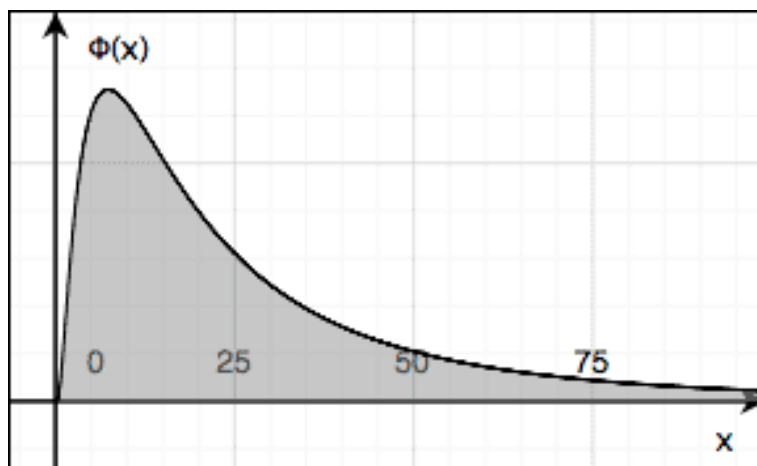
Другими словами, если имеет место нормальное распределение признака, то наиболее часто в выборке встречаются значения близкие по величине к среднему значению по выборке, и расположены они симметрично относительно среднего значения.

Равномерное распределение



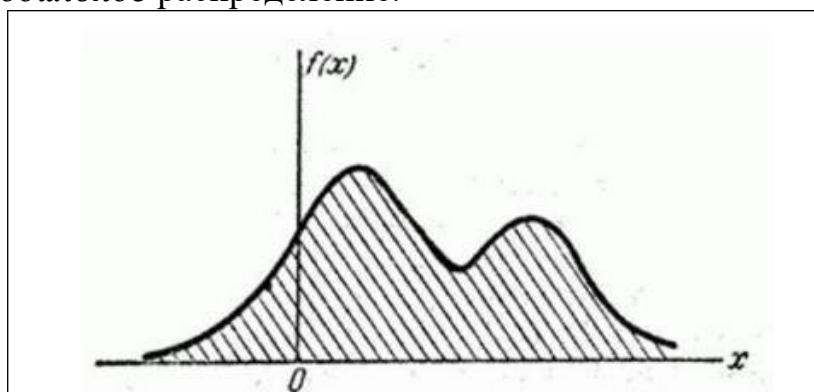
Равномерное распределение указывает на малое влияние переменной на исследуемый процесс или малое влияние процесса на показатели.

Асимметричное распределение (если асимметрия левосторонняя — **логнормальное распределение**).



Если функцию $f(x)$ логнормального распределения преобразовать на ее логарифм $\log(f(x))$, то в этом случае полученная функция будет иметь нормальное распределение и характеризоваться теми же параметрами.

Полимодальное распределение.



Полимодальное распределение может быть обусловлено действием нескольких скрытых факторов. Или о, возможно, неправильном построении исследования, например, выборка не является достаточно репрезентативной.

В зависимости от типа распределения выбираются соответствующие ему методы статистического анализа.

Основные параметры распределения

Если распределение является *нормальным*, то применяют методы так называемой **параметрической статистики**. При использовании методов параметрической статистики выборка практически исчерпывающе характеризуется параметрами: **средним значением (математическое ожидание)**, **дисперсией и стандартным отклонением (среднеквадратичное отклонение)**.

Среднее значение определяется формулой

$$\mu = \frac{\sum X}{N}.$$

Т. е. отношение суммы значений всех переменных к их количеству (N — количество для совокупности, n — для выборки, X — значение переменной).

Если распределение значений близко к нормальному, то большинство значений распределено возле *среднего значения*. Величина, характеризующая расброс значений от среднего называется **дисперсией**.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}.$$

Дисперсия — средний квадрат отклонения значений выборки от среднего по выборке. Т. к. оперировать квадратом размерности величины не удобно (например, если варьируемая величина имеет размерность *см*, то дисперсия измеряется в *см²*), на практике чаще используют корень квадратный от дисперсии называемый **стандартным отклонением**.

Дисперсия характеризует разброс значений относительно среднего значения по выборке.

Стандартное отклонение также характеризует разброс значений, но измеряется в той же размерности, что и сами значения (в случае с распределением роста — *сантиметры* или *метры*).

Другими словами, каждое значение признака может отличаться от среднего значения, причем либо на большую величину (большая разность по модулю), либо на меньшую (меньшая разность по модулю).



Стандартное (среднеквадратичное) отклонение характеризует эти различия по всей выборке и выражает их одним числом, что дает достаточное представление о **среднем разбросе значений от среднего**.

Для генеральной совокупности стандартное отклонение вычисляется по формуле:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X - \mu)^2}{N}}.$$

Для выборки формула имеет вид:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}.$$

Пример:

Стандартное отклонение — важный статистический показатель, но когда сообщаются статистические результаты, о нем часто забывают. Без этого показателя вы видите только часть информации относительно данных. Статистики часто приводят в пример историю о человеке, одной ногой стоящем в ведре с ледяной водой, а второй — в ведре с кипятком. В среднем несчастный должен чувствовать себя отлично! Но вспомните о разнице двух температур для каждой его ноги.

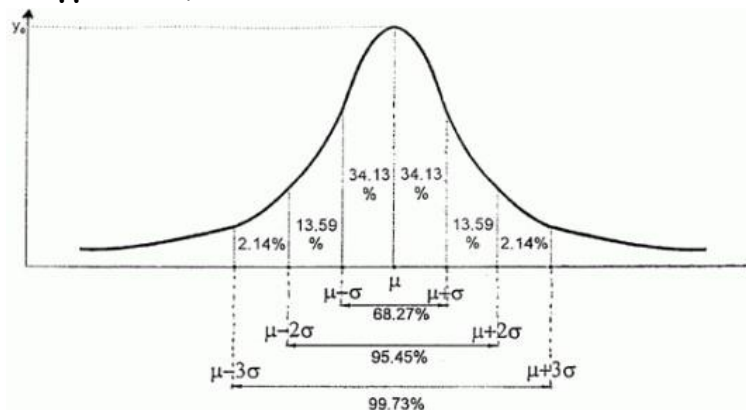
Другой пример. Средняя зарплата может не в полной мере отражать реальное положение дел в компании, если разброс окладов очень большой. Кто-то ест мясо, а кто-то капусту, в среднем вместе едят голубцы.

Важность замечаний в следующем: нельзя полагаться только на знание среднего значения без учета величины стандартного отклонения. Представления об объекте могут быть представлены в искаженном виде.

Нормальное распределение полностью характеризуется средним значением μ и стандартным отклонением σ .

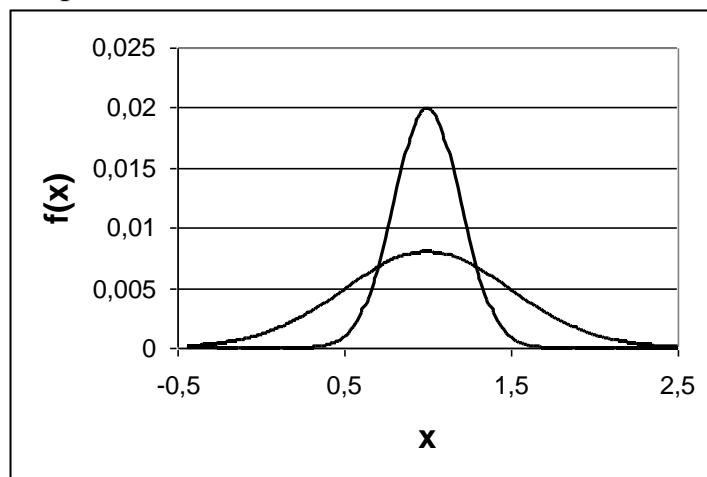
Правило трех сигм

Правило трёх сигм (трех стандартных отклонений σ) — практически все значения **нормально распределённой** случайной величины лежат в интервале от -3σ до $+3\sigma$. Более строго — приблизительно с 0,9973 вероятностью значение нормально распределённой случайной величины лежит в интервале от -3σ до $+3\sigma$.



В случае нормального распределения 68 % наблюдаемых значений отклоняются от среднего значения μ не более чем на величину стандартного отклонения σ , 95 % значений не выйдут из пределов $\mu \pm 2\sigma$ и практически все значения уместятся в пределы $\mu \pm 3\sigma$. Вероятность отклонения за пределы 3σ равна $0,0026 \approx 0,003$, т. е. такое событие наступит только в среднем в 3 случаях из 1000 испытаний.

С помощью графика нормального распределения можно представить зависимость ширины «колокола» от стандартного отклонения: *чем больше дисперсия (или стандартное отклонение), тем шире «колокол», т. е. разброс значений признака больше.*

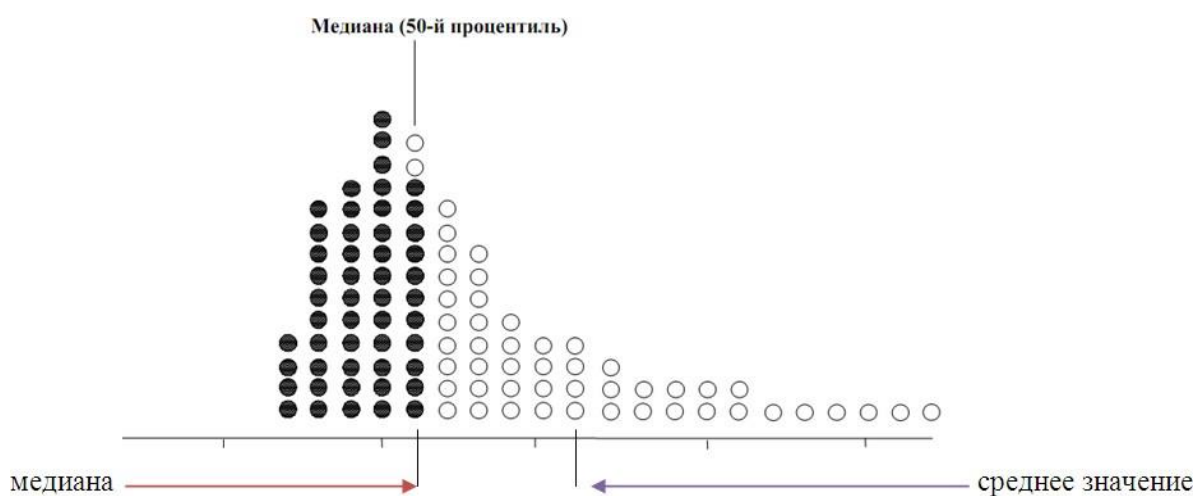


Медиана, мода, процентиля

Если же распределение отлично от нормального (например, значения распределены несимметрично относительно среднего) параметры «Среднее значение» и «Дисперсия» не являются информативными и могут ввести в заблуждение как самого исследователя, так и читателя результатов исследования. В случае другого типа распределения обычно совокупность описывается с помощью **моды** и **медианы**, а также **процентилей**.

Медиана — значение, которое делит распределение пополам, в результате справа и слева от него находится равное число значений.

На иллюстрации видно, что распределение отлично от нормального и имеет левостороннюю асимметрию.



Мода — наиболее часто встречающееся значение.

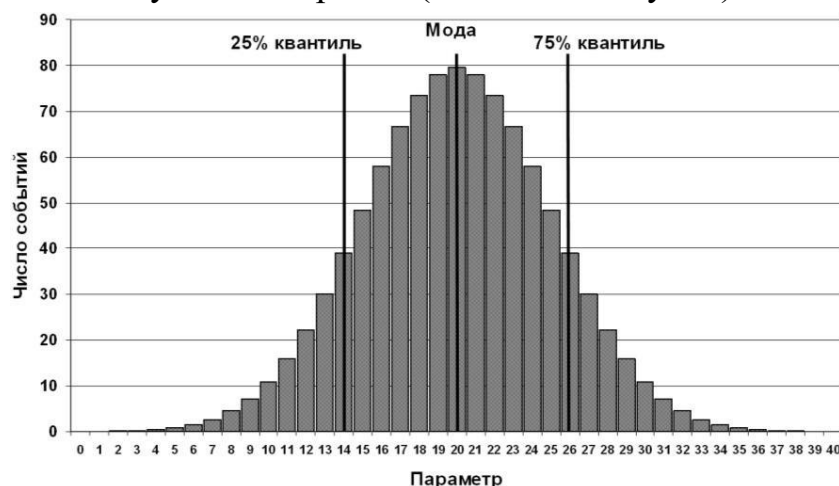
Иногда весь диапазон значений разбивают на четыре интервала — **процентили (квартили)**.



В природе наиболее часто встречается нормальное распределение. Однако в медицинских исследованиях так бывает не всегда. Очень часто речь идет о том, что исследователь не может однозначно сказать, что распределение является нормальным, этому могут быть несколько причин, например, недостаточное количество данных полученных в ходе эксперимента.

Для распределения, не являющегося нормальным, параметрические методы неприменимы (выборка не может характеризоваться параметрами «Среднее значение» и «Стандартное отклонение»), их использование может привести к серьезным ошибкам в выводах об исследуемой совокупности. В таких случаях разумнее воспользоваться **непараметрическими** или **ранговыми** методами, которые можно применять для любых распределений.

В случае с нормальным распределением значение медианы, моды и среднего значения обычно близки по своей величине. На графике это можно представить следующим образом (идеальный случай):



Как видно из графика, *медиана* и *среднее значение* равны 20, таким же по величине является и наиболее часто встречаемое в совокупности значение признака — *мода*.

Артефакты

Артефакты (или **выбросы**) — такие записанные значения признака, которые резко отличаются от всех других значений признака в группе.

Проверка артефактов должна проводиться всегда перед началом обработки полученных первичных данных. Если подтвердится, что *резко выделяющееся* значение (например, записанное значение *роста 263 см*) действительно не может относиться к объектам данной группы, и попало в записи вследствие ошибок внимания, следует такой артефакт исключить из обработки. Проверка артефактов может производиться по критерию, равному *нормированному отклонению выппада*.

Проверка выбросов может производиться по критерию, равному **нормированному отклонению выброса**:

$$T = \frac{X_i - \mu}{\sigma} \geq T_{st},$$

где: T — критерий выброса; X_i — выделяющееся значение признака (или очень большое или очень малое); μ, σ — средняя и стандартное отклонение, рассчитанные для группы, включающей артефакт; T_{st} — стандартные значения критерия выбросов, определяемых по таблице 1.

Таблица 1 — Стандартные значения критерия выбросов (T_{st})

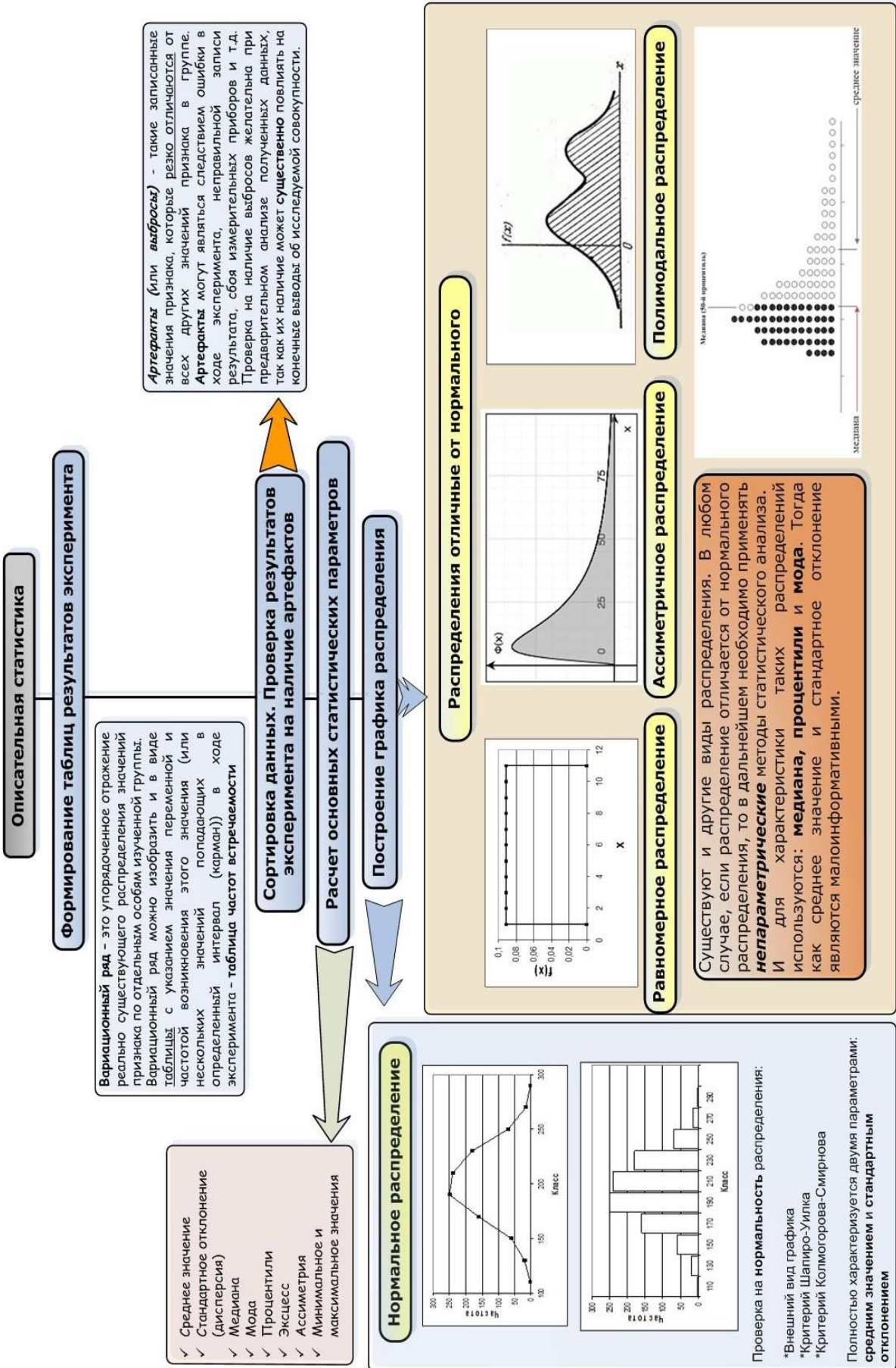
n	T_{st}	n	T_{st}	n	T_{st}	n	T_{st}
2	2,0	16–20	2,4	47–66	2,8	125–174	3,2
3–4	2,1	21–28	2,5	67–84	2,9	175–349	3,3
5–9	2,2	29–34	2,6	85–104	3,0	350–599	3,4
10–15	2,3	35–46	2,7	105–124	3,1	600–1500	3,5

Если $T \geq T_{st}$, то анализируемое значение признака является выбросом. Альтернатива $T < T_{st}$ не позволяет исключить из анализа значение признака.

Артефакты могут являться следствием ошибки в ходе эксперимента, неправильной записи результата, сбоя измерительных приборов и т. д. Проверка на наличие выбросов желательна при предварительном анализе полученных данных, так как их наличие может **существенно** повлиять на конечные выводы об исследуемой совокупности. В принципе, если исследователь знает границы возможных результатов и какие-то полученные значения сильно выбиваются из этих границ, он может исключить их из анализа, не проводя дополнительной проверки вышеописанным способом.

Стоит заметить, что иногда исследователя могут интересовать и сами выбросы как периодически возникающие аномальные явления, конечно, если они не являются следствием ошибки.

Ниже представлена краткая общая схема проведения процедуры описательной статистики:



Выполнение процедуры описательной статистики в MS Excel

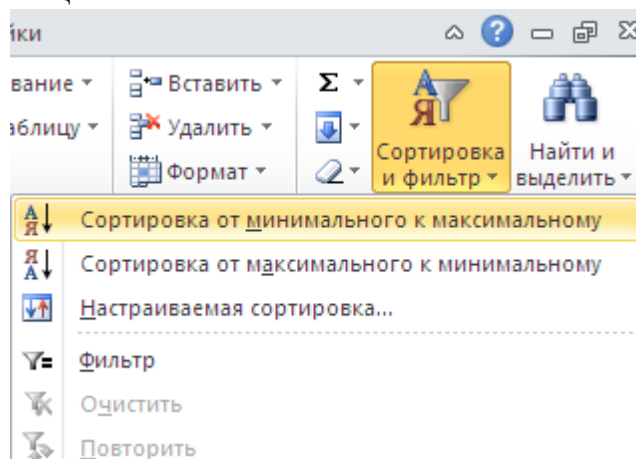
✓ Задача

Пусть целью вымышленного исследования является получение более детальной информации о населении открытой исследователями новой ранее неизведанной и неизученной народности. Первым делом ученые решили получить больше информации об антропологических данных, проживающих на территории людей. Одним из изучаемых признаков был выбран рост мужчин определенной возрастной группы (от 18 до 55 лет). Данные представлены в виде таблицы:

Рост
174
164
179
190
189
194
155
175
188
166
170
160
159
167
169
156
171
173
178
173
174
173
190
176
175
172
197
186
201
170
280
110

Необходимо описать данные и построить график распределения частот.

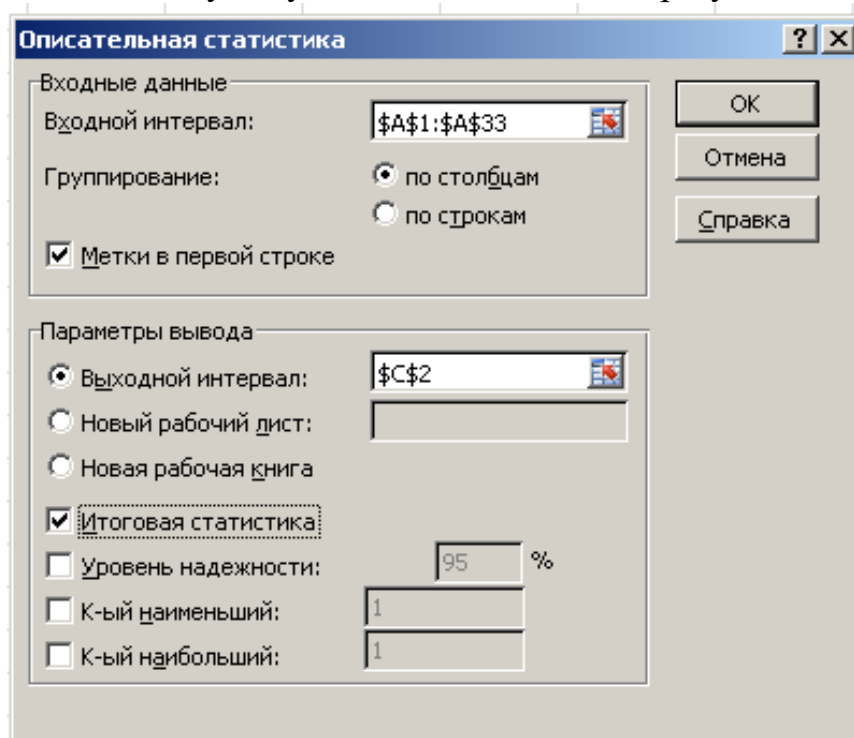
1. Скопируйте таблицу в книгу MS Excel. Первой подготовительной процедурой является сортировка исходных данных. Для этого выделите анализируемый массив, а затем в меню выберите «Сортировка от минимального к максимальному» или «Сортировка от максимального к минимальному» и щелкните левой кнопкой мыши.



Для статистического анализа данных в программе Excel воспользуйтесь «Пакетом анализа».

2. В модуле «Анализ данных» выберите «Описательная статистика», после чего щелкните мышкой «ОК».

3. В появившемся окне выполните установки, как показано на рисунке (первая строка в исходных данных (заглавие таблицы — *Рост*) — «Метка в первой строке»). «Входной интервал» — вся таблица (со словом *Рост*, при этом необходимо поставить маркер *Метки в первой строке*). «Выходной интервал» — ячейка, куда будет вставлена таблица результатов.



Полученные результаты представлены в виде таблицы.

Рост	
Среднее	176,7453182
Стандартная ошибка	4,40513267
Медиана	173,5375821
Мода	#Н/Д
Стандартное отклонение	24,91919346
Дисперсия выборки	620,9662029
Эксцесс	10,38020962
Асимметричность	1,788193469
Интервал	170
Минимум	110
Максимум	280
Сумма	5655,850183
Счет	32

Примечание: эта опция позволяет обрабатывать любое количество выборок одновременно.

4. Следующей процедурой является проверка данных на наличие выбросов. Из результатов обработки, представленных на рисунке, обращают на себя внимание высокие значения **эксцесса** и **асимметрии**. Можно предположить, что крайние значения (минимальные или максимальные) являются *выбросами*. Это значения 110 и 280 см. Проверяем эти значения по формуле, приведенной выше, и при помощи **таблицы 1**. Если значения являются артефактами, то они исключаются из выборки.

После исключения артефактов из таблицы результатов исследования повторяются *шаги 1-3*, но уже без учета значений признанных артефактами (значения признака, признанные артефактами, удаляются из таблицы).

5. Построение графика распределения.

Величина карманов. Для построения гистограммы необходимо определить величину класса (ширину кармана) по формуле:

$$k = \frac{X_{\max} - X_{\min}}{n},$$

где n — количество интервалов; X_{\max} , X_{\min} — максимальное и минимальное значение признака совокупности.

$$n = 1 + 3,322 \cdot \lg N,$$

где N — число наблюдений.

Количество интервалов n округляется до ближайшего целого вверх.

После вычисления ширины кармана определяется, какое количество значений признака попало в тот или иной карман, для упрощения выполнения данной операции можно воспользоваться функцией «ЧАСТОТА» программы MS Excel.

Если не рассчитывать размеры интервалов (величину карманов), то они будут определены программой автоматически и данный шаг можно пропустить

Если не вдаваться в столь строгое следование правилам расчета интервалов карманов, то можно задать их самостоятельно, начальной цифрой может являться минимальное значение по выборке, конечной — максимальное, например, в данной задаче это может выглядеть следующим образом:

Карманы
150
160
170
180
190
200
210

Создайте данную таблицу на том же листе MS Excel, и в поле «Интервал карманов» можно указывать ваши карманы, записав соответствующий диапазон ячеек.

6. Модуль »Анализ данных» и выберите опцию «Гистограмма», после чего щелкните мышкой «ОК».

7. Выполните установки как показано на рисунке (при необходимости укажите диапазон значений карманов, соответствующий таблице выше).

Гистограмма

Входные данные

Входной интервал:

Интервал карманов:

☒ Метки

Параметры вывода

☒ Выходной интервал:

☐ Новый рабочий лист:

☐ Новая рабочая книга

☐ Парето (отсортированная гистограмма)

☐ Интегральный процент

☒ Вывод графика

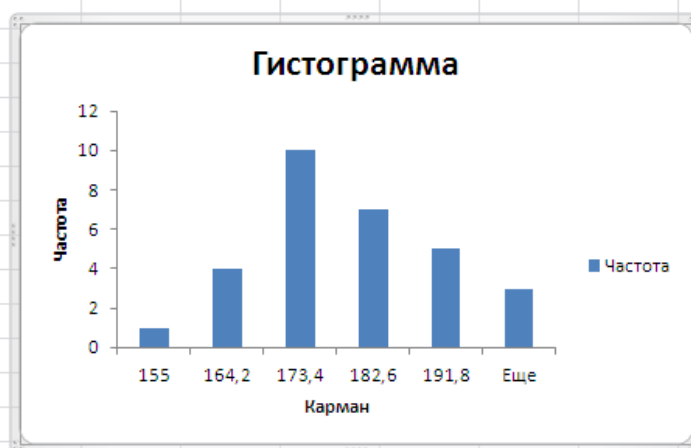
ОК

Отмена

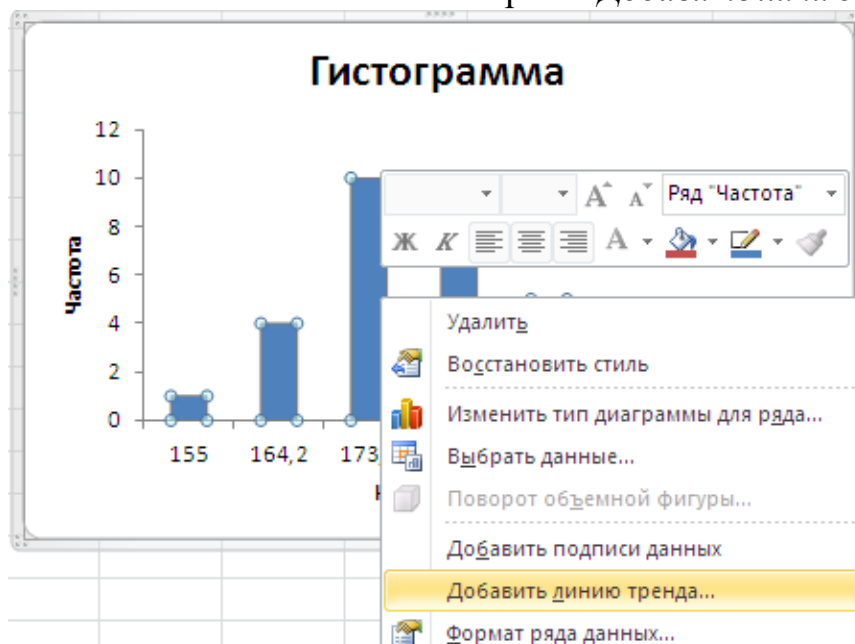
Справка

Результат обработки появится в указанной ячейке.

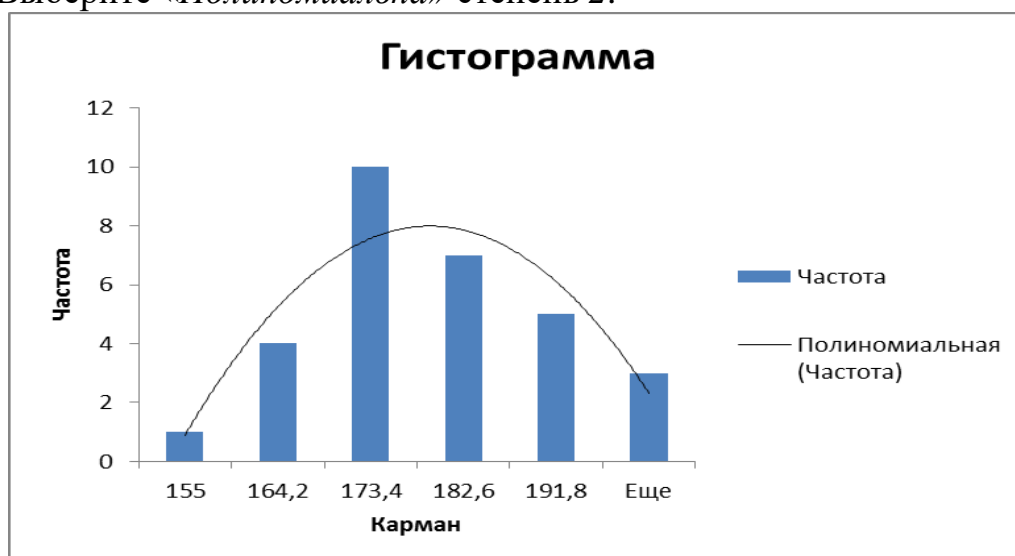
Карман	Частота
155	1
164,2	4
173,4	10
182,6	7
191,8	5
Еще	3



8. Построение линии тренда. Щелкните по столбцам диаграммы правой кнопкой мыши и в появившемся меню выберите «Добавить линию тренда».



Выберите «Полиномиальная» степень 2.



Исправьте названия осей на «Рост» (ось X) и «Частота встречаемости» (ось Y).

Оформите график. Название оси «Карман» переименуйте в название измеряемого параметра — «Рост». Уберите значение «Еще» и замените его на соответствующее вашей выборке, если вы использовали таблицу карманов, описанную выше и при этом появилось значение «Еще» и напротив него в поле «Частота» число 0, удалите обе ячейки.

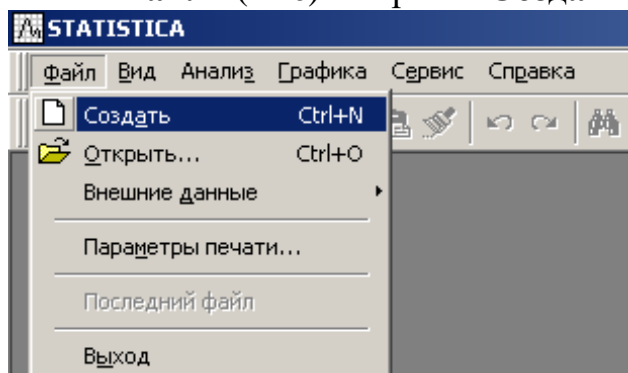
Порядок анализа данных в ПП «Statistica» 6

Подготовительные процедуры

Процедуры, связанные с сортировкой массивов данных и поисками выбросов выполняются в табличном редакторе Microsoft Excel.

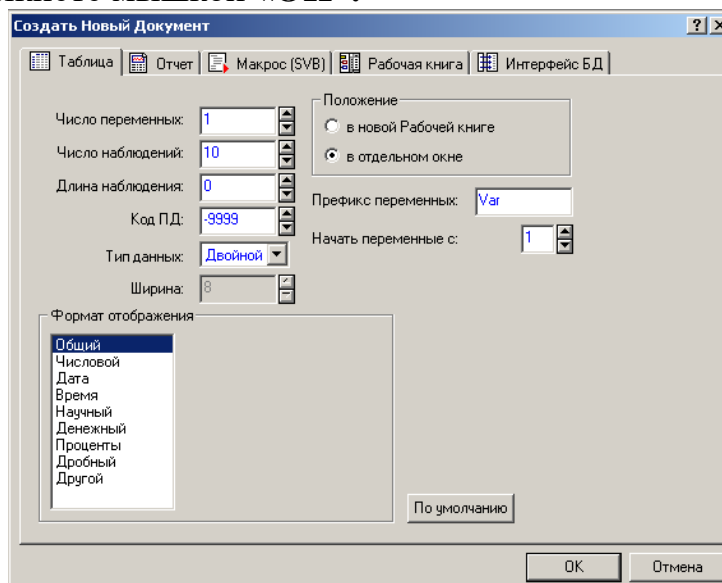
1. Запустите программный продукт «Statistica» 6.

Закройте появившиеся при запуске окна. Сформируйте таблицу исходных данных: в окне «Файл» (File) выбрать «Создать» (New).



Открытие таблицы

2. В появившемся окне задайте число строк (**Число переменных/Number of cases**) и столбцов (**Число наблюдений/Number of variables**). Щелкните мышкой «ОК».



Формирование таблицы

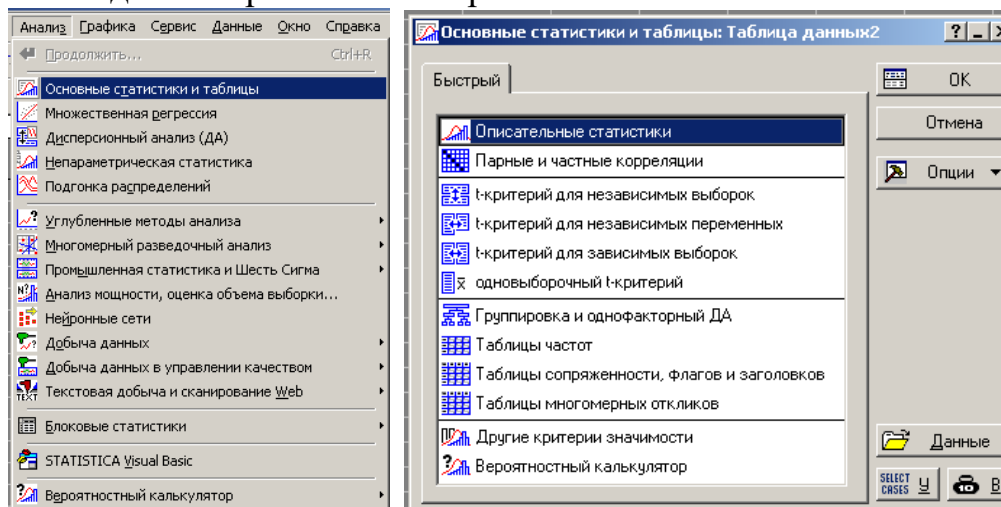
3. Теперь в таблицу необходимо внести исходные данные, осуществив набор непосредственно или вставку копии из файла, например, табличного редактора **MS Excel**. Измените название переменной на «Рост».

	1
	Рост
1	174
2	164
3	179
4	190
5	189
6	194
7	155
8	175
9	188
10	166
11	170
12	160
13	159
14	167
15	169
16	156
17	171
18	173
19	178
20	173
21	174
22	173
23	190
24	176
25	175
26	172
27	197
28	186
29	201
30	170

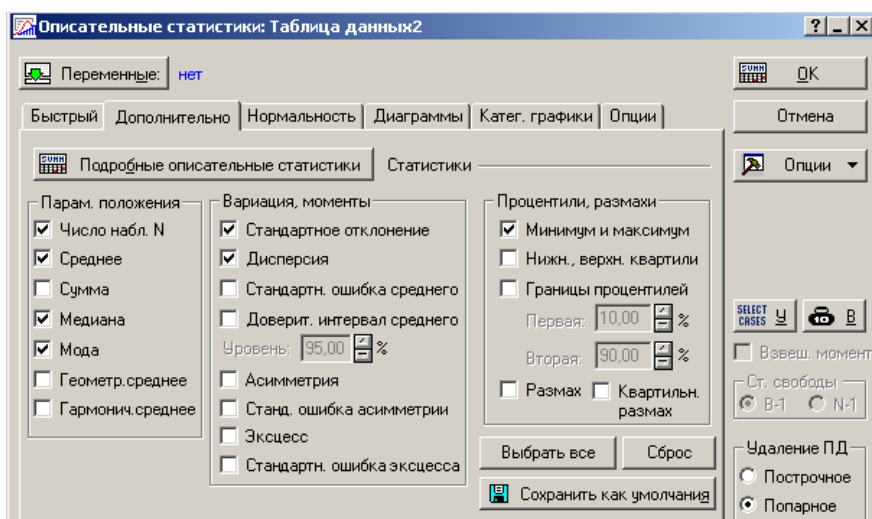
Первичный анализ статистических данных (описательная статистика)

4. В меню выберите «Анализ» (**Statistics**) и запустите модуль «Основные статистики» и таблицы (**Basic Statistics/Tables**). Высветите в стартовой панели модуля (**Basic Statistics/Tables**) «Основные статистики/таблицы» строку «Описательная статистика» (**Descriptive statistics**).

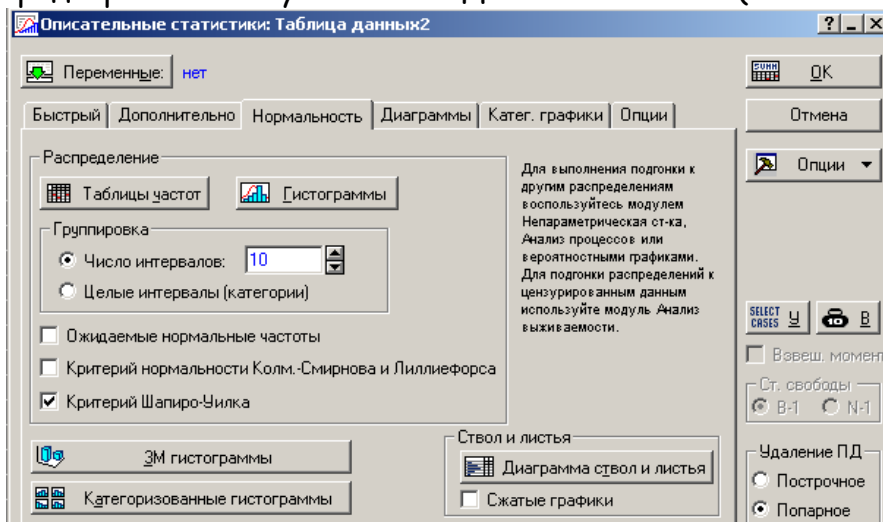
5. Нажмите кнопку «**OK**». Перед вами откроется окно «**Описательная статистика**» (**Descriptive statistics**). Выполните установки, как показано на рисунках. В окне «**Нормальность**» установите необходимое **число интервалов** для построения гистограмм.



Запуск модуля «Описательная статистика» (Descriptive statistics)

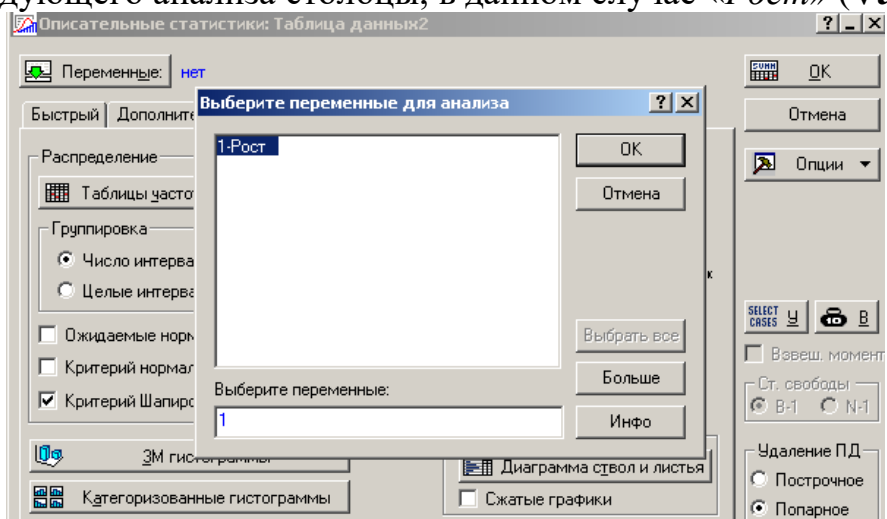


Предварительные установки «Дополнительно» (Advanced)



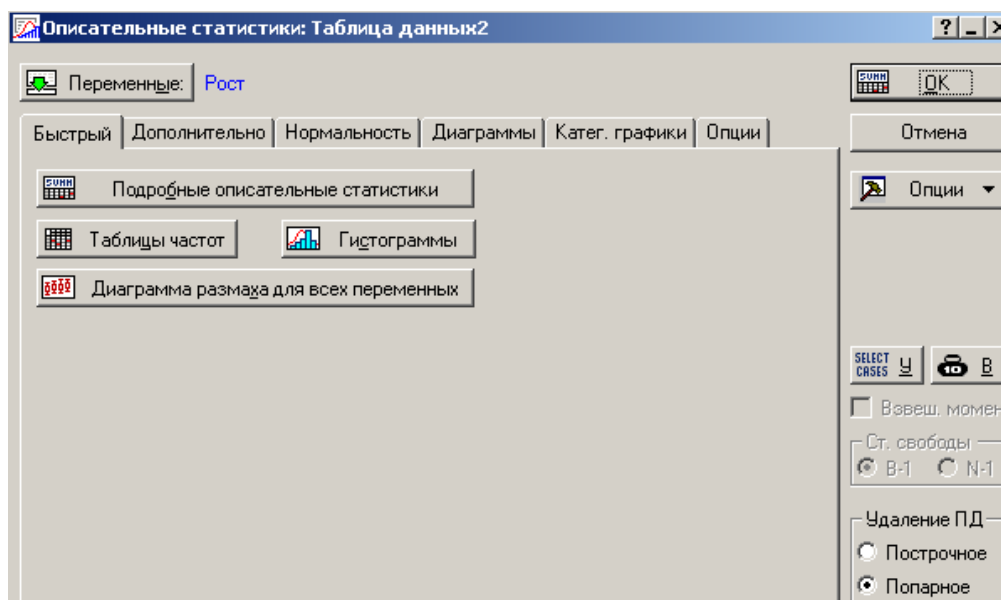
Предварительные установки Нормальность (Normality)

6. Загрузите в систему «**Statistica**» исходные данные. Для этого левой кнопкой щелкните по клавише «**Variables**» (**Переменные**). В появившемся окне, нажав и не отпуская левую кнопку мышки, пометьте необходимые для последующего анализа столбцы, в данном случае «**Рост**» (**Var1**).



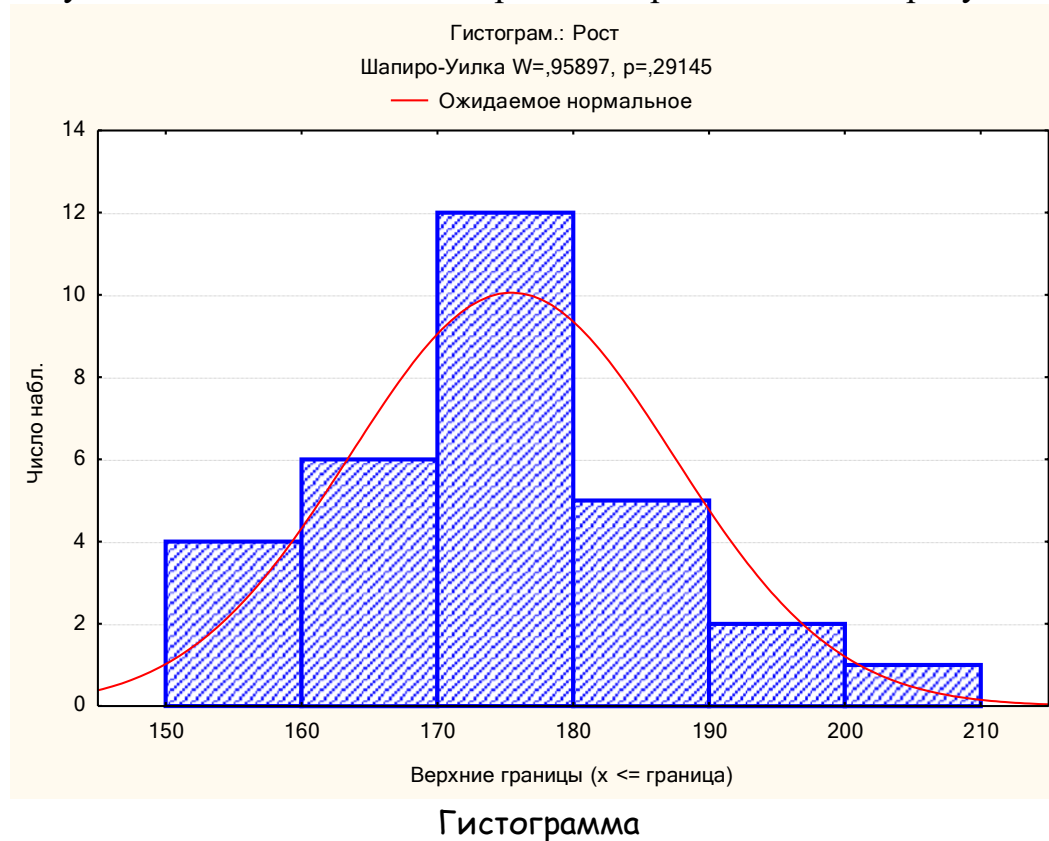
7. В диалоговом окне результатов **Быстрый** (Quick) последовательно нажмите кнопки:

- «Гистограммы» (Histograms).
- «Подробные описательные статистики» (Summary Descriptive statistics).
- «Таблица частот» (Frequency Tables).

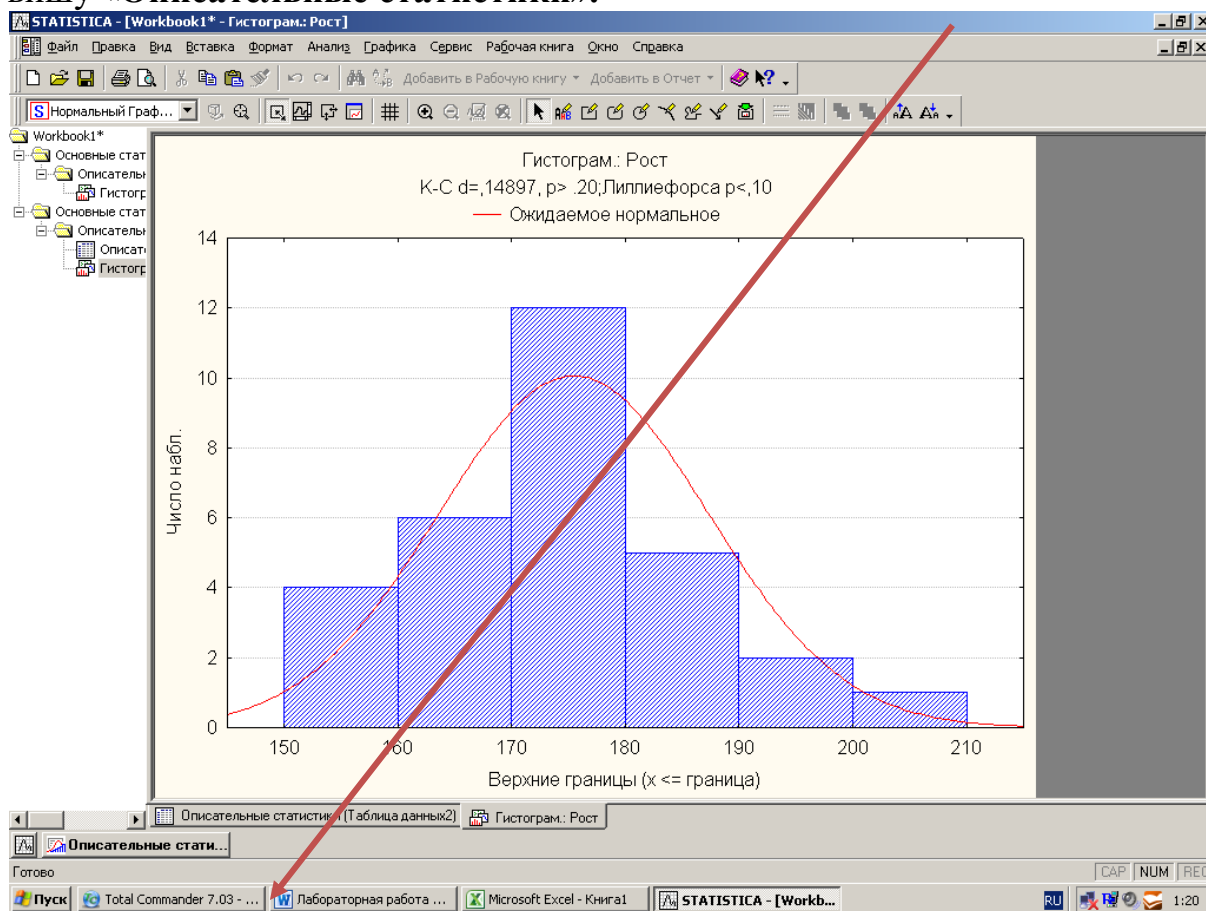


Диалоговое окно результатов «Быстрый» (Quick)

Результаты статистической обработки представлены на рисунках:



После вывода графика скопируйте его в документ Word. Для возвращения к окну выбора операций для дальнейшего анализа нажмите на клавишу **«Описательные статистики»**.



Подробные описательные статистики:

	N набл.	Среднее	Медиана	Мода	Частота	Минимум	Максимум	Дисперс.	Стд.откл.
Рост	30	175,4667	173,5000	173,0000	3	155,0000	201,0000	141,7057	11,90402

Описательная статистика

Чтобы скопировать таблицу результатов анализа можно щелкнув правой кнопкой мыши по выделенным значениям в таблице и выбрать **«Копировать с заголовками»**.

Проверка распределения на нормальность

Для проверки распределения на нормальность выберите **критерий Шапиро—илка**, если полученное значение p (см. график) больше чем 0,05 (критическое значение), то распределение является нормальным, если же меньше — нормальность распределения можно подвергнуть сомнению. В нашем случае вероятность $p=0,29$, что больше чем 0,05, это говорит о том, что распределение приближенно можно считать нормальным. Это можно заключить и по внешнему виду графика

✓ Задание

Произвести первичный анализ данных в программных продуктах MS Excel и «Statistica». Результаты оформить в отдельный документ формата Word, в виде отчета. **Отчет должен содержать:** исходные данные, полученные таблицы результатов, графики распределения, объяснение полученных результатов, выводы о типе распределения и о том какими статистическими методами следует пользоваться при дальнейшем исследовании данной выборки.

1. Давление при операции на открытом сердце (галотановая анестезия):

Галотановая анестезия
42
44
46
46
48
48
50
50
52
54
54
56
58
58
58
60
60
60
60
60
62
62
62
62
62
64
64
66
66
66
66
66
68
68

68
68
70
70
70
70
70
72
72
72
72
74
74
74
74
76
78
78
78
80
80
82
82
84
86
90
94
98

2. Таблица распределения возраста заболевших менингитом, вызванным гемофильной палочкой.

Возраст
1
1
1
1
1
1
1
1
1
1
1
1
1
1

1
1
1
1
1
1
1
1
1
1
1
20
50
70

Контрольные вопросы

1. Что такое генеральная совокупность?
2. Что такое выборка?
3. Что означает понятие «репрезентативная выборка»?
4. Что такое распределение случайной величины?
5. Какие виды распределения вы знаете?
6. График нормального распределения.
7. Какие выводы можно сделать, если данные имеют нормальное распределение?
8. Какие методы статистического анализа обычно используются, если распределение является нормальным?
9. Чем характеризуется нормальное распределение?
10. Что характеризует дисперсия случайной величины?
11. Что характеризует стандартное отклонение?
12. Что такое среднеквадратичное отклонение?
13. Какими параметрами характеризуется распределение, когда значения признака распределены несимметрично относительно среднего значения?
14. Что такое медиана и процентиля?
15. Что такое мода?
16. Сколько переменных было в решаемых задачах?
17. Почему среднее значение является недостаточно информативным параметром, если исследователь не указывает значение дисперсии и тип распределения?
18. Какие типы распределения наиболее часто встречаются в природе?
19. При каком условии можно пользоваться методами параметрической статистики?
20. При каком условии используют методы непараметрической статистики?
21. Что такое артефакты (выбросы)?
22. Зачем необходима проверка на выбросы?
23. Правило трех сигм.

Лабораторная работа № 3

Сравнение групп

Дисперсионный анализ

Краткие сведения из теории

Часто возникает задача сравнить группы между собой. Такая задача может возникнуть, если необходимо выявить какой метод лечения более эффективный или, при изучении нового препарата, поставить эксперимент, сравнив препарат с аналогом, или исследовать эффективность препарата, поставив эксперимент с плацебо и т. д.

Обычно в подобных исследованиях одна из групп испытуемых принимается за **контрольную** — например, группа, которую лечили эталонным методом лечения или давали плацебо вместо испытуемого лекарства, другая группа — **экспериментальная**. Сравнив между собой исследуемые показатели, исследователь делает вывод согласно цели эксперимента.

Нулевая гипотеза

(др.-греч. ὑπόθεσις — предположение; от ὑπό — снизу, под + θέσις — тезис) — утверждение, предполагающее доказательство.

Нулевая гипотеза — это предположение, **что исследуемые факторы не оказывают никакого влияния на параметры и полученные различия (исследуемых параметров) случайны**. Иначе: нулевая гипотеза предполагает отсутствие эффекта или различия между группами.

В качестве **нулевой гипотезы** может выступать гипотеза обратная утверждению о том, что значения признака распределены по нормальному закону распределения, или гипотеза о том, что различия между группами статистически не значимы или случайны и т. п. Например, сравнивая различающиеся средние значения артериального давления при применении двух разных анестетиков, мы выдвигаем **нулевую гипотезу** о том, что полученные различия не значительны или случайны, и анестетики воздействуют на измеряемый параметр организма в среднем одинаково.

Альтернативная гипотеза обратна нулевой гипотезе. В исследовании обычно формулируется именно нулевая гипотеза.

Получив результат, исследователь делает вывод о том, значим ли полученный результат с точки зрения статистического анализа. Или, другими словами, делает вывод о статистической значимости полученного результата.

Критерий значимости

Для того, чтобы сделать вывод о наличии или отсутствии статистической значимости используется так называемый **критерий значимости**. Полученное числовое значение критерия значимости указывает на то, принимается или отвергается нулевая гипотеза. Однако, **вывод** во многом **зависит** и от того, с какой вероятностью мы можем получить наблюдаемые результаты при верности нулевой гипотезы. Другими словами, мы всегда допускаем наличие ошибки, и максимально допустимую вероятность ее возникновения устанавливаем сами.

Если вероятность ошибки (обычно обозначается p) мала, то мы отвергаем нулевую гипотезу и заключаем, что различия между группами статистически значимы.

Максимальную приемлемую вероятность отвергнуть верную нулевую гипотезу называют **уровнем значимости** и обозначают α . Обычно принимают $\alpha = 0,05$ (5 %).

Малая вероятность ошибочно отвергнуть верную нулевую гипотезу еще не означает, что действие именно изучаемых факторов доказано (это зачастую вопрос планирования эксперимента), но, во всяком случае, маловероятно, что результат обусловлен случайностью.

Если в ходе исследования мы получили результат, который отвергает нулевую гипотезу при уровне значимости 5 %, то можно сказать следующее: *если бы нулевая гипотеза была справедлива, то вероятность получить наблюдаемые результаты была бы меньше 5 %*. В принятой системе обозначений это записывается как $P < 0,05$. P есть вероятность ошибочно отвергнуть верную нулевую гипотезу.

Ошибки первого и второго рода

Исходя из последнего утверждения, различают ошибки первого и второго рода.

Если мы ошибочно отклоняем нулевую гипотезу, например, находим различия там, где их нет, то это называется **ошибкой I рода**.

Максимальная приемлемая вероятность ошибки I рода и есть **уровень значимости**.

Обычно α принимают равной 0,05 (т. е. 5 %), однако можно взять и какой-нибудь другой уровень значимости, например 0,1 или 0,01.

Если мы не отклоняем нулевую гипотезу, когда она не верна, т. е. не находим различий там, где они есть, то это — **ошибка II рода**.

Алгоритм дисперсионного анализа

Для сравнения двух и более групп, при условии нормальности распределения значений признака в них, используют дисперсионный анализ. Задачей

которого является определение значимо или незначимо различие между группами. В ходе анализа вычисляется значение критерия значимости, которое указывает на значимость различий между группами. В начале анализа исследователь формулирует нулевую гипотезу и определяет уровень значимости, затем, получив числовое значение критерия значимости и вероятности ошибочного результата, делает вывод о принятии или отклонении нулевой гипотезы.

Дисперсию совокупности можно оценить двумя способами

*Дисперсия, вычисленная для каждой группы, — это оценка дисперсии совокупности. Поэтому дисперсию совокупности можно оценить на основании групповых дисперсий. Такая оценка не будет зависеть от различий групповых средних. В тоже время, **разброс выборочных средних** тоже позволяет оценить дисперсию совокупности. Такая оценка дисперсии зависит от различий выборочных средних.*

Если экспериментальные группы — это случайные выборки из одной и той же нормально распределенной совокупности, то обе оценки дисперсии совокупности дали бы примерно одинаковые результаты. Поэтому, если эти оценки оказываются близки, то мы не можем отвергнуть нулевую гипотезу. В противном случае мы отвергаем нулевую гипотезу, т. е., заключаем: маловероятно, что мы получили бы такие различия между группами, если бы они были просто случайными выборками из одной нормально распределенной совокупности.

Если верна нулевая гипотеза, то как внутригрупповая, так и межгрупповая дисперсии служат оценками одной и той же дисперсии и должны быть приближенно равны. Исходя из этого, вычислим критерий F (критерий Фишера):

$$F = \frac{\text{Дисперсия совокупности, оцененная по выборочным средним}}{\text{Дисперсия совокупности, оцененная по выборочным дисперсиям}}.$$

И числитель, и знаменатель этого отношения — это оценки одной и той же величины — дисперсии совокупности σ^2 , поэтому значение F должно быть близко к 1.

Если F значительно превышает 1, нулевую гипотезу следует отвергнуть. Если же значение F близко к 1, нулевую гипотезу следует принять.

Критическое значение F

Если извлекать случайные выборки из нормально распределенной совокупности, значение F будет меняться от опыта к опыту.

Значение критерия, начиная с которого мы отвергаем нулевую гипотезу, называется **критическим значением**. Вероятность ошибочно отвергнуть верную нулевую гипотезу, т. е. найти различия там, где их нет, обозначается P .

Как правило, считают достаточным, чтобы эта вероятность не превышала 5 %. (Максимальная приемлемая вероятность ошибочно отвергнуть нулевую гипотезу называется **уровнем значимости и обозначается α**). Критическое значение F однозначно определяется уровнем значимости (обычно 0,05 или 0,01) и еще двумя параметрами, которые называются **внутригрупповым и межгрупповым числом степеней свободы** и обозначаются греческой буквой ν («ню»).

Межгрупповое число степеней свободы — это число групп минус единица $\nu_{\text{меж}} = m - 1$. **Внутригрупповое число степеней свободы** — это произведение числа групп на численность каждой из групп минус единица $\nu_{\text{вну}} = m(n - 1)$. Вычислить критическое значение F довольно сложно, поэтому пользуются таблицами критических значений F для разных α , $\nu_{\text{меж}}$ и $\nu_{\text{вну}}$.

Таким образом, на основании значения уровня значимости и степеней свободы определяется величина **критического значения критерия Фишера $F_{\text{кр}}$** (значение критерия, начиная с которого мы отвергаем нулевую гипотезу).

Для определения статистической значимости различий между группами сравниваются между собой F и $F_{\text{кр}}$, а также P и α ,

если $F > F_{\text{кр}}$ и $P < \alpha$, то нулевая гипотеза отвергается — различия между группами статистически значимы, в противном случае различия случайны или незначимы. Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но допустимая вероятность **равна уровню значимости** (т. е. незначительна).

➤ **Общая последовательность действий:**

1. Формирование исследуемых групп (например, первая группа принимает препарат, вторая — плацебо).
2. Формирование таблиц результатов измерения (снимаются исследуемые показатели первой группы и показатели второй группы).
3. Выполняется первичный анализ полученных данных (описательная статистика).
4. Если данные в обеих группах распределены по нормальному закону распределения (или близкому к таковому), то выбирается метод анализа, в нашем случае — дисперсионный анализ.

5. Выбирается значение уровня значимости (в зависимости от строгости исследования: 0,1 или 0,05 или 0,01). Следует помнить об ошибках I и II рода.

6. Формулируется нулевая гипотеза (различие между группами незначимо или является следствием случайности).

7. Рассчитывается критерий Фишера F и вероятность ошибочного результата P (вероятность ошибочно отвергнуть верную нулевую гипотезу, т. е. найти различия там, где их нет).

8. На основании значения уровня значимости и степеней свободы определяется величина критического значения критерия Фишера $F_{кр}$ (значение критерия, начиная с которого мы отвергаем нулевую гипотезу). Сравниваются между собой F и $F_{кр}$, а также P и α , если $F > F_{кр}$ и $P < \alpha$, то нулевая гипотеза отвергается и различия между группами статистически значимы, в противном случае различия случайны или незначимы. Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но допустимая вероятность равна уровню значимости (т. е. незначительна).

✓ Задача

Позволяет ли правильное лечение сократить срок госпитализации?

Стоимость пребывания в больнице — самая весомая статья расходов на здравоохранение. Сокращение госпитализации без снижения качества лечения дало бы значительный экономический эффект. Способствует ли соблюдение официальных схем лечения сокращению госпитализации? Чтобы ответить на этот вопрос, Кнапп и соавторы изучили истории болезни лиц, поступивших в бесплатную больницу с острым пиелонефритом. Острый пиелонефрит был выбран как заболевание, имеющее четко очерченную клиническую картину и столь же четко регламентированные методы лечения.

Чтобы избежать возможных ловушек исследования, **чрезвычайно важно в явном виде задать критерии, по которым больных относили к той или иной группе**. Самому исследователю это поможет избежать невольного самообмана, читателю работы это даст возможность судить, насколько результаты исследования приложимы к его больным. Кнапп сформулировал следующие критерии включения в исследование:

1. Диагноз при выписке — острый пиелонефрит.
2. При поступлении — боли в пояснице, температура выше 37,8 °С.
3. Бактериурия более 100 000 колоний/мл, определена чувствительность к антибиотикам.
4. Возраст от 18 до 44 лет (больных старше 44 лет не включали в связи с высокой вероятностью сопутствующих заболеваний, ограничивающих выбор терапии).

5. Отсутствие почечной, печеночной недостаточности, а также заболеваний, требующих хирургического лечения (эти состояния тоже ограничивают выбор терапии).

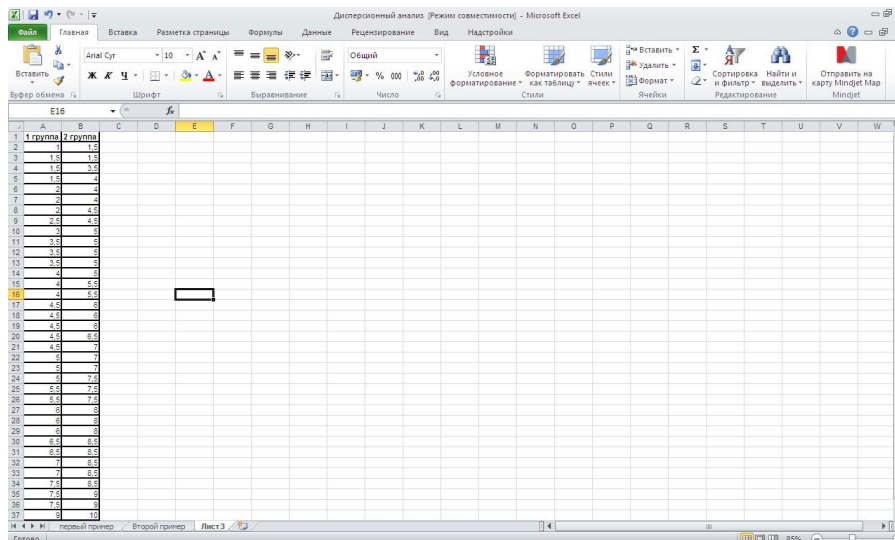
6. Больной был выписан в связи с улучшением (т. е. не покинул больницу самовольно, не умер и не был переведен в другое лечебное учреждение).

Кроме того, исследователи сформулировали критерий того, что считать «правильным» лечением. Правильным считалось лечение, соответствующее рекомендациям авторитетного справочника по лекарственным средствам «Physicians' Desk Reference» («Настольный справочник врача»). По этому критерию больных разделили на две группы: леченных правильно (1-я группа) и неправильно (2-я группа). В обеих группах было по 36 больных. Результат представлен в таблице:

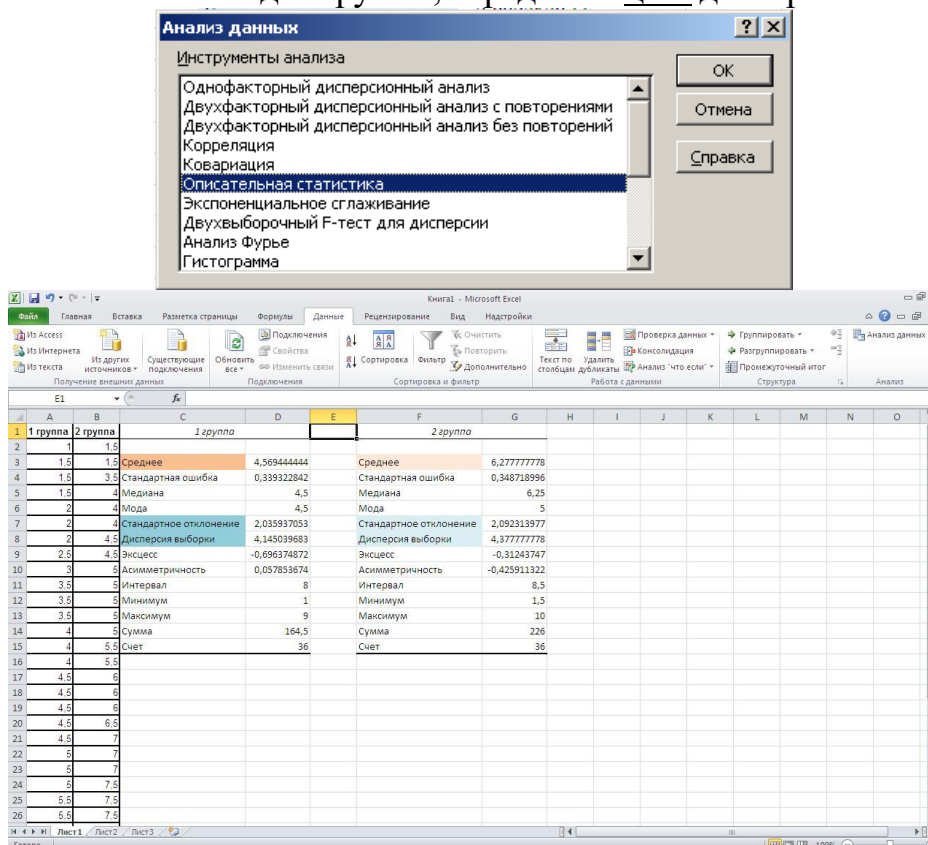
1 группа	2 группа
1	1,5
1,5	1,5
1,5	3,5
1,5	4
2	4
2	4
2	4,5
2,5	4,5
3	5
3,5	5
3,5	5
3,5	5
4	5
4	5,5
4	5,5
4,5	6
4,5	6
4,5	6
4,5	6,5
4,5	7
5	7
5	7
5	7,5
5,5	7,5
5,5	7,5
6	8
6	8
6	8
6,5	8,5
6,5	8,5
7	8,5
7	8,5
7,5	8,5
7,5	9
7,5	9
9	10

Решение задачи в MS Excel

1. Нулевая гипотеза: правильность лечения не влияет на сроки госпитализации, различия в группах случайны и статистически не значимы.
2. Скопируйте таблицу с данными в табличный редактор MS Excel



3. При помощи модуля «**Описательная статистика**» (последовательность действий вам известна из предыдущих работ) получите значения основных статистических параметров (среднее значение, стандартное отклонение, дисперсия) исходных данных для каждой группы. Проследите разницу между средними значениями каждой группы, определите цель дисперсионного анализа.

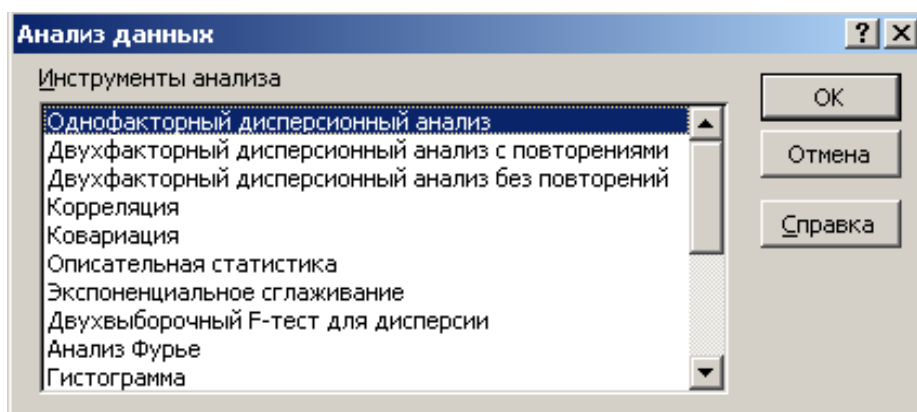


Средние значения отличаются друг от друга по величине, следовательно при помощи дисперсионного анализа необходимо проверить статистическую значимость этих различий. **Нулевая гипотеза:** *различие средних значений статистически не значимо.*

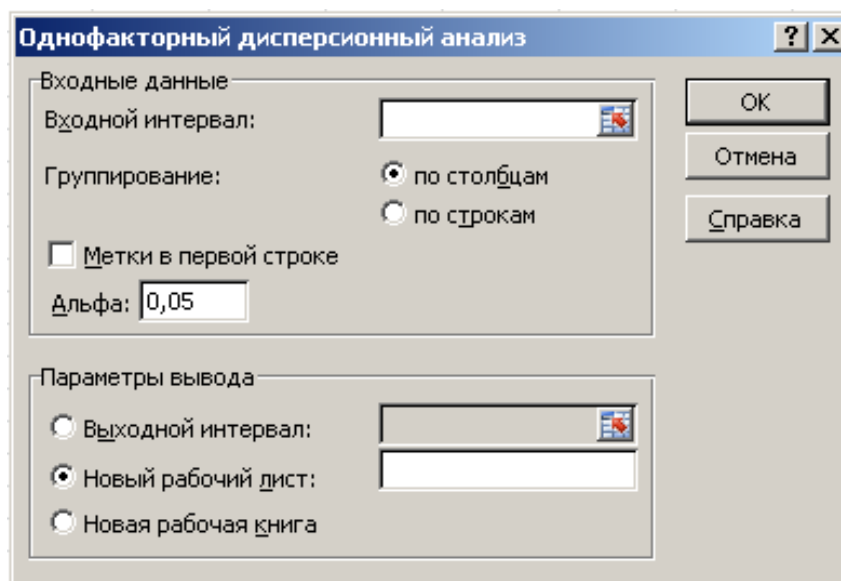
Следует указать, что для проведения дисперсионного анализа необходимо убедиться в **нормальности распределения** значений в обеих группах (построить график распределения). Небольшое отклонение от этого правила допустимо. Также величина дисперсии в первой и второй группах должна приблизительно совпадать.

4. Порядок проведения дисперсионного анализа в *MS Excel*.

Так как в задаче проверяется влияние одного фактора (правильность лечения), то необходимо воспользоваться **однофакторным дисперсионным анализом**: Вкладка «Данные — Анализ данных — Однофакторный дисперсионный анализ».



Входной интервал — данные 1 и 2 групп (выделяются при помощи мыши, с заголовками), ставим галочку «*Метки в первой строке*», «*Альфа — 0,05*»; «*Выходной интервал*» — любая свободная ячейка.



Образец:

- ✓ Что такое альфа?
- ✓ Почему значение альфа выбрано 0,05?
- ✓ Можно ли было выбрать другое значение, например 0,01?

Результат дисперсионного анализа. Интересующие значения оцениваемых параметров выделены цветом:

Однофакторный дисперсионный анализ					
ИТОГИ					
Группы	Счет	Сумма	Среднее	Дисперсия	
1 группа	36	164,5	4,569444444	4,145039683	
2 группа	36	228	6,277777778	4,377777778	
Дисперсионный анализ					
Источник вариации	SS	df	MS	F	P-Значение F критическое
Между группами	52,53125	1	52,53125	12,32720289	0,000785693 3,977779393
Внутри групп	298,2986111	70	4,26140873		
Итого	350,8298611	71			

В рассмотренном примере эмпирический **F-критерий** (критерий Фишера) показывает, что различие между средними статистически значимо (значимо на уровне $p = 0,00078$, меньшем, чем критическое значение 0,05).

Вывод: Так как расчетное значение **критерия Фишера F** больше его **критического значения $F_{кр}$** при **уровне значимости (альфа) - 0,05**, то нулевая гипотеза **отвергается**. **Вероятность ошибки P** меньше уровня значимости. Таким образом, правильность лечения влияет на длительность срока пребывания пациента в клинике.

Порядок анализа в «Statistica»

Введите исходные данные таблицы (из предыдущего примера) так, как это показано на иллюстрации:

	A	B
1	a	1
2	a	1.5
3	a	1.5
4	a	1.5
5	a	2
6	a	2
7	a	2
8	a	2.5
9	a	3
10	a	3.5
11	a	3.5
12	a	3.5
13	a	4
14	a	4
15	a	4
16	a	4.5
17	a	4.5
18	a	4.5
19	a	4.5
20	a	4.5
21	a	5
22	a	5
23	a	5
24	a	5.5
25	a	5.5
26	a	6
27	a	6
28	a	6
29	a	6
30	a	6.5
31	a	7
32	a	7
33	a	7.5
34	a	7.5
35	a	7.5
36	b	1.5
37	b	1.5
38	b	1.5
39	b	3.5

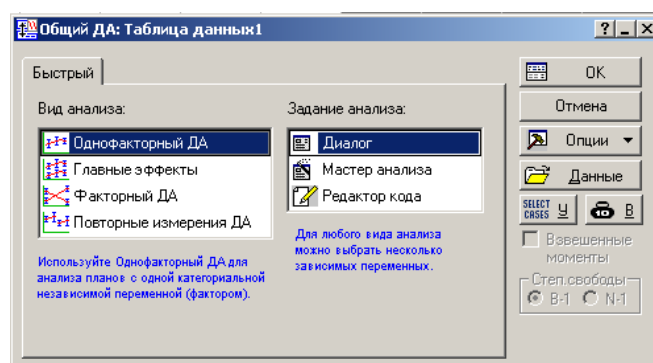
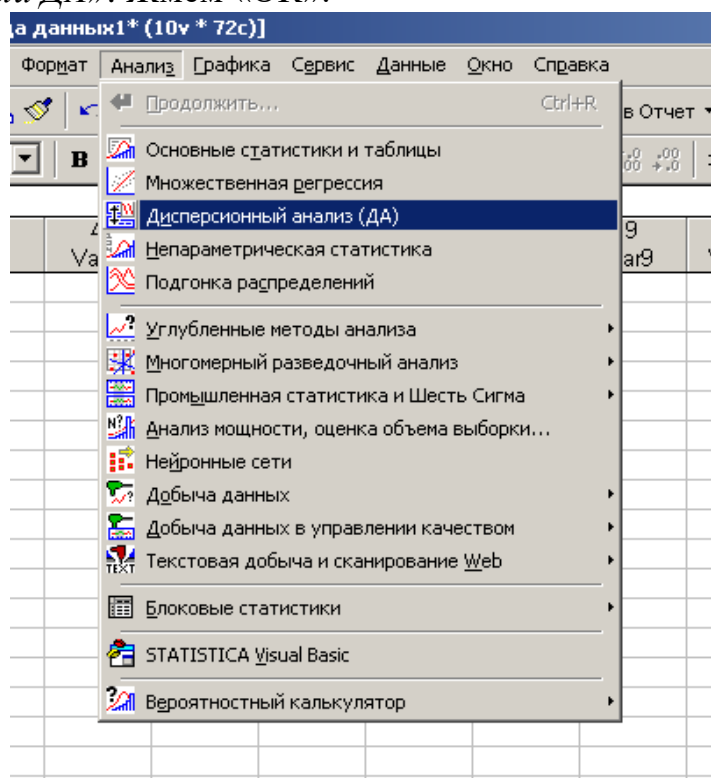
Т. е. напротив значений параметра из группы 1 ставим индекс (предиктор) **a**, напротив значений группы 2 — индекс (предиктор) **b**.

Копируем получившуюся таблицу в «Statistica» **6**.

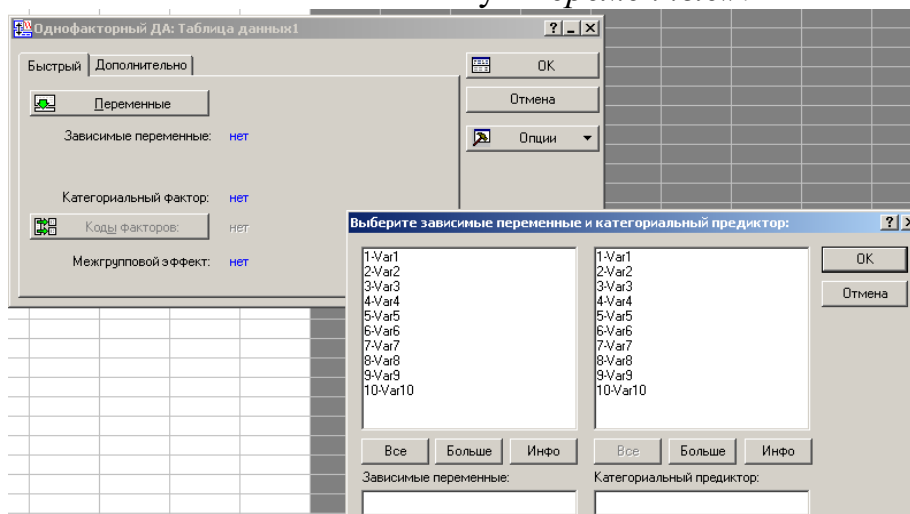
	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8	9 Var9	10 Var10
1	a	1								
2	a	1.5								
3	a	1.5								
4	a	1.5								
5	a	2								
6	a	2								
7	a	2								
8	a	2.5								
9	a	3								
10	a	3.5								
11	a	3.5								
12	a	3.5								
13	a	4								
14	a	4								
15	a	4								
16	a	4.5								
17	a	4.5								
18	a	4.5								
19	a	4.5								
20	a	4.5								
21	a	5								
22	a	5								
23	a	5								
24	a	5.5								
25	a	5.5								
26	a	6								
27	a	6								
28	a	6								
29	a	6.5								
30	a	6.5								
31	a	7								
32	a	7								
33	a	7.5								

Проведем «Дисперсионный анализ» (ANOVA).

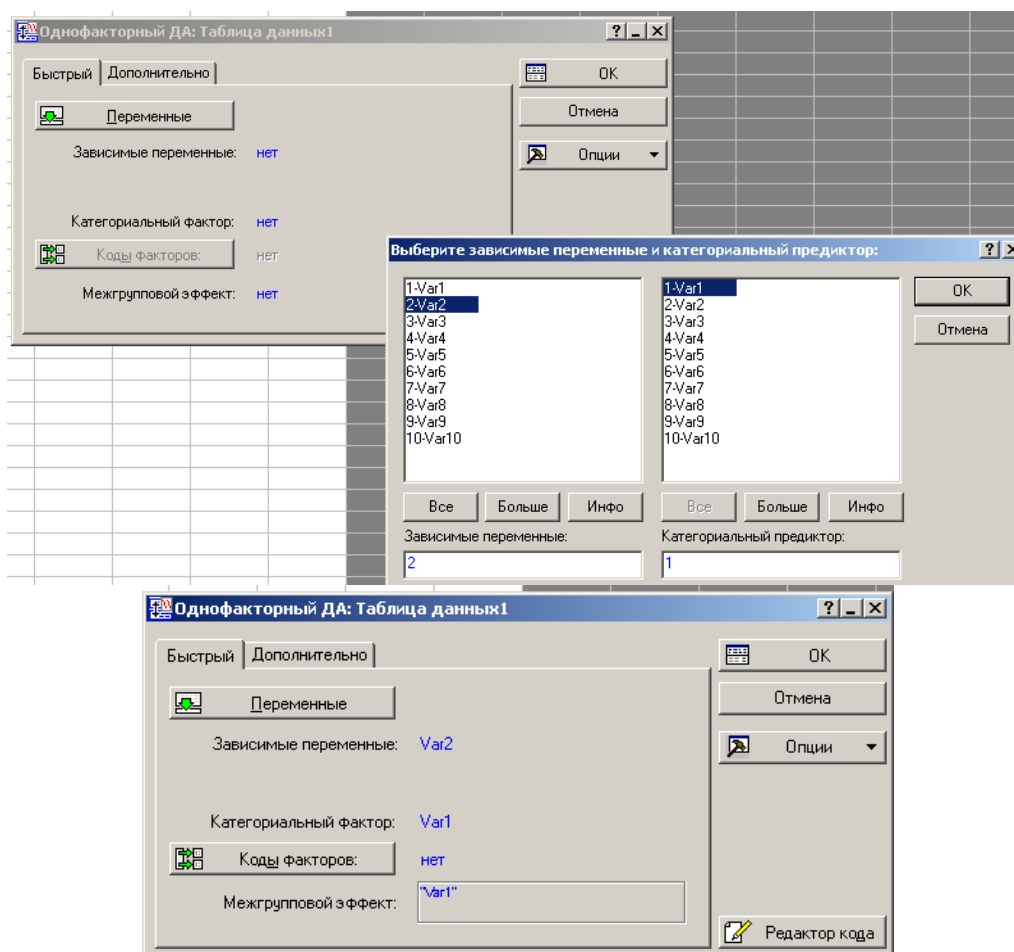
Для этого выбираем: вкладка «Анализ — Дисперсионный анализ — Однофакторный ДА». Жмем «ОК».



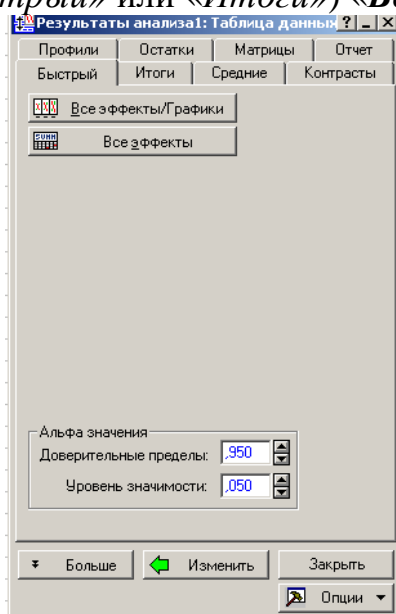
В появившемся окне жмем кнопку «Переменные».



Указываем: «*Зависимые переменные*» — числовые значения (в нашем случае это столбец Var 2), «*Категориальный предиктор*» — буквы **a** и **b** — Var 1. Как показано на иллюстрации.



Жмем «ОК» и в появившемся окне «Результаты анализа» нажимаем кнопку (во вкладке «Быстрый» или «Итоги») «**Все эффекты**»:



В результате получаем таблицу результата анализа (Строка *Var 1*).
Интересующие параметры: **критерий Фишера F** и **вероятность ошибки p**

STATISTICA - [Workbook1* - Одномерный критерий значимости для Var2 (Таблица данных1)]

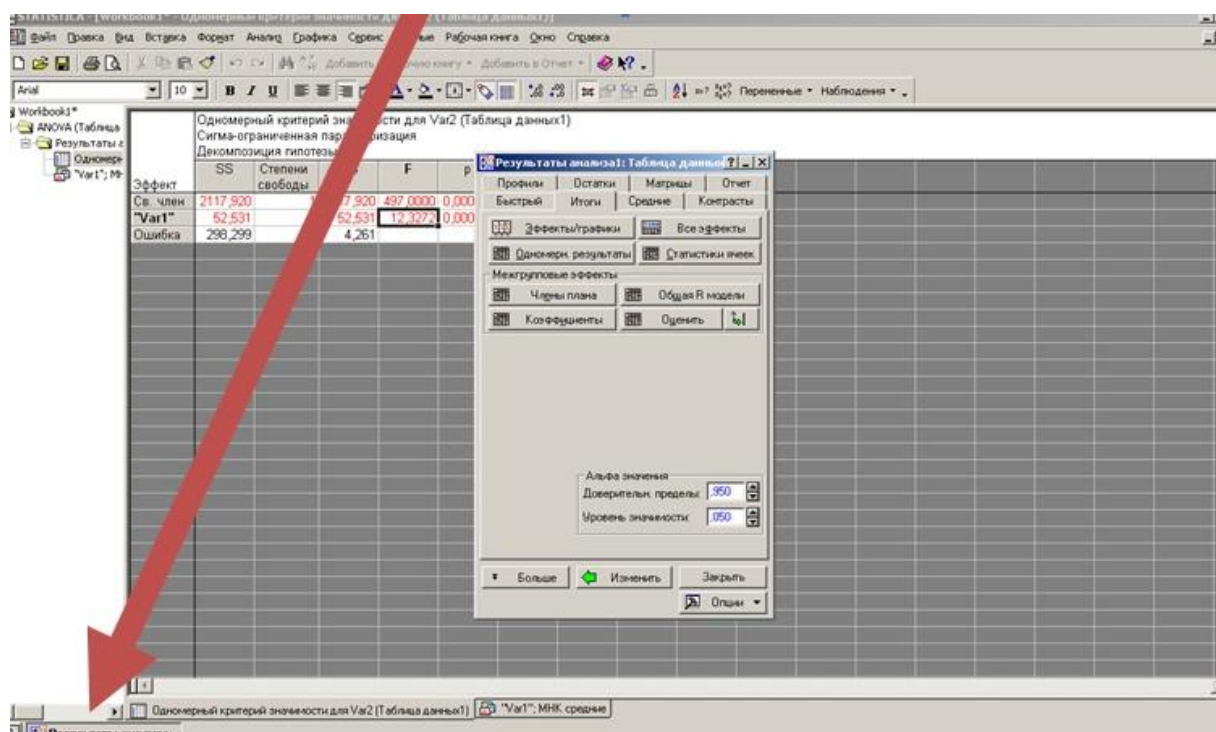
Файл Правка Вид Вставка Формат Анализ Графика Сервис Данные Рабочая книга Окно Справка

Workbook1*
ANOVA (Таблица)
Результаты анализа
Одномерный критерий значимости для Var2 (Таблица данных1)

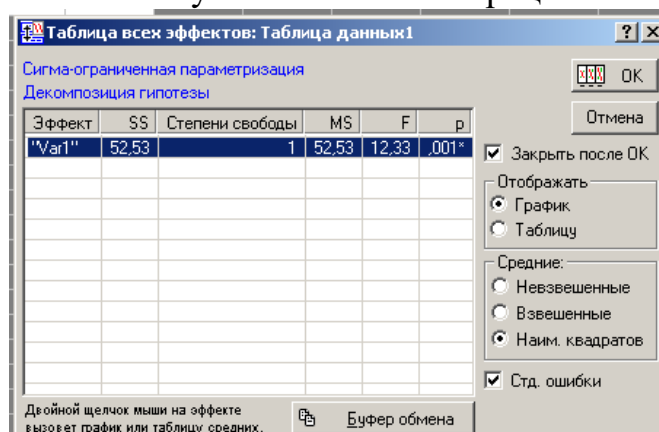
Сигма-ограниченная параметризация
Декомпозиция гипотезы

Эффект	SS	Степени свободы	MS	F	p
Св. член	2117,920	1	2117,920	497,0000	0,000000
"Var1"	52,531	1	52,531	12,3272	0,000786
Ошибка	298,299	70	4,261		

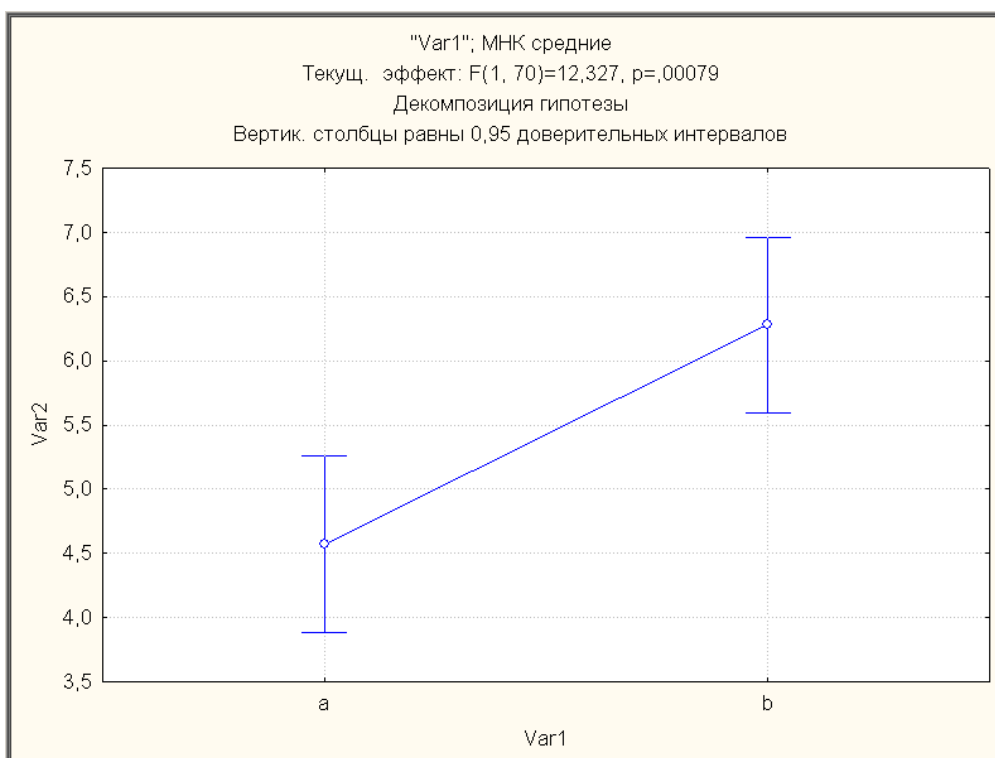
Возвращаемся в окно «**Результаты анализа**».



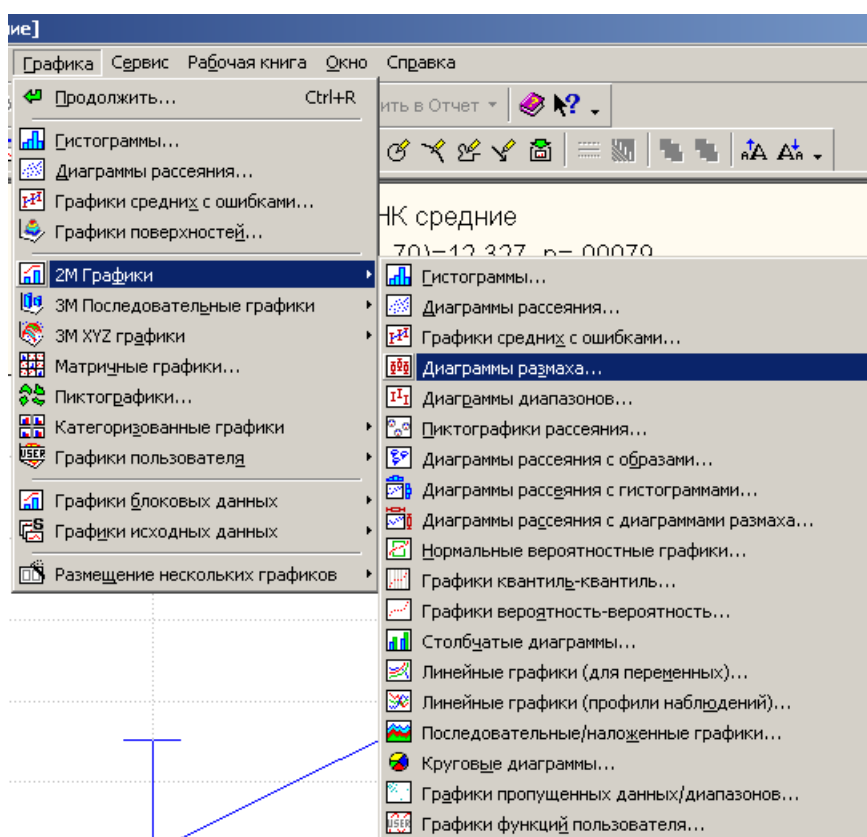
Зайдите во вкладку «**Итоги**» и жмем кнопку «**Эффекты/графики**».
Устанавливаем отметки как указано на иллюстрации.



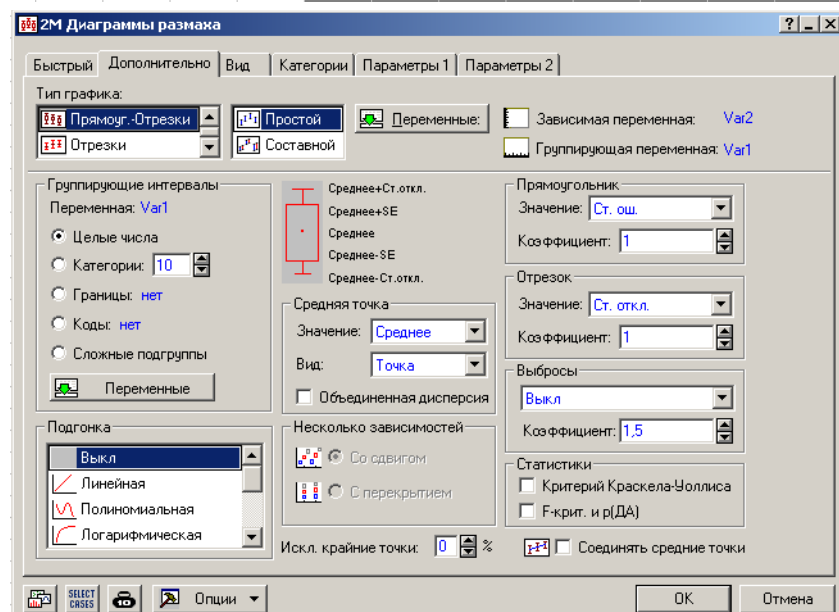
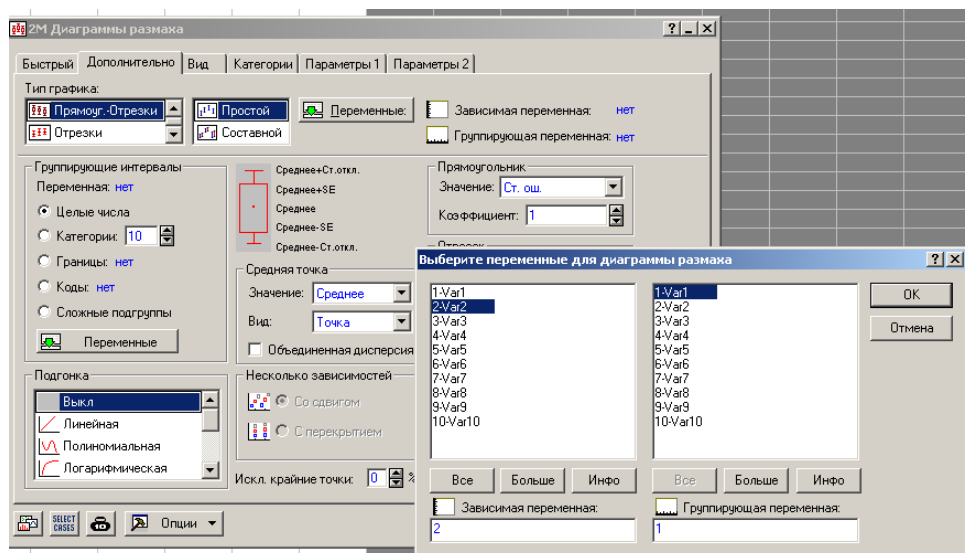
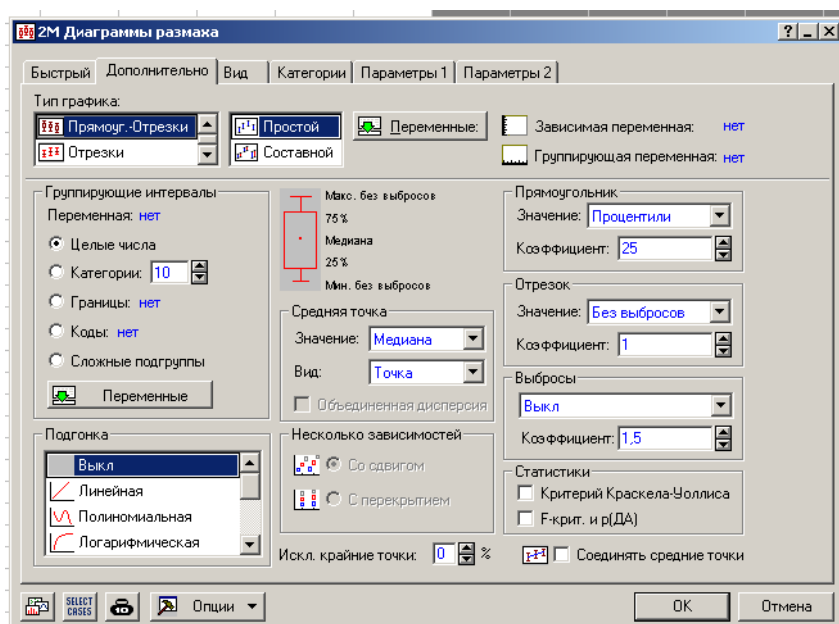
Графическая интерпретация результата.



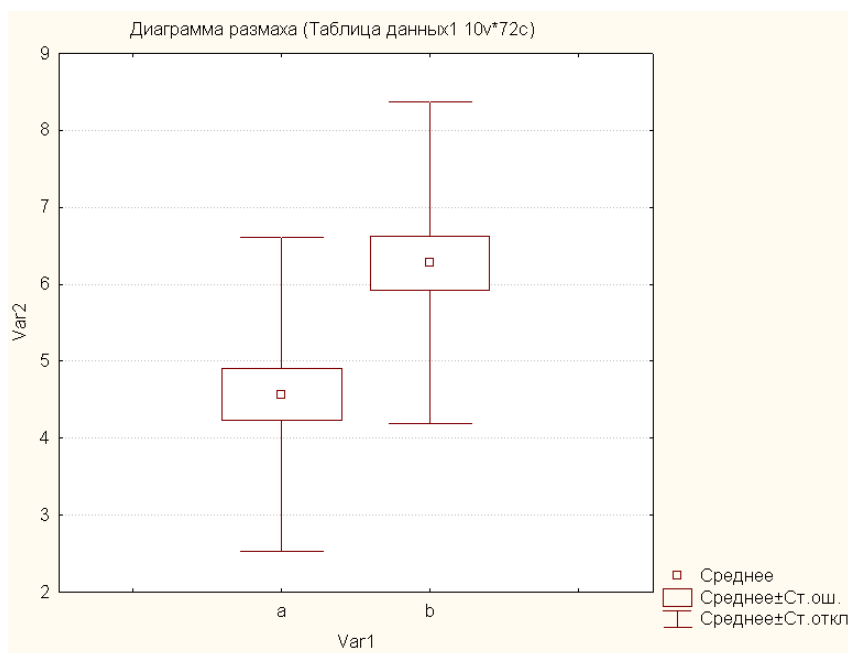
Графическую интерпретацию можно получить и следующим образом:



Указываем переменные и устанавливаем флажки как указано на следующих иллюстрациях.



Полученная графическая интерпретация результата исследования:



В рассмотренном примере **F-критерий** показывает, что различие между группами статистически значимо (значимо на уровне 0,00079, т. е. меньше, чем критическое значение 0,05). Поскольку различие между средними значениями значимо, **нулевая гипотеза отвергается** и принимается альтернативная гипотеза о значимости различий между группами (результат в строке **Var1** подсвечивается **красным цветом**).

Вывод: Так как расчетное значение **критерия Фишера F** больше его **критического значения на уровне значимости (альфа) равного 0,05**, то нулевая гипотеза **отвергается**. **Вероятность ошибки P** меньше уровня значимости. Таким образом, можно утверждать, что правильность лечения влияет на длительность срока пребывания пациента в клинике.

✓Задание

Галотан и морфин при операциях на открытом сердце

Препарат галотан, широко используемый при общей анестезии. Он обладает сильным действием, удобен в применении и очень надежен. Галотан — газ, его можно вводить через респиратор. Поступая в организм через легкие, галотан действует быстро и кратковременно, поэтому, регулируя подачу препарата можно оперативно управлять анестезией. Однако галотан имеет существенный недостаток — он угнетает сократимость миокарда и расширяет вены, что ведет к падению АД. В связи с этим было предложено вместо галотана для общей анестезии применять морфин, который не снижает АД. Т. Конахан и соавт. сравнили галотановую и морфиновую анестезию у больных, подвергшихся операции на открытом сердце. В исследование включали больных, у которых не было противопоказаний ни к галотану, ни к морфину. Способ анестезии (галотан или мор-

фин) выбирали случайным образом. Регистрировали следующие показатели: параметры гемодинамики на разных этапах операции, длительность пребывания в реанимационном отделении и общую длительность пребывания в больнице после операции; а также послеоперационную летальность. Сосредоточим внимание на артериальном давлении между началом анестезии и началом операции. Именно в этот период артериальное давление наиболее адекватно отражает гипотензивное действие анестетика, поскольку в дальнейшем начинает сказываться гипотензивный эффект самой операции. Артериальное давление между началом анестезии и началом операции измеряли многократно, каждый раз вычисляя среднее артериальное давление (данные заносились в таблицу). В исследование вошло 122 больных. У половины больных использовали галотан (1-я группа), у половины — морфин (2-я группа). Результаты представлены в таблице. Данные округлены до ближайшего четного числа.

Галотановая	Морфиновая
42	42
44	46
46	46
46	52
48	56
48	58
50	58
50	58
52	58
54	60
54	60
56	60
58	62
58	62
58	62
60	62
60	62
60	64
60	64
60	64
62	66
62	66
62	66
62	66
62	68
64	68
64	68
66	68

66	70
66	70
66	70
66	72
68	74
68	74
68	76
68	76
70	76
70	78
70	80
70	80
70	80
72	82
72	82
72	82
72	84
74	84
74	84
74	84
74	84
74	86
76	86
78	86
78	88
78	88
80	90
80	90
82	92
82	96
84	96
86	98
90	98
94	98
98	100

Выполнить дисперсионный анализ данных путем выше изложенного алгоритма и сделать вывод о том, какой вид анестезии — галотановая или морфиновая — благоприятнее влияет на организм человека при операции на открытом сердце.

Контрольные вопросы

1. В каких случаях необходимо производить сравнение групп?
2. Какая группа обычно называется контрольной?
3. Какая экспериментальной?

4. Что такое гипотеза?
5. Что такое нулевая гипотеза?
6. Что называется уровнем значимости?
7. Какие значения уровня значимости наиболее часто встречаются в медицинских исследованиях?
8. Что значит «ошибочно отвергнуть верную нулевую гипотезу»?
9. Что подразумевается, когда говорят, что уровень значимости равен 0,05?
10. Алгоритм дисперсионного анализа.
11. Что такое критическое значение критерия F?
12. Если расчетное значение критерия F больше чем $F_{\text{критическое}}$ это означает, что ...?

Лабораторная работа № 4

Сравнение групп Критерий Стьюдента

Краткие сведения из теории

Результатом исследования обычно является **утверждение**, имеющее важное значение, например, какой метод лучше для лечения болезни, каковы самые распространенные побочные эффекты определенного вида хирургического вмешательства, каков коэффициент выживаемости после определенного лечения и действительно ли новое экспериментальное лекарство улучшает состояние организма.

Однако, прежде чем получить результат (утверждение или вывод), необходимо правильно спланировать ход исследования, одним из важных этапов является *формулировка исходного утверждения, правильность которого и проверяется в ходе исследования*. Такое утверждение называется **гипотезой**.

Гипотеза — утверждение, предполагающее доказательство.

Проверка гипотезы — это статистическая процедура, предназначенная для проверки утверждения. Обычно утверждение касается исследуемого параметра совокупности (т. е. некой величины, которая характеризует какую-то интересующую нас сторону объекта).

К примеру, среднее время госпитализации у первой исследуемой группы составило 4,5 суток, у второй — 6,3 суток (**переменная** — время госпитализации, **параметр** — среднее значение дней госпитализации). В этом примере под гипотезой может выступать утверждение: различие между средними значениями статистически значимо

После того как исследователь определил переменную и параметр, влияние на которые того или иного фактора проверяется в исследовании, он формулирует *нулевую гипотезу*.

Например, при лечении пациентов согласно официальным схемам лечения (первая группа) и лечение той же болезни, но не согласующееся полностью с официальными схемами (вторая группа), среднее время госпитализации составило: у первой исследуемой группы 4,5 суток, у второй — 6,3 суток. В этом случае влияющий фактор — лечение согласно официальным схемам лечения. А **нулевая гипотеза**: различие между группами является случайным и статистически незначимо.

Нулевая гипотеза — это предположение, что исследуемые факторы не оказывают никакого влияния на измеряемый признак и полученные различия (исследуемых параметров, средних значений количества суток госпитализации у двух групп выше) случайны.

Например, сравнивая различающиеся средние значения артериального давления при применении двух разных анестетиков, мы выдвигаем нулевую гипотезу о том, что полученные различия случайны и значимой разницы между двумя анестетиками нет.

Статистическая значимость результатов. Уровень значимости

Получив результат, исследователь делает вывод о том, значим ли полученный результат с точки зрения статистического анализа. Другими словами, он делает вывод о **статистической значимости** полученного результата.

Для того, чтобы сделать вывод о наличии или отсутствии статистической значимости используется так называемый **критерий значимости**.

Полученное числовое значение критерия значимости указывает на то, принимается или отвергается нулевая гипотеза. Однако, **вывод** во многом зависит и от того с какой вероятностью мы можем получить наблюдаемые результаты при верности нулевой гипотезы.

Статистическое исследование всегда допускает возможное наличие ошибки, и необходимо учитывать максимальную вероятность ее возникновения. Если эта вероятность **мала**, то нулевая гипотеза отвергается и делается заключение о значимости полученного результата (например, различие средних значений двух групп статистически значимо и не обусловлено случайностью).

Максимальную приемлемую вероятность отвергнуть верную нулевую гипотезу называют **уровнем значимости** и обозначают **α (альфа)**.

Обычно принимают **$\alpha = 0,05$ (5 %)**. Однако эта величина может быть и другой: 0,01 (1 %), 0,1(10 %) и т. д., в зависимости от строгости исследования.

Если в ходе исследования мы получили результат, который отвергает нулевую гипотезу, при уровне значимости 5 %, то можно сказать следующее: *если бы нулевая гипотеза была справедлива, то вероятность получить наблюдаемые результаты была бы меньше 5 %*. В принятой системе обозначений это записывается как $P < 0,05$. *Р* есть вероятность ошибочно отвергнуть нулевую гипотезу. Это еще не означает что влияние изучаемых факторов доказано (это вопрос тесно связан с планированием эксперимента), но, в тоже время, маловероятно, что результат обусловлен случайностью.

Стандартное заключение может выглядеть следующим образом: нулевая гипотеза об отсутствии влияния исследуемого препарата, например, на давление, вряд ли справедлива, и различия между группами статистически значимы при 5 % уровне значимости. Разумеется, этот вывод по сути своей **носит вероятностный характер**. Не исключено, что ошибочно признается неэффективный препарат эффективным, т. е. находятся различия там, где их нет. Однако можно утверждать, что вероятность подобной ошибки **не превышает 5 %**.

Критерий Стьюдента

Критерий Стьюдента используется для сравнения только двух групп. Это частный случай дисперсионного анализа. Критерий Стьюдента чрезвычайно популярен, он используется более чем в половине медицинских публикаций.

Следует помнить, что критерий Стьюдента предназначен для сравнения только двух групп, а не нескольких групп попарно.

Ошибочное использование критерия Стьюдента увеличивает вероятность «выявить» не существующие различия.

Например, вместо того чтобы признать несколько методов лечения равно эффективными (или неэффективными), один из них объявляют «лучшим».

Общая формула для вычисления критерия Стьюдента:

$$t = \frac{\text{Разность выборочных средних}}{\text{Стандартная ошибка разности выборочных средних}}.$$

Для двух случайных выборок извлеченных из одной нормально распределенной совокупности это отношение, как правило, будет близко к нулю. **Чем меньше** (по абсолютной величине, по модулю) t , **тем больше вероятность справедливости нулевой гипотезы**. **Чем больше t , тем больше оснований отвергнуть нулевую гипотезу и считать, что различия статистически значимы**. Величина критерия Стьюдента, начиная с которой отвергается нулевая гипотеза, называется **критическим значением** критерия Стьюдента ($t_{кр}$).

Если значение t критерия Стьюдента по модулю **больше** чем критическое значение критерия Стьюдента (найденное по таблице критических значений или рассчитанное при помощи программного обеспечения) для заданного уровня значимости, то нулевая гипотеза **отвергается**, и различия считаются **статистически значимыми**.

Это означает, что если бы группы представляли собой две случайные выборки из одной и той же совокупности, то вероятность получить наблюдаемые различия (или более сильные) равна 0,05 (или другому значению выбранного уровня значимости). Следовательно, ошибочный вывод о существовании различий мы будем делать в 5 % случаев. Застраховаться от подобных ошибок можно приняв уровень значимости не 0,05, а к примеру 0,01. Однако даже в этом случае ошибочные выводы о существовании различий все же не исключены — их вероятность снижается до 1 %. И в тоже время вероятность не найти различия, там где они есть, теперь повысилась.

Ошибки в использовании критерия Стьюдента

Критерий Стьюдента предназначен для сравнения двух групп. Однако на практике он широко используется для оценки различий большего числа групп посредством попарного их сравнения. При этом вступает в силу *эффект множественных сравнений*.

Пример

Исследуется влияние препаратов А и Б на уровень глюкозы плазмы. Исследование проводят на трех группах — получавших препарат А, получавших препарат Б и получавших плацебо В. С помощью критерия Стьюдента проводят 3 парных сравнения: группу А сравнивают с группой В, группу Б — с группой В и наконец А с Б. Получив достаточно высокое значение t в каком-либо из трех сравнений, сообщают, что « $P < 0,05$ ». Это означает, что вероятность ошибочного заключения о существовании различии не превышает 5 %. Но это неверно: вероятность ошибки значительно превышает 5 %. Разберемся подробнее. В исследовании был принят 5 % уровень значимости. Значит, вероятность ошибиться при сравнении групп А и В — 5 %. Казалось бы все правильно. Но точно также мы ошибемся в 5 % случаев при сравнении групп Б и В. И, наконец, при сравнении групп А и Б ошибка возможна также в 5 % случаев. Следовательно, вероятность ошибиться хотя бы в одном из трех сравнений составит не 5%, а значительно больше. Итак, в нашем исследовании вероятность ошибиться хотя бы в одном из сравнений составляет примерно 15 %. При сравнении четырех групп число пар и соответственно возможных попарных сравнений равно 6. Поэтому при уровне значимости в каждом из сравнений 0,05 вероятность ошибочно обнаружить различие хотя бы в одном равна уже не 0,05, а примерно $6 \times 0,05 = 0,30$. И когда исследователь, выявив таким способом «эффективный» препарат, будет говорить про 5 % вероятности ошибки, на самом деле эта вероятность равна 30 %.

Правила использования критерия Стьюдента:

- Критерий Стьюдента может быть использован только в случае выборок с **нормально распределенными** значениями признака.
- Критерий Стьюдента может быть использован для проверки гипотезы о различии средних **только для двух групп**.
- Если схема эксперимента предполагает большее число групп, необходимо воспользоваться дисперсионным анализом.
- Если критерий Стьюдента был использован для проверки различий между несколькими группами, то истинный уровень значимости можно получить, умножив уровень значимости, приводимый авторами на число возможных сравнений.

➤ **Общая последовательность действий при использовании критерия Стьюдента:**

1. Формирование таблиц результатов измерения (записываются и организуются в таблицу исследуемые показатели первой группы и показатели второй группы).
2. Выполняется анализ полученных данных (описательная статистика)
3. Проверка соответствия распределения данных нормальному закону распределения.
4. Выбирается значение уровня значимости (в зависимости от строгости исследования: 0,1 или 0,05, или 0,01).
5. Формулируется нулевая гипотеза (различие между группами незначимо или является следствием случайности).
6. Рассчитывается значение критерия Стьюдента t (значение критерия, начиная с которого мы отвергаем нулевую гипотезу) и вероятность ошибочного результата P (вероятность ошибочно отвергнуть верную нулевую гипотезу, т. е. найти различия там, где их нет).
7. На основании значения уровня значимости и количества элементов выборки определяется величина критического значения критерия $t_{кр}$ (значение критерия, начиная с которого мы отвергаем нулевую гипотезу).
8. Сравниваются между собой t и $t_{кр}$, а также P и α , если $t > t_{кр}$ и $P < \alpha$, то нулевая гипотеза отвергается и различия между группами статистически значимы, в противном случае различия — случайны или незначимы. Следует учитывать, что даже при обнаруженной статистической значимости различий исследователь может ошибаться, но допустимая вероятность равна уровню значимости.

✓ **Задача**

Позволяет ли правильное лечение сократить срок госпитализации?

Стоимость пребывания в больнице — самая весомая статья расходов на здравоохранение. Сокращение госпитализации без снижения качества лечения

дало бы значительный экономический эффект. Способствует ли соблюдение официальных схем лечения сокращению госпитализации? Чтобы ответить на этот вопрос, Кнапп и соавторы изучили истории болезни лиц, поступивших в бесплатную больницу с острым пиелонефритом. Острый пиелонефрит был выбран как заболевание, имеющее четко очерченную клиническую картину и столь же четко регламентированные методы лечения.

Чтобы избежать ловушек обсервационного (и особенно ретроспективного) исследования, **чрезвычайно важно в явном виде задать критерии, по которым больных относили к той или иной группе**. Самому исследователю это поможет избежать невольного самообмана, читателю работы это даст возможность судить, насколько результаты исследования приложимы к его больным. Кнапп и соавт. сформулировали следующие критерии включения в исследование:

1. Диагноз при выписке — острый пиелонефрит.
2. При поступлении — боли в пояснице, температура выше 37,8 °C.
3. Бактериурия более 100 000 колоний/мл, определена чувствительность к антибиотикам.
4. Возраст от 18 до 44 лет (больных старше 44 лет не включали в связи с высокой вероятностью сопутствующих заболеваний, ограничивающих выбор терапии).
5. Отсутствие почечной, печеночной недостаточности, а также заболеваний, требующих хирургического лечения (эти состояния тоже ограничивают выбор терапии).
6. Больной был выписан в связи с улучшением (т. е. не покинул больницу самовольно, не умер и не был переведен в другое лечебное учреждение).

Кроме того, исследователи сформулировали критерий того, что считать «правильным» лечением. Правильным считалось лечение, соответствующее рекомендациям авторитетного справочника по лекарственным средствам «Physicians' Desk Reference» («Настольный справочник врача»). По этому критерию больных разделили на две группы: леченных правильно (1-я группа) и неправильно (2-я группа). В обеих группах было по 36 больных. Результат представлен в таблице:

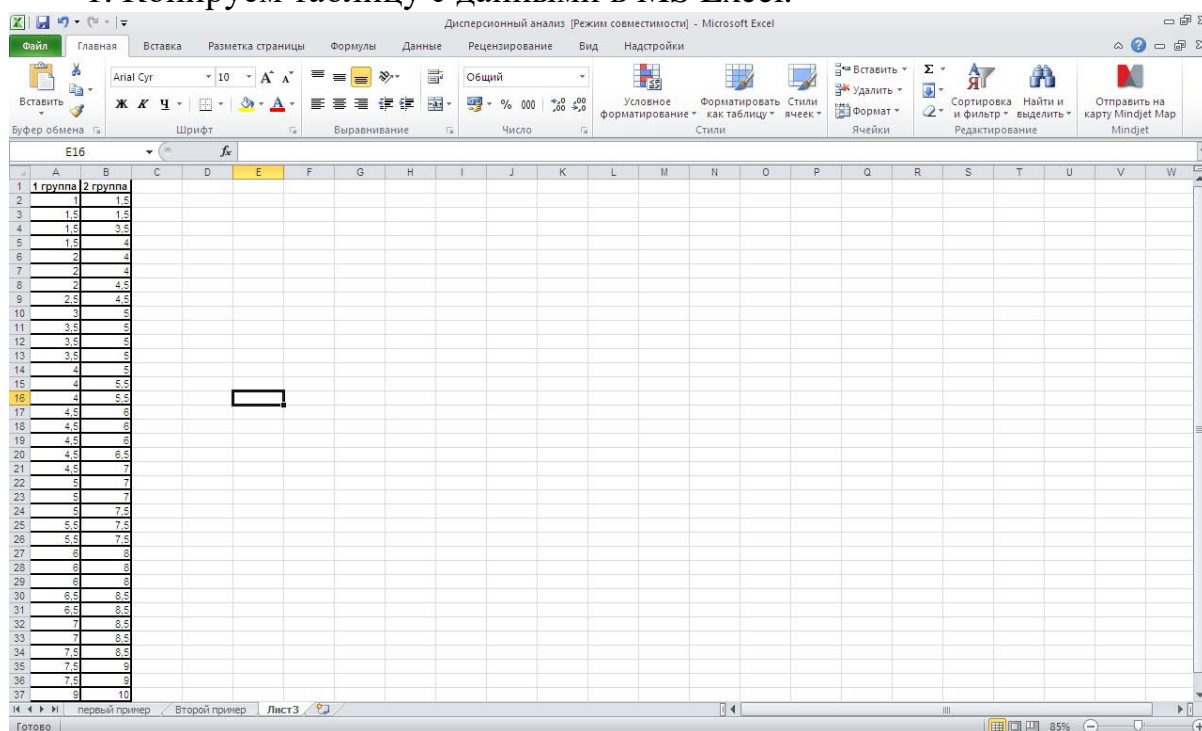
1 группа	2 группа
1	1,5
1,5	1,5
1,5	3,5
1,5	4
2	4
2	4
2	4,5
2,5	4,5
3	5
3,5	5
3,5	5

3,5	5
4	5
4	5,5
4	5,5
4,5	6
4,5	6
4,5	6
4,5	6,5
4,5	7
5	7
5	7
5	7,5
5,5	7,5
5,5	7,5
6	8
6	8
6	8
6,5	8,5
6,5	8,5
7	8,5
7	8,5
7,5	8,5
7,5	9
7,5	9
9	10

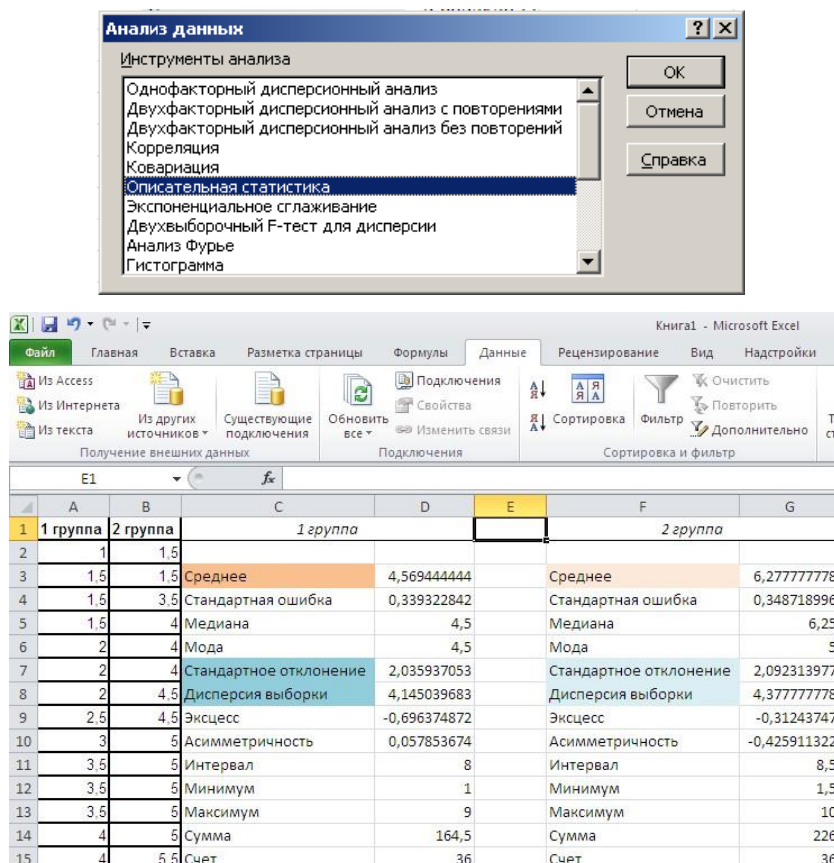
Нулевая гипотеза: правильность лечения не влияет на сроки госпитализации.

Порядок анализа в MS Excel

1. Копируем таблицу с данными в MS Excel.



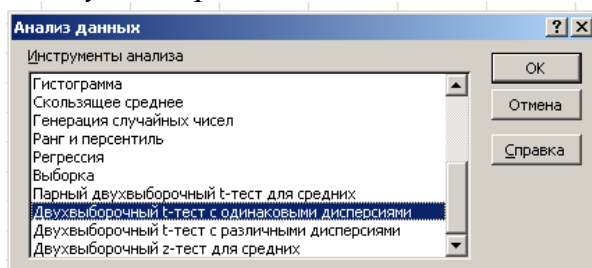
2. При помощи модуля «**Описательная статистика**» (последовательность действий вам известна) получить значения основных статистических параметров (*среднее значение, стандартное отклонение, дисперсия*) исходных данных для каждой группы. Проследите разницу между средними значениями каждой группы, определите цель анализа и то, насколько он необходим.



Средние отличаются друг от друга по величине, следовательно, при помощи *критерия Стьюдента* необходимо проверить статистическую значимость этих различий. **Нулевая гипотеза:** *отличие средних значений статистически не значимо.*

Важно, что критерий Стьюдента можно применять если дисперсии в обеих группах приблизительно равны. В нашем случае это условие выполняется.

3. Воспользуемся *двухвыборочным t-тестом для одинаковых дисперсий.*



Исходные данные:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	1 группа	2 группа											
2	1	1,5											
3	1,5	1,5											
4	1,5	3,5											
5	1,5	4											
6	2	4											
7	2	4											
8	2	4,5											
9	2,5	4,5											
10	3	5											
11	3,5	5											
12	3,5	5											
13	3,5	5											
14	4	5											
15	4	5,5											
16	4	5,5											
17	4,5	6											
18	4,5	6											

Двухвыборочный t-тест с одинаковыми дисперсиями

Входные данные

Интервал переменной 1: \$A\$1:\$A\$37

Интервал переменной 2: \$B\$1:\$B\$37

Гипотетическая средняя разность:

☒ Метки

Альфа: 0,05

Параметры вывода

☒ Выходной интервал: \$G\$1

☐ Новый рабочий лист:

☐ Новая рабочая книга

OK Отмена Справка

Результат выводится в виде таблицы. Интересующие исследователя значения выделены цветом.

Двухвыборочный t-тест с одинаковыми дисперсиями		
	1 группа	2 группа
Среднее	4,569444444	6,277777778
Дисперсия	4,145039683	4,377777778
Наблюдения	36	36
Объединенная дисперсия	4,26140873	
Гипотетическая разность средних	0	
df	70	
t-статистика	-3,511011661	
P(T<=t) одностороннее	0,000392847	
t критическое одностороннее	1,666914479	
P(T<=t) двухстороннее	0,000785693	
t критическое двухстороннее	1,994437112	

4. Из таблицы видно, что расчетное значение критерия Стьюдента (**t-статистика**) равно -3,51, что больше (по модулю) его критического значения (**t критическое двухстороннее**) равного 1,99. Вероятность ошибки $P = 0,00078$, что меньше заданного уровня значимости (альфа) равного 0,05. Следовательно, нулевая гипотеза **отвергается**. И различия средних значений показателей статистически значимы: правильность лечения влияет на длительность срока пребывания пациента в клинике.

Порядок анализа в «Statistica» 6

Условием возможности использования критерия Стьюдента для сравнения групп является **нормальность распределения значений в каждой из групп и приблизительное равенство дисперсий**, поэтому необходимо проверить распределение на соответствие нормальному распределению, например, с помощью *критерия Шапиро–Уилка* (в данной лабораторной работе проверка отсутствует, данные взяты из книги «Медико-биологическая статистика» Стентон Гланц) и с помощью *критерия Левена* оценить равенство дисперсий.

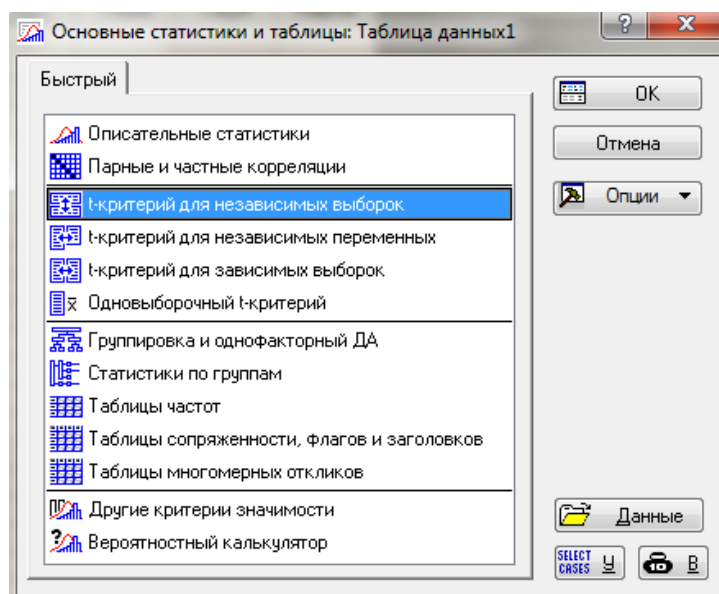
Подготовим данные, как это было сделано в предыдущей работе:

a	1
a	1,5
a	1,5
a	1,5
a	2
a	2
a	2
a	2,5
a	3
a	3,5
a	3,5
a	3,5
a	4
a	4
a	4
a	4,5
a	4,5
a	4,5
a	4,5
a	4,5
a	5
a	5
a	5
a	5,5
a	5,5
a	6
a	6
a	6
a	6,5
a	6,5
a	7
a	7
a	7,5

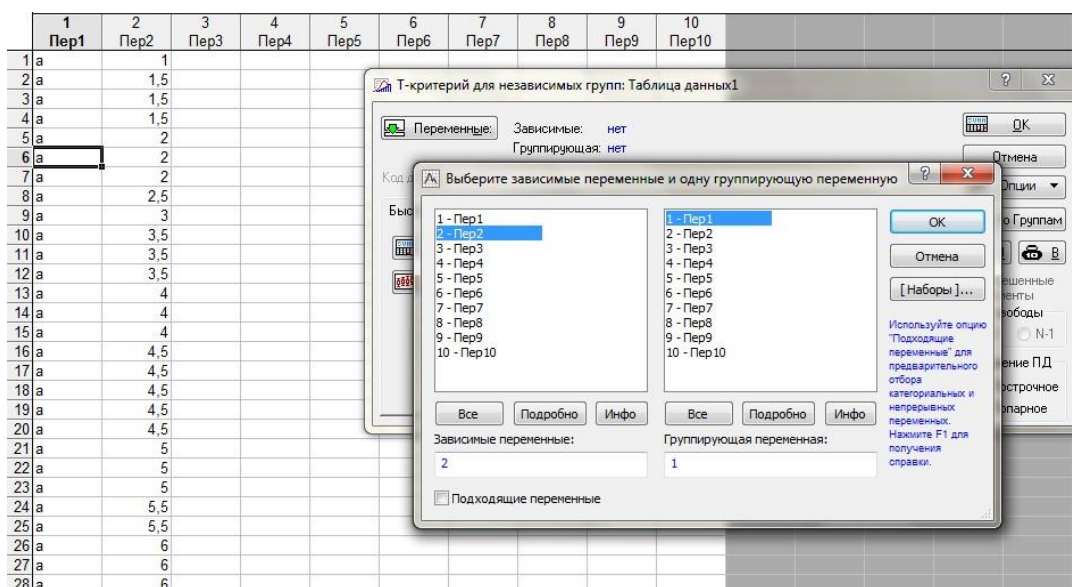
a	7,5
a	7,5
a	9
b	1,5
b	1,5
b	3,5
b	4
b	4
b	4
b	4,5
b	4,5
b	5
b	5
b	5
b	5
b	5
b	5,5
b	5,5
b	6
b	6
b	6
b	6,5
b	7
b	7
b	7
b	7,5
b	7,5
b	7,5
b	8
b	8
b	8
b	8,5
b	8,5
b	8,5
b	8,5
b	8,5
b	9
b	9
b	10

И скопируем в окно программы «Statistica».

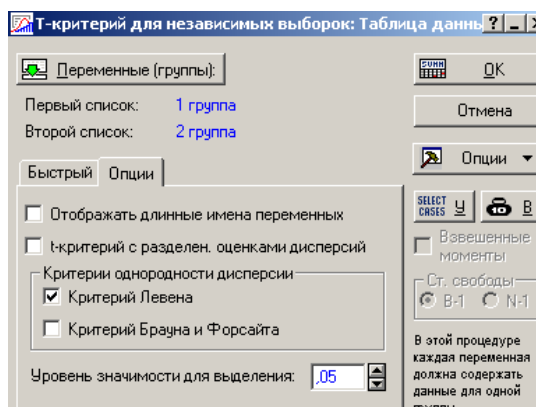
Воспользуемся критерием *Стьюдента* для двух независимых выборок: Анализ — Основные статистики и таблицы — *t*-критерий для независимых выборок.



В окне ввода данных нажимаем кнопку «Переменные (группы)» и выбираем переменные (зависимые и группирующие) как это указано на иллюстрации.



По критерию Левена необходимо проверить однородность дисперсий.



Жмем «ОК». Результаты представлены в виде таблицы:

		Т-критерий независимых выборок (Таблица данных1)													
		Замечание: Переменные рассм. как независимые выборки													
Группа 1 и Группа 2		Среднее	Среднее	t-знач.	ст. св.	p	N набл.	N набл.	Ст. откл.	Ст. откл.	F-отн.	p	Левена	Ст. св.	p
Группа 1	Группа 2	Группа 1	Группа 2				Группа 1	Группа 2	Группа 1	Группа 2	Дисперсии	Дисперсии	F(1,cc)	Левена	Левена
1 группа vs. 2 группа		4,569444	6,277778	-3,51101	70	0,000786	36	36	2,035937	2,092314	1,056149	0,872539	0,188225	70	0,665732

При $p < 0,05$ для критерия Левена следует сделать вывод о **различии дисперсий** и произвести анализ для ***t-критерия с разными оценками дисперсии***. В нашем случае $p_{\text{Левена}} = 0,6657$, что больше 0,05, и говорит о равенстве дисперсий, следовательно во внимание принимается ***t-значение*** и ***вероятность ошибки p*** для критерия Стьюдента.

Оценим полученный результат: ***t-значение (критерий Стьюдента)*** равно -3,511, что больше (по модулю) табличного значения при числе степеней свободы 70 ($t_{\text{критическое}} = 1,994$). Значение $p = 0,000786$, что значительно меньше выбранного **уровня значимости (альфа) 0,05**. Следовательно **нулевую гипотезу об отсутствии** статистически значимых различий между средними следует **отклонить**.

Вывод: Так как расчетное значение **критерия Стьюдента t** больше его **критического значения** при **уровне значимости (альфа) равном 0,05**, то нулевая гипотеза **отвергается**. **Вероятность справедливости нулевой гипотезы P** меньше выбранного уровня значимости. Таким образом, можно говорить, что правильность лечения влияет на длительность срока пребывания пациента в клинике.

✓Задание

Кокаин чрезвычайно вреден для сердца, он может вызвать инфаркт миокарда даже у молодых людей без атеросклероза. Кокаин сужает коронарные сосуды что приводит к уменьшению притока крови к миокарду, кроме того, он ухудшает насосную функцию сердца. **Нифедипин** (препарат из группы антагонистов кальция) обладает способностью расширять сосуды, его применяют при ишемической болезни сердца. Ш. Хейл и соавт. предположили, что **нифедипин** можно использовать и при поражении сердца, вызванном кокаином. Собакам вводили кокаин, а затем **нифедипин** либо **физиологический раствор**. Показателем насосной функции сердца служило среднее артериальное давление. Были получены следующие данные:

Среднее артериальное давление после приема кокаина, мм рт. ст.

Плацебо	Нифедипин
156	73
171	81
133	103
102	88
129	130
150	106
120	106
110	111
112	122
130	108
105	99

Влияет ли нифедипин на среднее артериальное давление после приема кокаина?

Контрольные вопросы

1. Что такое гипотеза?
2. Что такое нулевая гипотеза?
3. Что называется уровнем значимости?
4. Какие значения уровня значимости наиболее часто встречаются в медицинских исследованиях?
5. Что значит «ошибочно отвергнуть верную нулевую гипотезу»?
6. Что подразумевается, когда говорят, что уровень значимости равен 0,05?
7. Условия применения критерия Стьюдента.
8. Алгоритм применения критерия Стьюдента.
9. Что такое критическое значение критерия Стьюдента?
10. Ошибки применения критерия Стьюдента.
11. Если расчетное значение критерия t больше чем $t_{\text{критическое}}$ при заданном уровне значимости, это означает, что...?

Лабораторная работа № 5

Анализ зависимостей. Корреляционный и регрессионный анализ. Парная корреляция

Краткие сведения из теории

Часто исследователя интересует возможность предсказать поведение (принимаемые значения) одной переменной в зависимости от поведения (принимаемых значений) другой.

В качестве простейшего примера можно привести взаимосвязь между ростом человека и его весом: при увеличении роста значение веса также возрастает.

Зная характер зависимости, исследователь в процессе практического применения данных эксперимента опирается уже не только на свой личный опыт или опыт своих коллег, или же интуитивное знание, но и на знания, подтвержденные достаточно точными методами статистического анализа, сведенные в таблицы и представленные графиками.

Если вернуться к примеру с ростом и весом человека, то изучив предполагаемую зависимость, исследователь с достаточной достоверностью может говорить о том, что при определенном значении роста среднестатистического человека вес его будет иметь соответствующее значение, незначительно отклоняющееся в ту или иную сторону. Другим примером может являться исследование воздействия дозы

препарата на какие-либо определенные показатели состояния организма человека. Исследователю необходимо проверить, как ведет себя организм (наблюдаемые показатели) при увеличении дозы препарата. При обнаружении зависимости дальнейшими действиями является определение ее (зависимости) характера и проверка на практике, что, при положительном результате, позволяет достаточно точно говорить о том, что, например, увеличив дозу препарата, врач с большой вероятностью будет видеть **соответствующие** изменения показателей здоровья или, опираясь на **исходные показатели** состояния организма, подбирать соответствующую дозу препарата.

Выявление и измерение связи между признаками, характеризующими изучаемые явления или процессы, является важнейшей частью исследования.

Виды связей между переменными

Различают **функциональную** и **корреляционную** связи. При наличии **функциональной** связи изменение величины одного признака неизбежно вызывает совершенно определенные изменения величины другого признака.

Примером такой связи может служить однозначная зависимость площади круга от его радиуса. При конкретном значении радиуса величина площади круга всегда принимает строго конкретное и известное значение.

Функциональная связь между явлениями присуща неживой природе. В биологических науках (и в медицинских исследованиях) чаще приходится иметь дело со связью между явлениями, когда одной и той же величине одного признака соответствует ряд варьирующих (разных, но близких по величине) значений другого признака, что обусловлено чрезвычайным многообразием взаимодействия различных явлений живой природы.

В примере с ростом и весом: одному и тому же значению роста могут соответствовать разные, но чаще всего близкие по величине значения веса.

Такого рода связь носит название **корреляционной** (correlation — соответствие, соотносительность). В то время как функциональная связь имеет место в каждом отдельном наблюдении, корреляционная связь проявляется только при многочисленном сопоставлении признаков. Исследователю следует помнить, что *обнаружение корреляции между сопоставляемыми явлениями не говорит еще о существовании однозначной причинной связи между ними.*

Другими словами, имея дело с подобными явлениями, исследователь не может однозначно определить вид (формулу) связи между ними, как это выглядит при функциональной зависимости. Возвращаясь к примеру с ростом и весом: исследователь не может определить математическую формулу, при подстановке в которую значения роста, он получил бы совершенное определенное и единственное значение веса; каждому определенному значению роста будет соответствовать несколько близких по величине значений веса, при условии, что выборка однородна. Часто исследователь видит имеющуюся взаимосвязь, но предсказать точное (как при вычислении математического уравнения) поведение объекта при изменении параметров другого объекта (даже при явно замеченной взаимосвязи) исследователь не может. Причиной может являться как сложные механизмы самого взаимодействия, так и влияющие факторы, о которых исследователь не знает или не может исключить из эксперимента или повседневного практического опыта.

Популярный пример

Популярным примером корреляционной зависимости является замеченная взаимосвязь между частотой трелей сверчка и температурой окружающей среды: когда на улице холодно, сверчки поют не так часто.

Еще один пример корреляции — набор кадров в полицию. Часто оказывается, что количество преступлений (на душу населения) связано с количеством полицейских на данной территории.

Учитывая вышеуказанную особенность корреляционной связи, исследователь вынужден использовать приблизительную характеристику зависимости, каковая не становится менее значимой в силу своей неточности, с точки зрения зависимости функциональной.

Регрессионный и корреляционный анализы

Анализ предполагаемой зависимости имеет поэтапную структуру и состоит из следующей последовательности действий:

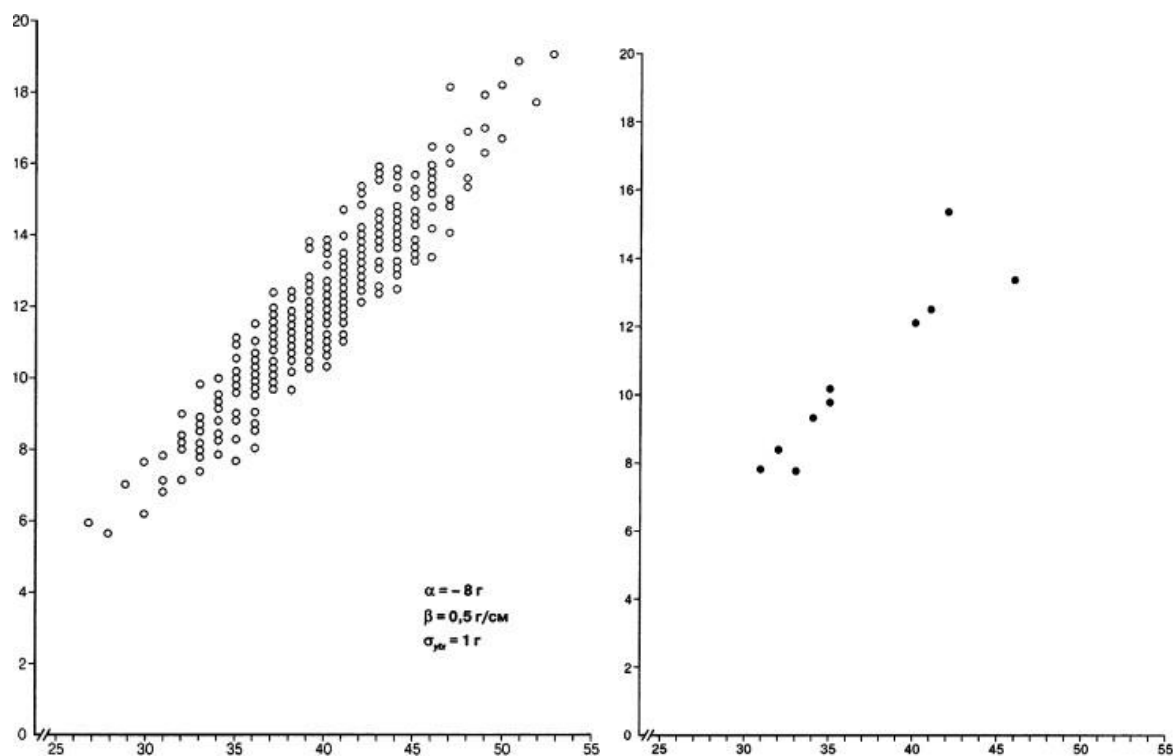
➤ Этапы статистического анализа зависимостей:

- Сбор данных.
- Внесение данных в таблицы (попарно). Каждому значению одной переменной соответствует определенное значение другой.
- Первичный анализ предполагаемой зависимости.
- Исключение из выборки артефактов (выбросов), если таковые имеют место быть.

- Регрессионный анализ.
- Корреляционный анализ.
- Проверка соответствия модели экспериментальным данным.

Первичный анализ

Первичный анализ предполагаемой связи подразумевает нанесение результатов эксперимента на *график*. Уже по первичному анализу можно судить о наличии какой-либо взаимосвязи.



Такой выборка (рисунок справа) представляется исследователю, который не может наблюдать всю совокупность (рисунок слева).

Из графика видно как может проявляться взаимосвязь одной переменной (зависимой) от другой (независимой), также можно проследить характер зависимости: линейная, криволинейная; прямая (при увеличении одной переменной увеличивается и другая) или обратная (при увеличении одной переменной, другая уменьшается); сильная или слабая.

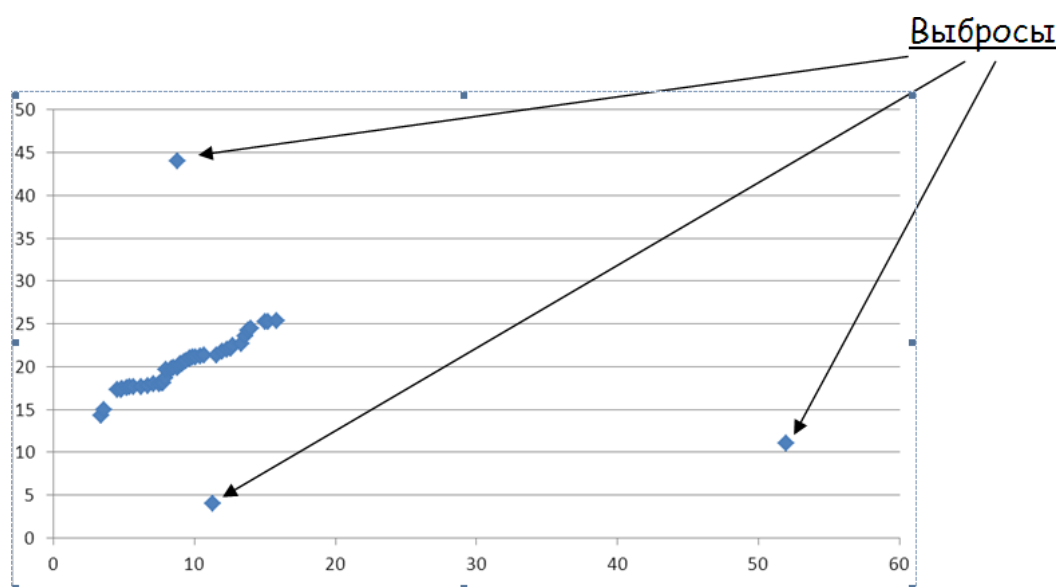
- ✓ Как вы считаете, как можно охарактеризовать взаимосвязь на приведенном выше графике:
 - ✓ линейная или криволинейная?
 - ✓ прямая или обратная?
 - ✓ сильная или слабая?

Выбросы

В силу возможных ошибок эксперимента в таблицу результатов могут попасть данные явно нетипичные для конкретного эксперимента — **выбросы**.

По определению, **выбросы** являются нетипичными, резко выделяющимися наблюдениями. Единичный выброс способен существенно изменить дальнейшие расчеты по характеристике предполагаемой зависимости. Обычно считается, что выбросы представляют собой случайную ошибку, которую следует контролировать. К сожалению, не существует общепринятого метода автоматического удаления выбросов. Чтобы не быть введенными в заблуждение полученными значениями, необходимо проверить на диаграмме рассеяния каждый случай, характеризующийся как выброс.

Выбросы могут не только искусственно усилить впечатление о взаимосвязи, но также реально уменьшить существующую корреляцию.



Выбросы исключаются из выборки попарно.

На графике видно, что имеется явная взаимосвязь между переменными, причем взаимосвязь сильная, прямая, линейная, однако подозрение вызывают некоторые (указанные) точки, явно выделяющиеся из общего характера связи между переменными. Эти пары данных можно обозначить как **выбросы** и исключить из таблицы результатов попарно.

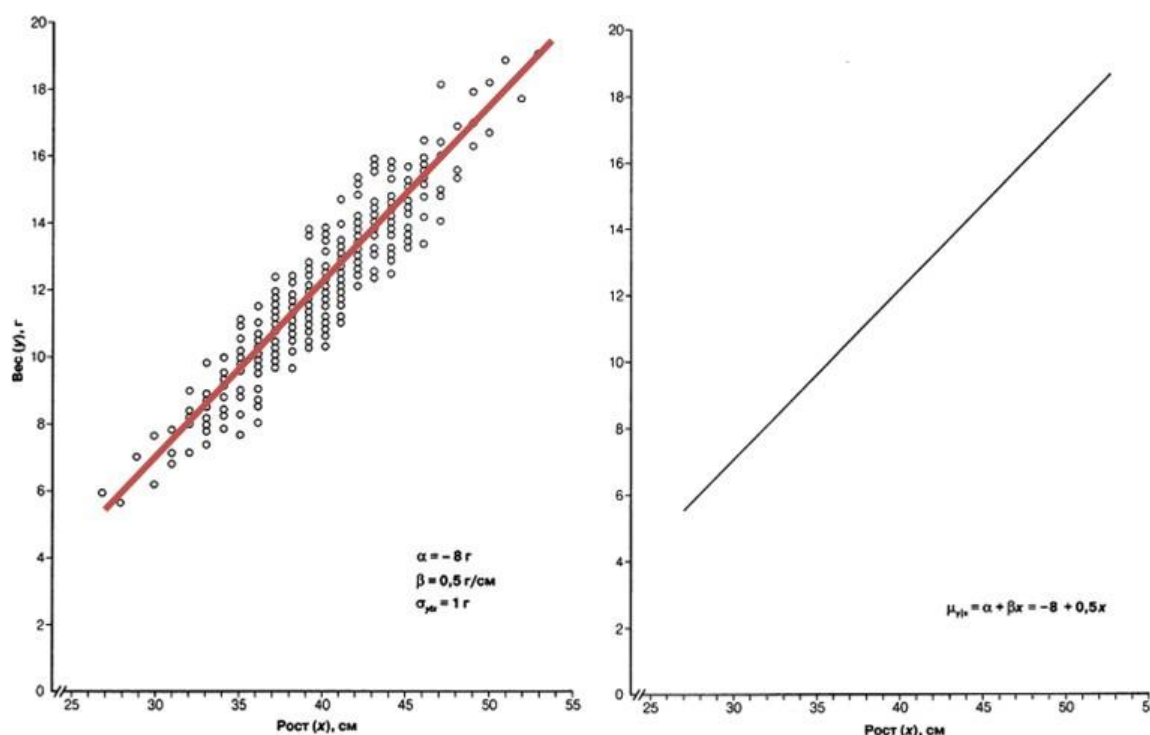
Регрессионный анализ

Регрессионный анализ предполагает построение кривой регрессии, построенной методом наименьших квадратов и описывающей имеющуюся зависимость уравнением кривой с определенной точностью.

Другими словами, весь массив данных (точек на графике) мы заменяем уравнением (прямой или кривой линией), характеризующим наличествующую связь.

Расположение линии определяется **методом наименьших квадратов**. Метод наименьших квадратов. Сумма квадратов расстояний (вычисленных по оси Y) от наблюдаемых точек до прямой является минимальной.

В уравнении регрессии одна из переменных, x , называется **независимой переменной**, а другая, y , — **зависимой**. Это не означает, что одна переменная однозначно определяет другую. Просто *по значению одного признака предсказывает значение второго*. В условиях эксперимента произвольно изменяется независимая переменная и отслеживается, как меняется зависимая. При достаточном количестве данных эксперимента можно делать выводы о наличии взаимосвязи и адекватности полученной модели.



Прямая и уравнение регрессии

Пример уравнения регрессии (прямая линия):

$$y = a + bx .$$

Корреляционный анализ

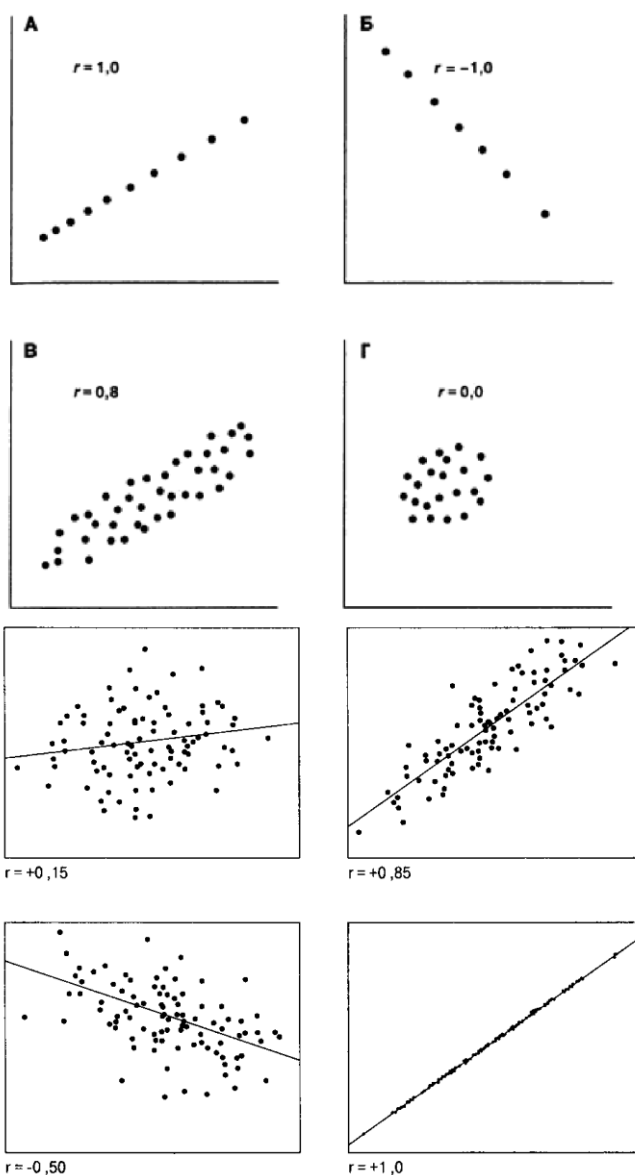
Корреляционный анализ. Исследователя может интересовать не только предсказание поведения одной переменной по значению другой, но и просто *характеристика тесноты (силы) связи между ними, при этом выраженная одним числом*. Эта характеристика называется **коэффициентом корреляции**, обычно его обозначают буквой **r**. **Корреляция** означает, что между двумя числовыми переменными наблюдается определенная линейная взаимосвязь.

Коэффициент корреляции может принимать значения от -1 до +1.

Знак коэффициента корреляции показывает направление связи (прямая или обратная), а абсолютная величина — тесноту связи.

Коэффициент, равный -1 , определяет столь же **жесткую** связь, что и равный 1 . В **отсутствии** связи коэффициент корреляции равен **нулю**. Промежуточные значения ($0,8$; $-0,5$; $0,25$) указывают на **наличие** связи в большей или меньшей степени, при этом, чем ближе значение по модулю к нулю, тем менее тесная связь между переменными. Знак коэффициента указывает на **направление** связи: « $-$ » — обратная связь, « $+$ » — прямая.

Коэффициент корреляции Пирсона предназначен для описания линейной связи количественных признаков; как и регрессионный анализ, он требует нормальности распределения.



✓ Охарактеризуйте силу взаимосвязи между переменными на вышеприведенных графиках

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

После построения модели необходимо тщательно проверить ее надежность. В процессе регрессионного анализа точность модели оценивается с помощью дисперсионного анализа, коэффициенты уравнения регрессии оцениваются при помощи **критерия Стьюдента**.

Уровень значимости (p), вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции. Значимость определенного коэффициента корреляции зависит от объема выборок. Критерий значимости основывается на предположении, что распределение остатков (т. е. отклонений наблюдений от регрессионной прямой) для зависимой переменной Y является нормальным (с постоянной дисперсией для всех значений независимой переменной X).

Порядок анализа в MS Excel

Корреляционный анализ

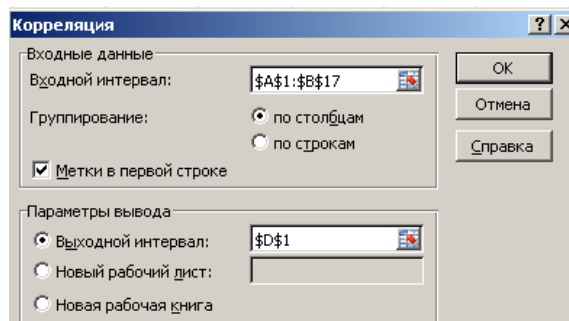
Выполнение корреляционного анализа покажем на примере результатов измерения массы (Y) и длины туловища (X) исследуемой группы животных. Исходные данные представлены в таблице (в табличном редакторе **MS Excel** данные представлены двумя столбцами). Скопируйте любую пару значений (X и Y) на «Лист 1» программы MS Excel.

Исходные данные

X	Y
3,4	14,3
3,6	14,9
4,5	17,3
4,8	17,3
4,9	17,4
5,2	17,5
5,4	17,6
5,7	17,6
6,2	17,6
6,7	17,8
7,1	18
7,5	18
7,7	18,1
7,8	18,1
7,9	18,6
8	19,7

Откройте модуль «Анализ данных», выберите опцию «Корреляция», после чего щелкните мышкой «ОК».

В появившемся окне выполните операции и установки, как показано на рисунке:



Стартовая панель (ваши значения в ячейке **входной** и **выходной** интервал могут отличаться, **метки в первой строке** ставятся при условии, что входной интервал включает заголовки таблицы X и Y).

Щелкните мышкой «ОК». Результат обработки появится в указанном поле (выходной интервал \$E\$1).

Результат обработки:

	X	Y
X	1	
Y	0,85	1

В полученной таблице нас интересует значение в ячейке на пересечении X и Y — 0,85. Это и есть значение *коэффициента корреляции*.

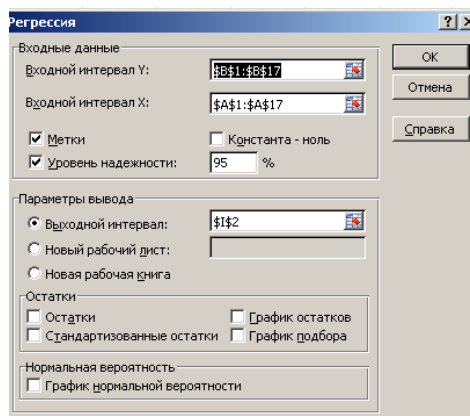
✓ О чем говорит данное полученное числовое значение коэффициента корреляции?

Регрессионный анализ

Для выполнения регрессионного анализа использовались исходные данные таблицы.

1. Откройте модуль «Анализ данных» и выберите опцию «Регрессия», после чего щелкните мышкой «ОК».

2. В появившемся окне выполните операции и установки, как показано на рисунке ниже.



Стартовая панель регрессионного анализа

3. Щелкните мышкой «ОК». Результат обработки появится в указанном поле (выходной интервал \$I\$1).

Результат обработки:

Параметры	Значения
Множественный R	0,847
R-квадрат	0,719
Нормированный R-квадрат	0,699
Стандартная ошибка	0,701
Наблюдения	16

Дисперсионный анализ

Параметры	df	SS	MS	F	Значимость F
Регрессия	1	17,59	17,59	35,759	0,000..
Остаток	14	6,887	0,492		
Итого	15	24,478			

Регрессионный анализ

Параметры	Коэффициенты	Стандартная ошибка	t-статистика	P-значение
Y-пересечение	13,289	0,724	18,365	0,000..
X	0,697	0,117	5,98	0,000..

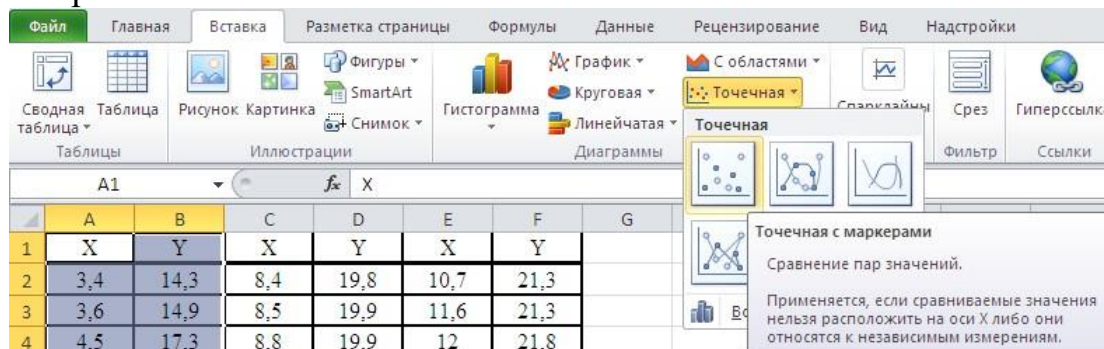
Таким образом, корреляционная связь между массой и длиной туловища исследуемой группы животных характеризуется высоким ($r = 0,85$) и достоверным коэффициентом корреляции (достоверность проверяется критерием Фишера, из таблицы: критерий Фишера $F = 35,759$ при уровне значимости F существенно меньше 0,05). Получена очень надежная регрессия, о чем свидетельствует t-статистика из таблицы (уровень значимости (значимость F и p-значение) существенно меньше 0,05).

Уравнение линейной регрессии Пирсона и коэффициент корреляции:

$$Y = 13,289 + 0,697 \times X; r = 0,85.$$

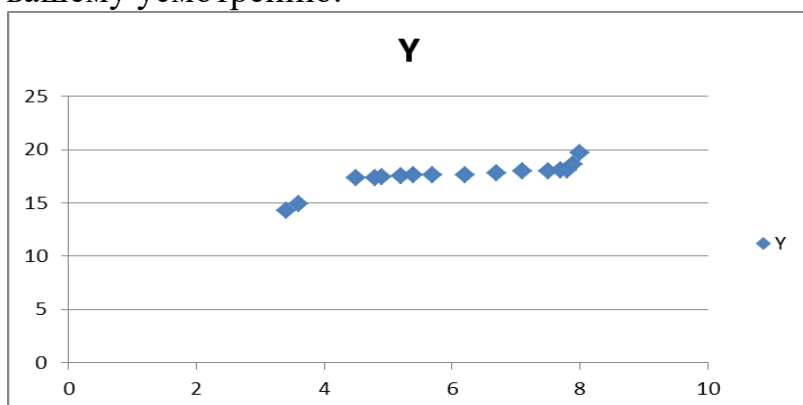
С менее подробной информацией операции регрессии и корреляции можно выполнить в системе MS Excel, используя модуль «Мастер диаграмм».

1. В системе MS Excel откройте модуль «Мастер диаграмм» (Вставка-Диаграммы-Точечная — для Excel 2010), предварительно выделив значения переменных X и Y.

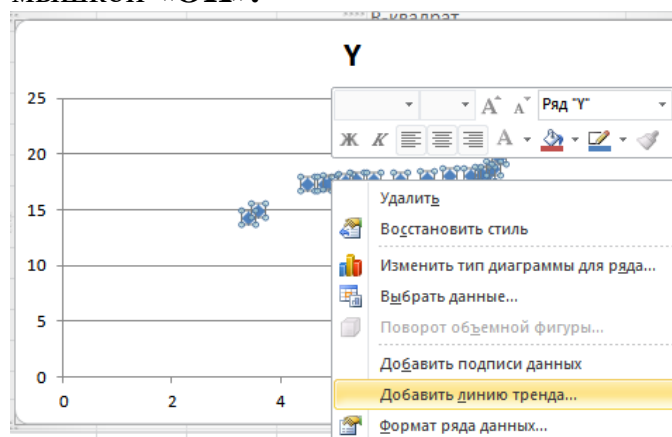


Вставка — Диаграммы — Точечная — для Excel 2010

2. Выберите «Тип» диаграммы «Точечная».
3. Оформите график, используя инструменты по работе с диаграммами согласно вашему усмотрению.



4. На графике щелкните правой кнопкой по любой точке диаграммы.
5. Выберите опцию «Добавить линию тренда» и «Тип — линейная». Щелкните мышкой «ОК».



Формат линии тренда

Параметры линии тренда

Построение линии тренда (аппроксимация и сглаживание)

☐ Экспоненциальная

☒ Линейная

☐ Логарифмическая

☐ Полиномиальная Степень: 2

☐ Степенная

☐ Линейная фильтрация Точки: 2

Название аппроксимирующей (сглаженной) кривой

☒ автоматическое: Линейная (Y)

☐ другое:

Прогноз

вперед на: 0,0 периодов

назад на: 0,0 периодов

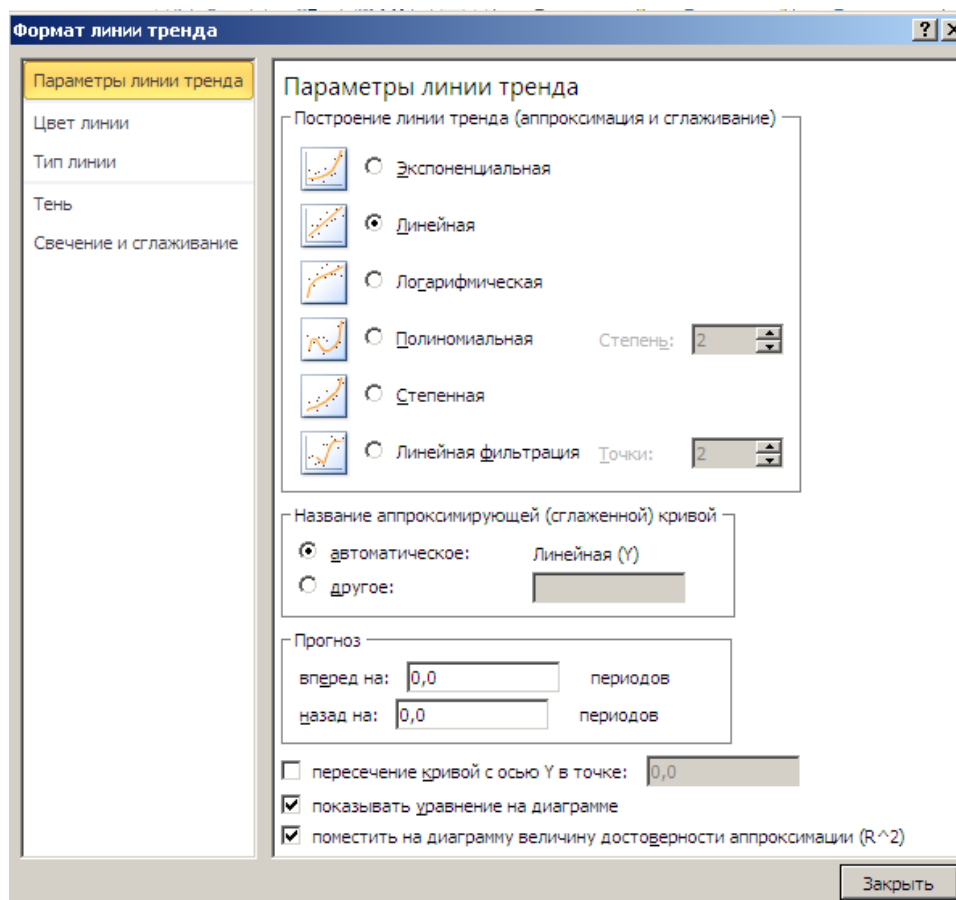
☐ пересечение кривой с осью Y в точке: 0,0

☒ показывать уравнение на диаграмме

☒ поместить на диаграмму величину достоверности аппроксимации (R^2)

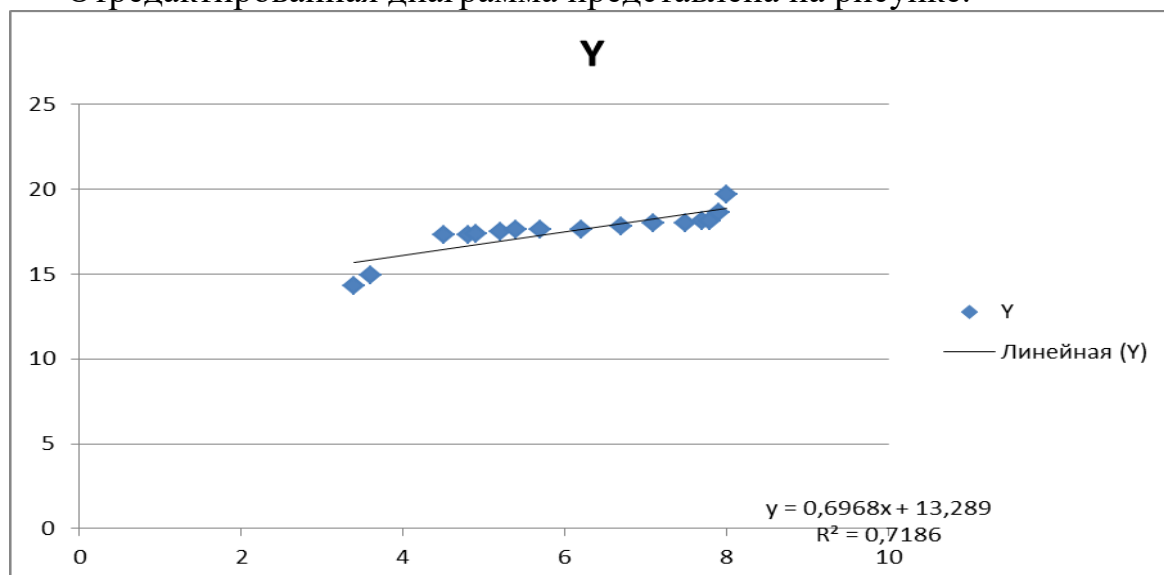
Заккрыть

6. Выберите установки, как показано на рисунке. Щелкните мышкой «ОК».



Тип аппроксимации кривой и параметры кривой (Excel 2010)

Отредактированная диаграмма представлена на рисунке.



Отредактированная диаграмма с уравнением регрессии и коэффициентом детерминации

Уравнение регрессии и коэффициент детерминации R^2 находятся в правом нижнем углу диаграммы. Как видно, что они такие же, как и при выполнении регрессионного анализа в пакете «Анализ данных — Регрессия».

Порядок анализа в «Statistica» 6

Корреляционный и регрессионный анализ

Исходные данные необходимо взять из предыдущей части работы (для сравнения результатов обработки рекомендуется использовать те же значения), которые из табличного редактора **MS Excel** копируются в таблицу модуля «**Statistica**» 6. После копирования измените названия переменных на **X** и **Y** соответственно.

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

Исходные данные

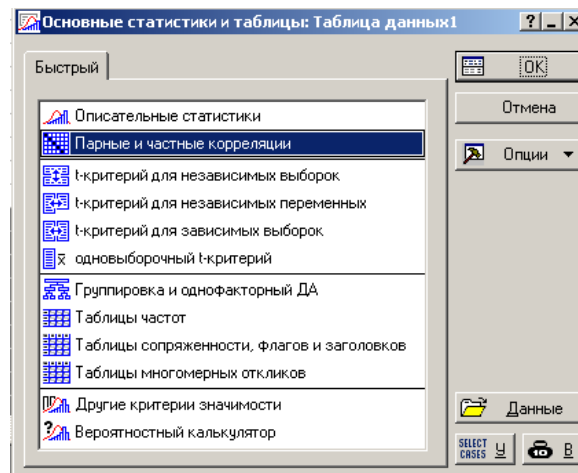
X — независимая переменная

Y — зависимая переменная

Проведем анализ в модуле «**Основные статистики и таблицы**». Исследуем предполагаемую связь между **X** и **Y**.

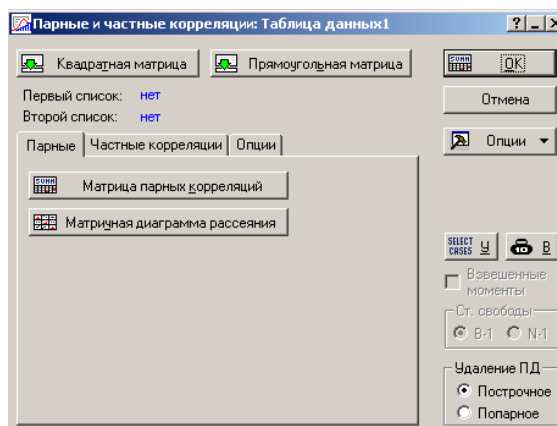
1. Из Переключателя модулей «Statistica» откройте модуль «**Основные статистики и таблицы**».

2. Щелкните мышью по названию «Correlation matrices» (*Парные и частные корреляции*).



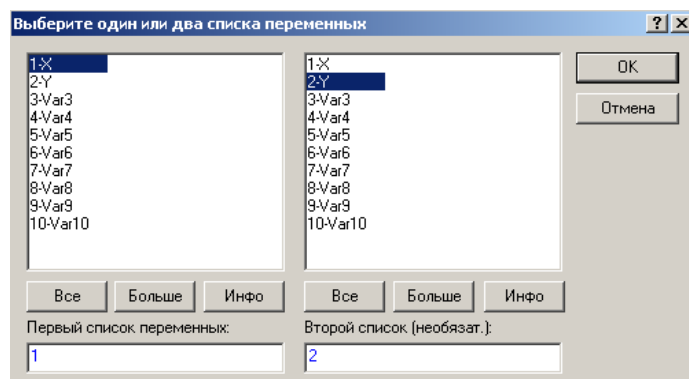
Панель запуска

3. Выберите переменные для анализа. Выбор переменных осуществляется с помощью кнопки **«Прямоугольная матрица»**, находящейся в центре верхней части панели.



Стартовая панель модуля «Парные и частные корреляции»

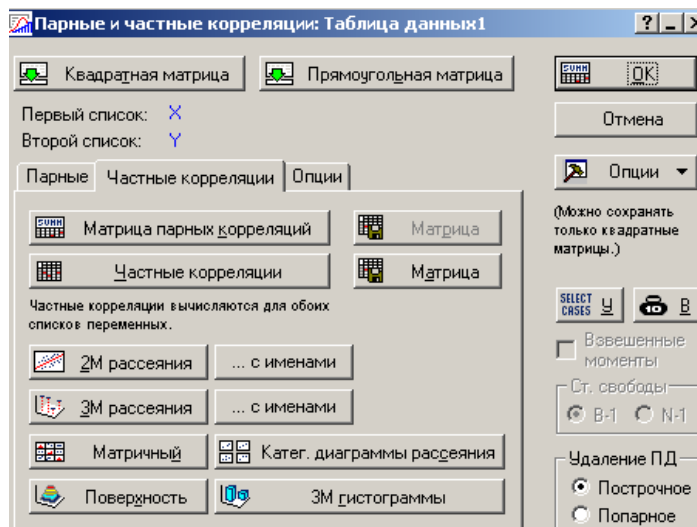
После того, как кнопка будет нажата, диалоговое окно **«Выбрать списки зависимых и независимых переменных»** появится на вашем экране.

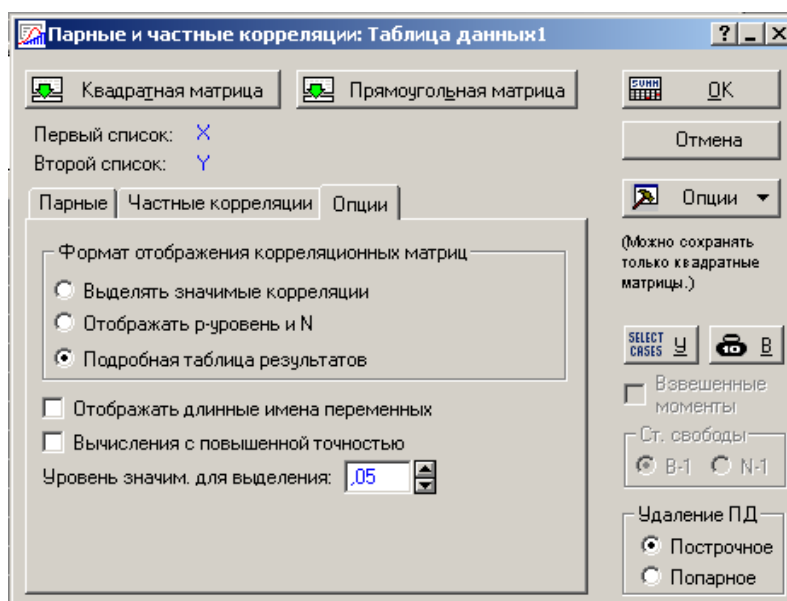


Окно выбора переменных для анализа

4. Высветив имя переменной в правой части окна, выберите переменную в левой части окна.

После нажатия кнопки **«ОК»** выполните установки, показанные на рисунке ниже.





Окно предварительных установок

5. После нажатия кнопки «ОК» программа произведет расчет корреляции между X и Y , и на экране появится окно результатов.

Корреляции (Таблица данных1)											
Отмеченные корреляции значимы на уровне $p < ,05000$											
(Построчное удаление ПД)											
Пер. X и Пер. Y	Среднее	Стд. откл.	$r(X, Y)$	r^2	t	p	N	Св. член завис. Y	Наклон завис. Y	Св. член завис. X	Наклон завис. X
X	6,02500	1,554134									
Y	17,48750	1,277432	0,847730	0,718647	5,979923	0,000034	16	13,28929	0,696798	-12,0108	1,031355

Результат расчета корреляции

На рисунке представлена следующая информация:

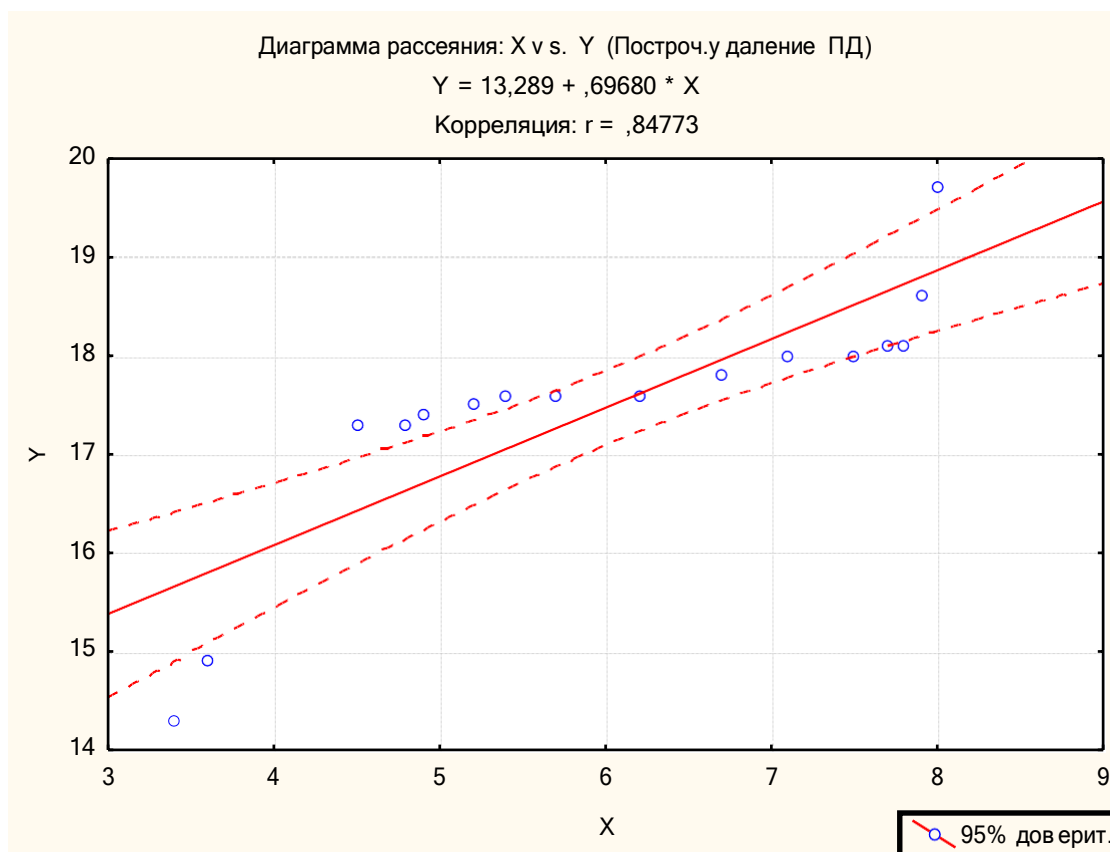
- среднее значение;
- стандартное отклонение;
- значение коэффициента корреляции r ;
- значение коэффициента детерминации r^2 ;
- t — критерий;
- p — уровень значимости;
- число коррелируемых пар;
- 13,289 — свободный член уравнения регрессии.
- 0,696 — коэффициент при независимой переменной уравнения регрессии.

В этом примере $r = 0,847$. Это очень высокое значение (подсвечено красным цветом), показывающее, что построенная регрессия объясняет более 90 % разброса значений переменной X относительно среднего.

Из таблицы видно, что оцененная модель имеет вид:

$$Y = 13,289 + 0,696 \cdot X; \quad r = 0,847$$

6. Вернувшись обратно в окно Предварительные установки (*Парные и частные корреляции* (внизу слева)) нажать кнопку «**2М рассеяния**» появится график, на котором данные с подогнанной прямой имеют вид:



Линейная регрессия для данных X и Y

Красной пунктирной линией на рисунке изображен **доверительный интервал** для линии регрессии.

Таблицу результатов с заголовками и график скопировать в отдельный документ Word и сделать выводы о результатах анализа.

✓Задание

Исследуя проницаемость сосудов сетчатки, Дж. Фишман и соавторы решили выяснить, связан ли этот показатель с электрической активностью сетчатки. Позволяют ли полученные данные говорить о существовании связи?

Электрическая активность сетчатки (независимая переменная)	Проницаемость сосудов сетчатки (зависимая переменная)
0	19,5
38,5	15
59	13,5
97,4	23,3
119,2	6,3
129,5	2,5
198,7	13
248,7	1,8
318	6,5
438,5	1,8

1. Произвести корреляционный и регрессионный анализ данных в любой на ваш выбор программе.
2. Построить график зависимости и линию регрессии.
3. Определить и записать уравнение регрессии.
4. Исходя из величины коэффициента корреляции, сделать выводы о силе зависимости между переменными и ее направлении.
5. Сделать выводы о значимости корреляции.
6. Получившийся результат скопировать в документ Word и сделать выводы о результатах анализа.

Контрольные вопросы

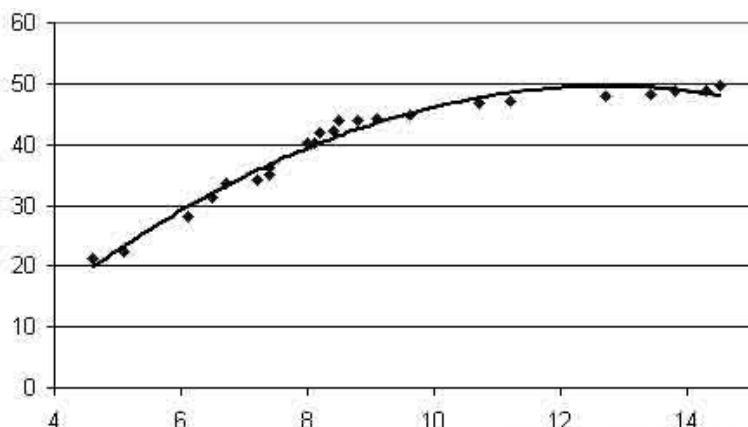
1. Какие цели преследуются при изучении зависимости между переменными?
2. Какие виды связей между переменными вы знаете?
3. Что означает функциональная зависимость?
4. Что означает корреляционная связь?
5. Приведите примеры корреляционной связи между переменными.
6. Что означает коэффициент корреляции Пирсона?
7. Приведите примеры графиков зависимостей между переменными с разными коэффициентами корреляции.
8. Принцип регрессионного анализа?
9. Объясните смысл уравнения регрессии и линии регрессии.
10. Что означает уровень значимости корреляции?

Лабораторная работа № 6

Криволинейная корреляция и регрессия

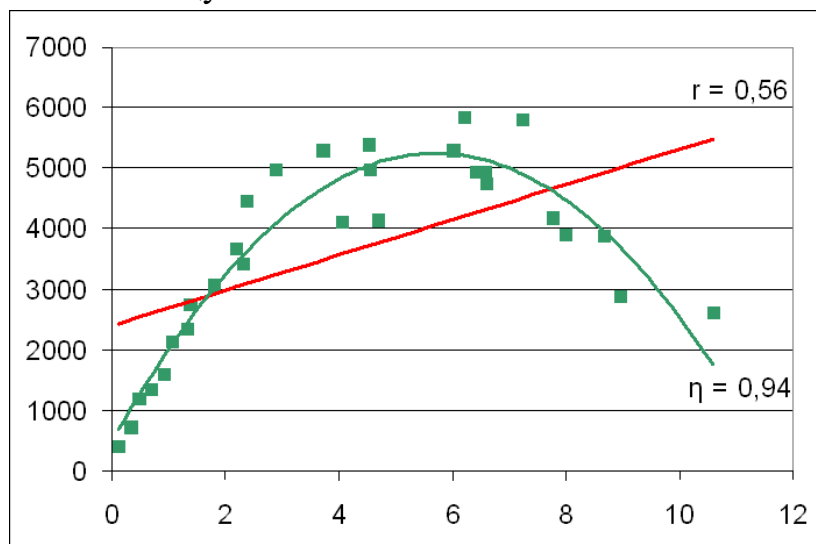
Краткие сведения из теории

Если связь между изучаемыми явлениями существенно отклоняется от пропорциональной (носит характер не прямой а кривой линии, см. рисунок), что легко установить по графику, то коэффициент корреляции непригоден в качестве меры связи.



В этом случае коэффициент корреляции может указать на отсутствие сопряженности («криволинейности») там, где налицо сильная криволинейная зависимость.

Естественно и коэффициент корреляции, рассчитанный исходя из предполагаемой линейной зависимости, как и сама прямая линия не будут характеризовать имеющуюся зависимость:



Поэтому необходим новый показатель, который правильно измерял бы степень криволинейной зависимости. Таким показателем является **корреляционное отношение**, обозначаемое греческой буквой η (эта).

Корреляционное отношение измеряет степень корреляции при любой ее форме.

Корреляционное отношение измеряет степень криволинейных и прямолинейных связей.

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

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

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

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

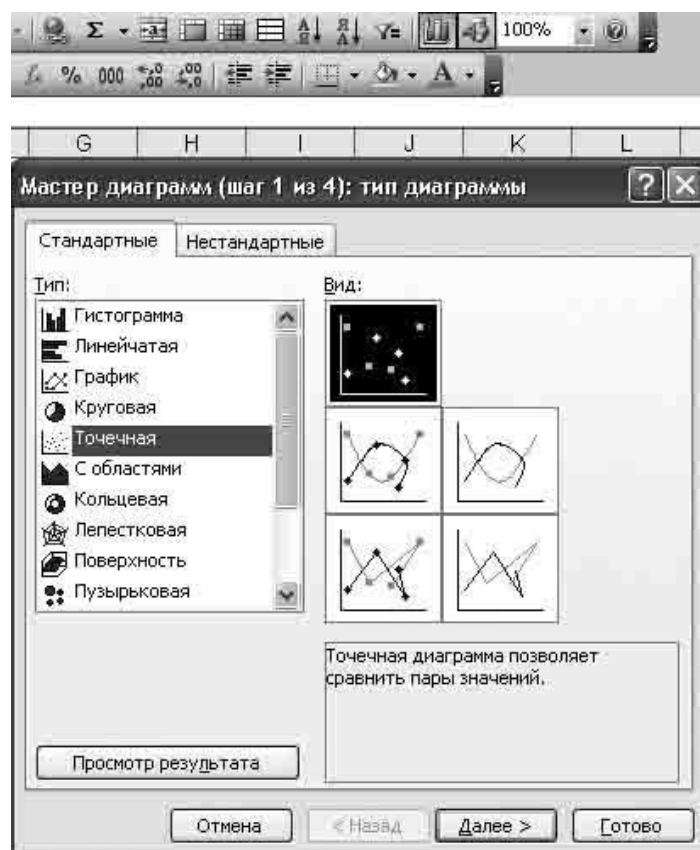
Задачей исследователя является подобрать вид функции, которая бы наиболее четко ложилась на поле регрессии, иначе: значение квадрата корреляционного отношения было бы максимально возможным.

Проведение анализа в MS Excel

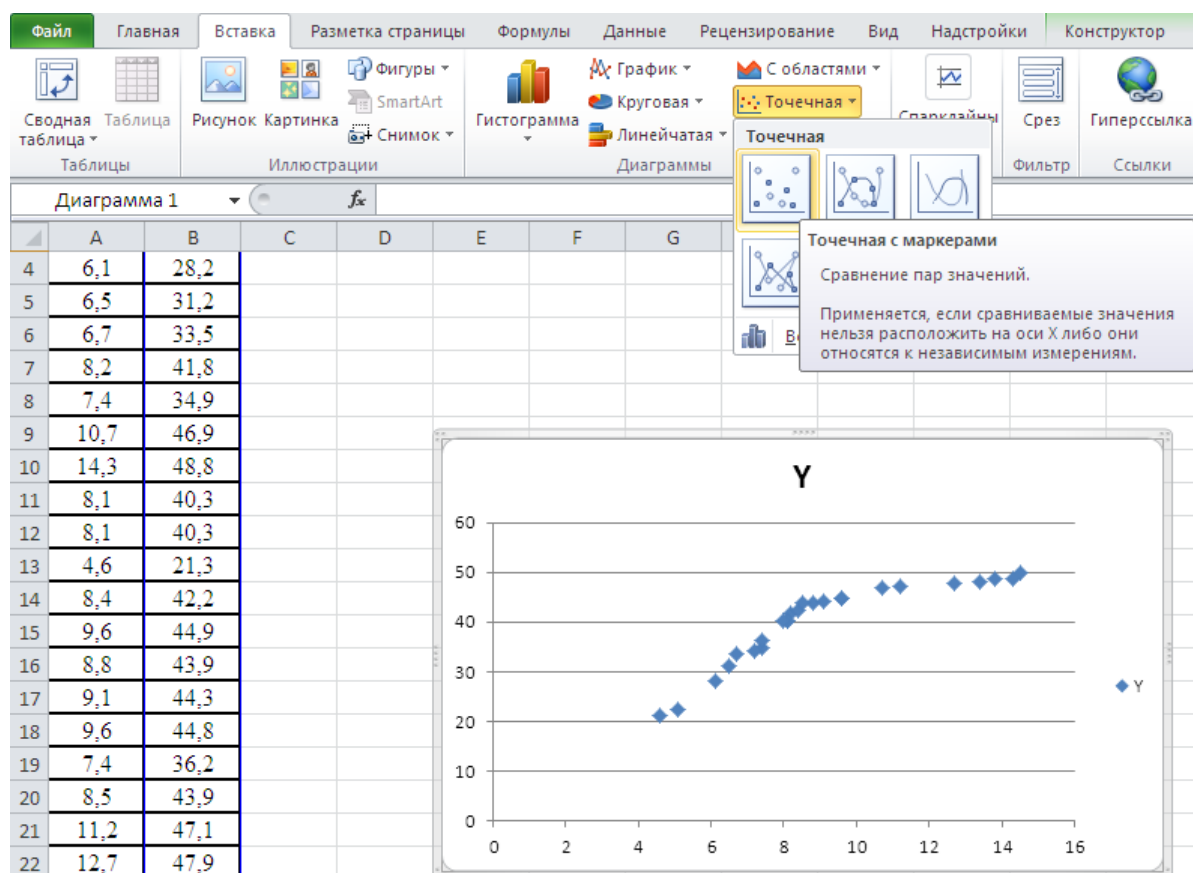
Исходные данные представлены в таблице:

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

1. Скопируйте исходные данные на «Лист 1» книги MS Excel. Выделите все значения переменных (всю таблицу с заголовками) и постройте график (тип *диаграммы* — «Точечная»):

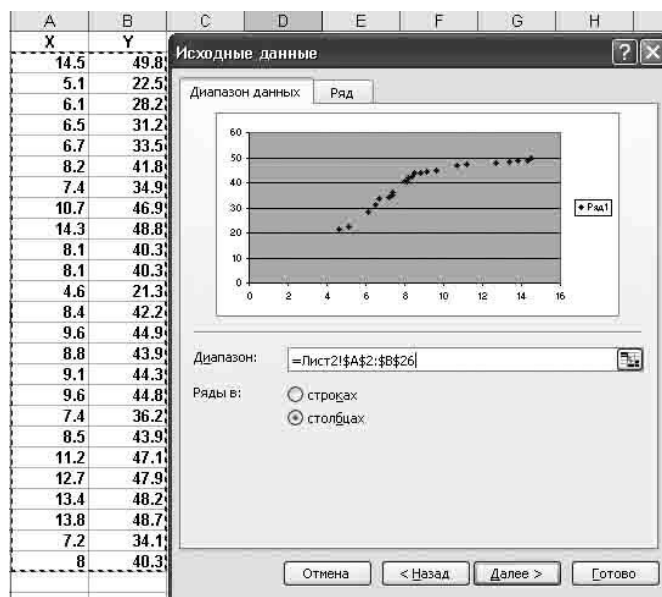


Выбор типа и вида диаграммы (Excel 2003)



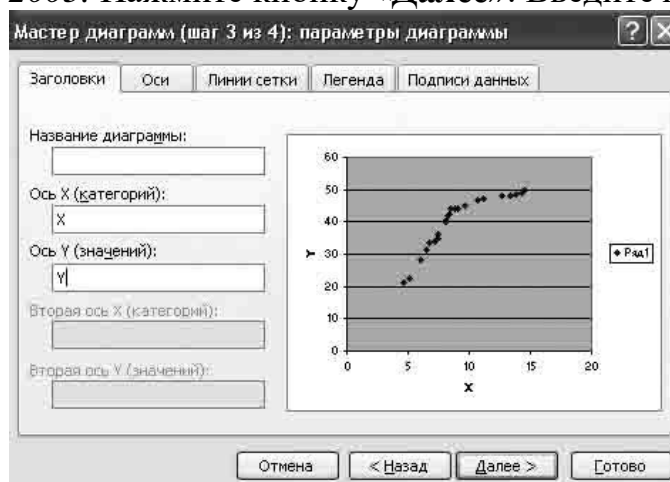
Выбор типа и вида диаграммы (Excel 2010)

2. Для *Excel 2003*. Нажмите кнопку «Далее». В окно «Диапазон» установите маркер, выберите в таблице анализируемые выборки.



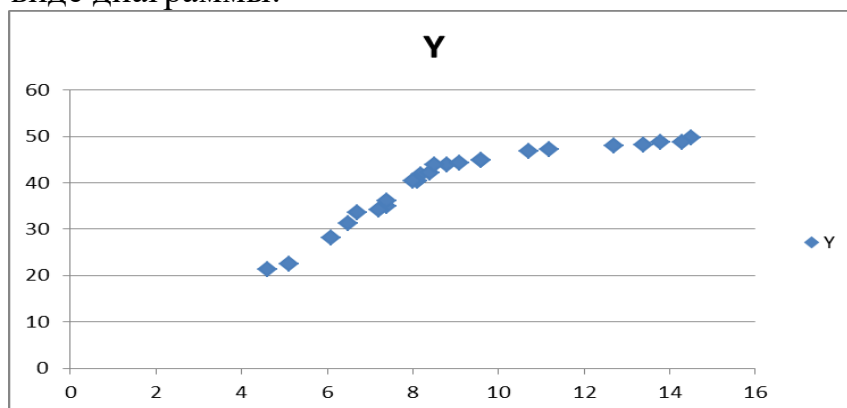
Установка диапазона *Excel 2003*

3. Для *Excel 2003*. Нажмите кнопку «Далее». Введите названия осей.

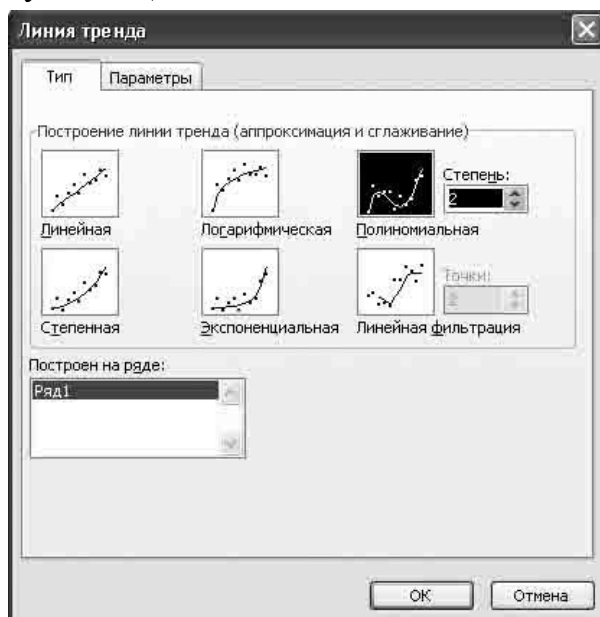


Оформление диаграммы

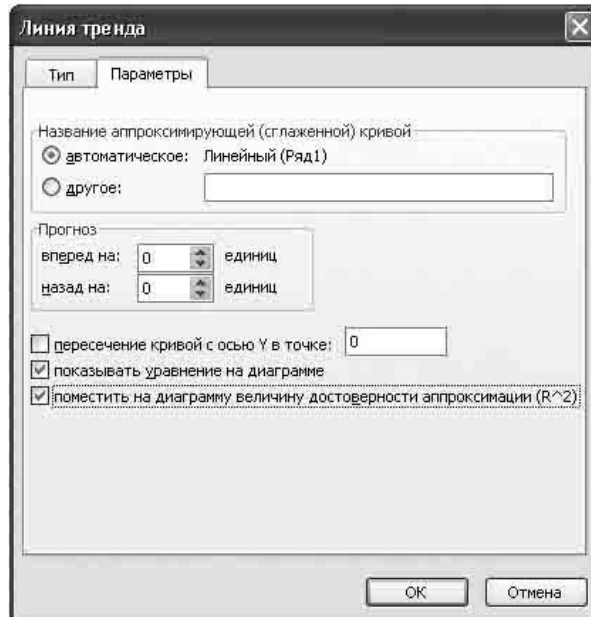
4. Для *Excel 2003*. Щелкните мышкой **Готово**. Результат обработки появится в виде диаграммы:



5. На графике щелкните правой кнопкой по любой точке диаграммы.
6. Выберите опцию «Добавить линию тренда» и «Тип». В меню представлены четыре типа аппроксимации: логарифмическая, полиномиальная, степенная и экспоненциальная. В данном случае **полином 2 степени** (однако, вид кривой может быть другим, например, логарифмическая зависимость).
7. В опции «Параметры» выберите установки, как показано на рисунке. Щелкните мышкой «ОК».

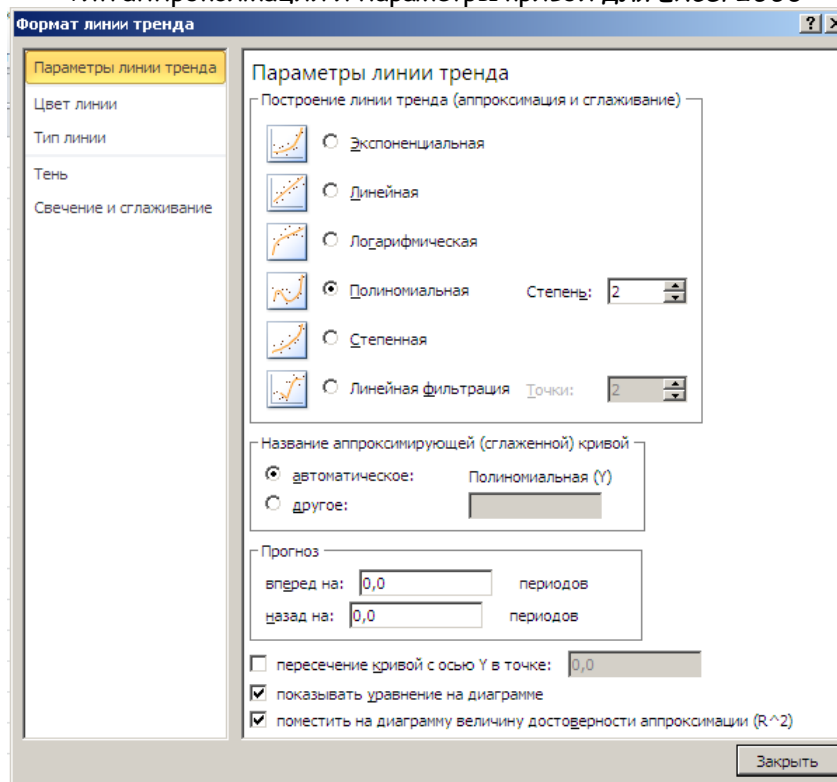


а)



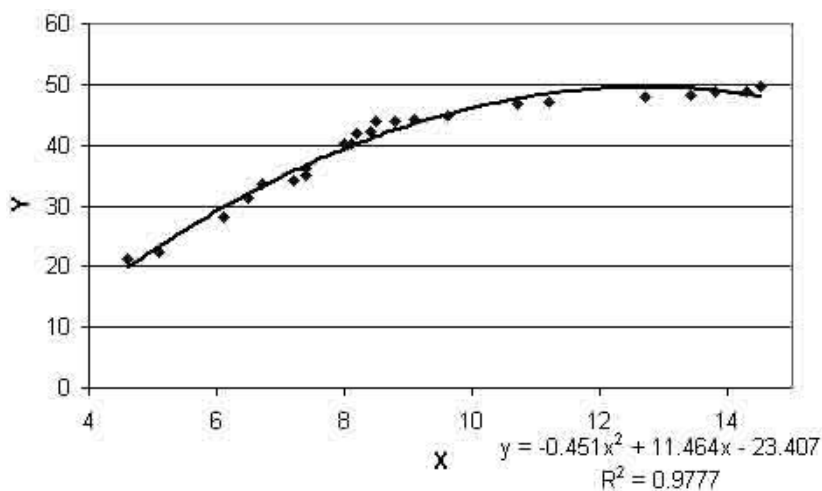
б)

Тип аппроксимации и параметры кривой для Excel 2003



Тип аппроксимации и параметры кривой для Excel 2010

Отредактированная диаграмма представлена на рисунке:



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

Результаты анализа показали, что корреляционная связь между переменными велика $r = 0,97$ и описывается полиномом 2 степени:

$$Y = -0,45 \cdot X^2 + 11,46 \cdot X - 23,41$$

Результат анализа и график скопируйте в документ Word.

Криволинейная корреляция и регрессия в «Statistica» 6

Проведение анализа

Условия задачи такие же, как и в первой части лабораторной работы.

Обязательно создайте новый документ для анализа!

Введите исходные данные, как показано ниже:

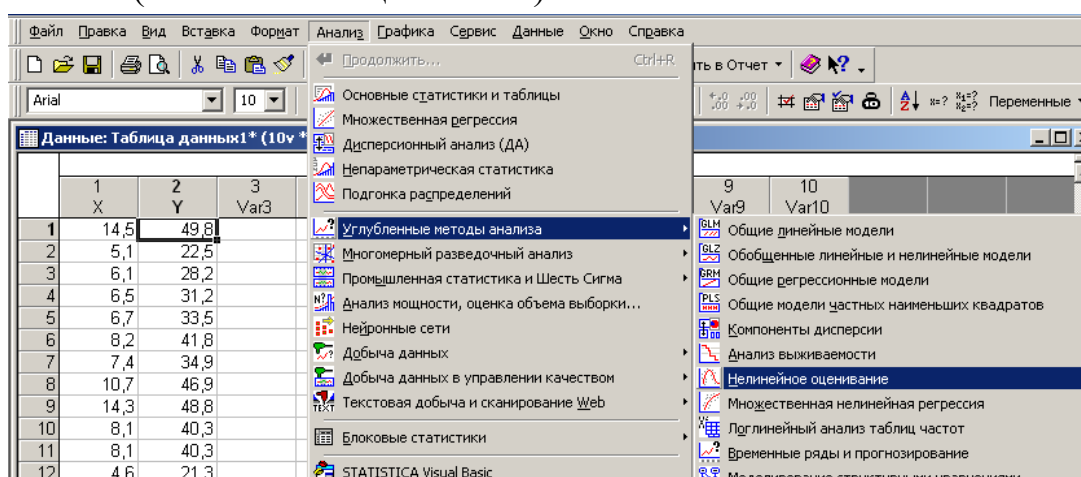
	1 X	2 Y
1	14,5	49,8
2	5,1	22,5
3	6,1	28,2
4	6,5	31,2
5	6,7	33,5
6	8,2	41,8
7	7,4	34,9
8	10,7	46,9
9	14,3	48,8
10	8,1	40,3
11	8,1	40,3
12	4,6	21,3
13	8,4	42,2
14	9,6	44,9
15	8,8	43,9
16	9,1	44,3
17	9,6	44,8
18	7,4	36,2
19	8,5	43,9
20	11,2	47,1
21	12,7	47,9
22	13,4	48,2
23	13,8	48,7
24	7,2	34,1
25	8	40,3

Var1 — независимая переменная — X; **Var2** — зависимая переменная — Y.

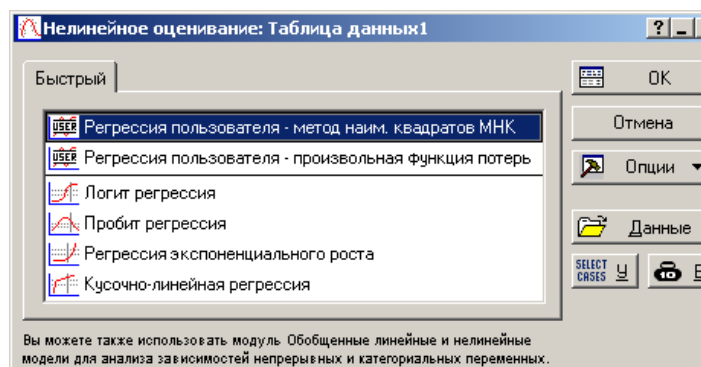
Переименовывать переменные **Var1** и **Var2** не обязательно, иногда это может привести к возникновению ошибки при построении графика криволинейной зависимости во время процедуры нелинейного оценивания. Рекомендуется не менять имя переменной.

Проведем анализ в модуле «**Nonlinear estimation**» (Нелинейная оценка).

1. Из Переключателя модулей «Statistica» откройте модуль «Nonlinear estimation» (Нелинейное оценивание).

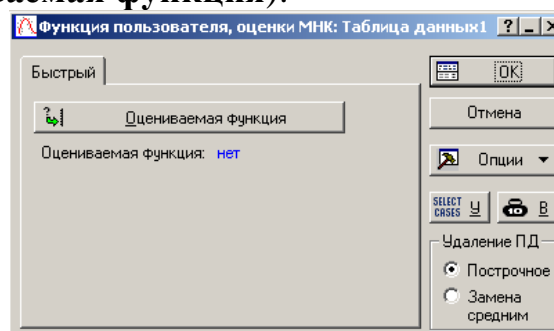


2. На экране появится стартовая панель модуля. Выберите опцию «**User specified regression, least squares**» (Метод наименьших квадратов) и далее щелкните мышью по названию модуля.



Стартовая панель модуля «Nonlinear estimation» (Метод наименьших квадратов)

3. В появившемся окне щелкните мышью по кнопке «**Function of estimated**» (Оцениваемая функция):



Панель ввода функции

4. В окне с помощью клавиатуры введите предполагаемую функцию.

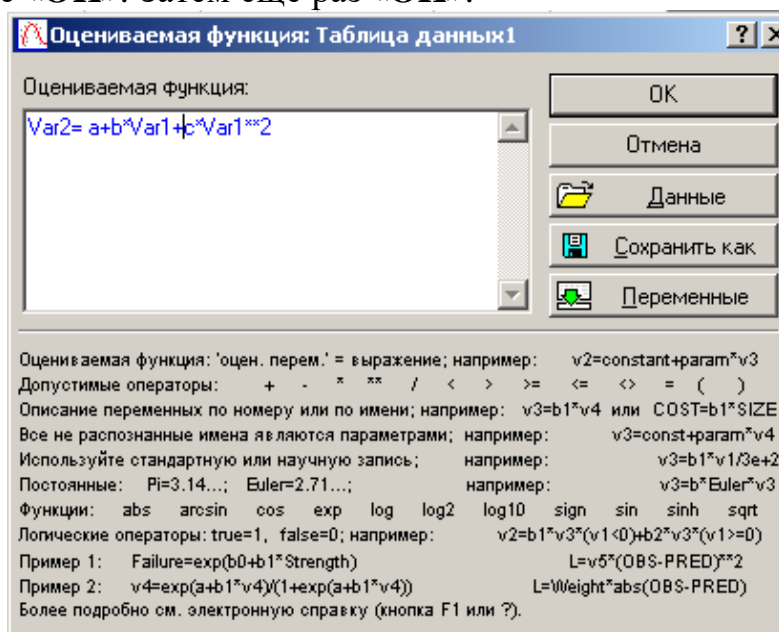
В отличие от подобной операции в табличном редакторе MS Excel в «Statistica» 6 вы можете ввести любую формулу, связывающую зависимую и независимую переменные.

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

$$Y = a + b \cdot X + c \cdot X^2 \text{ или в конкретном случае:}$$
$$Var2 = a + b * Var1 + c * Var1 ** 2$$

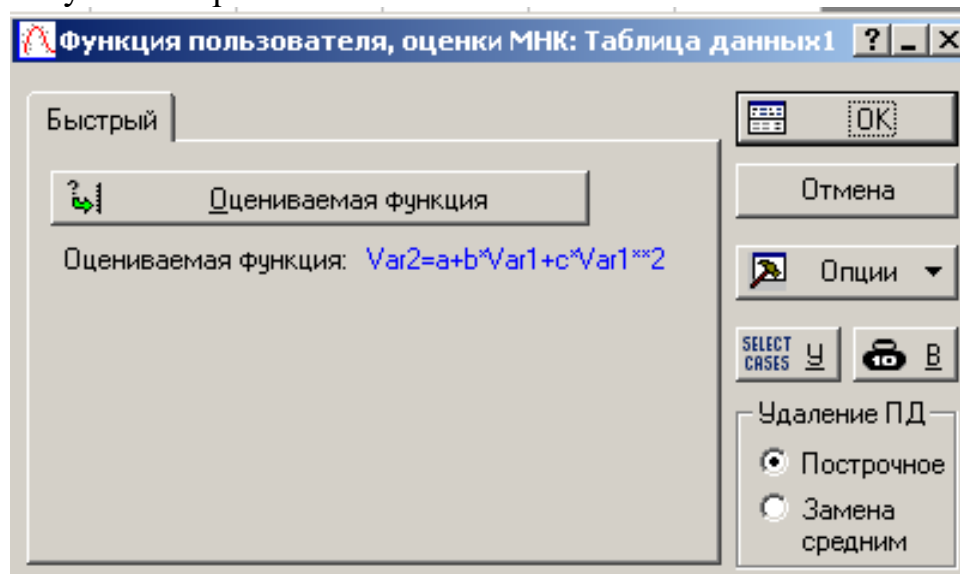
В нижней части рисунка приведен перечень алгебраических и функциональных символов, которые воспринимаются программой.

Нажмите «ОК». Затем еще раз «ОК».

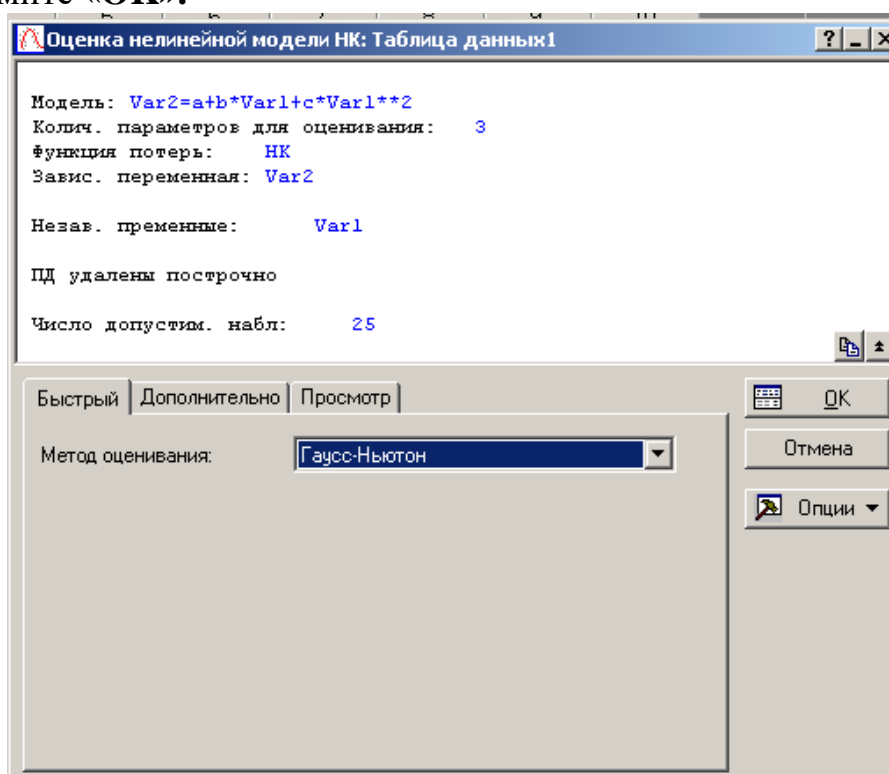


Ввод функции

Результат обработки:

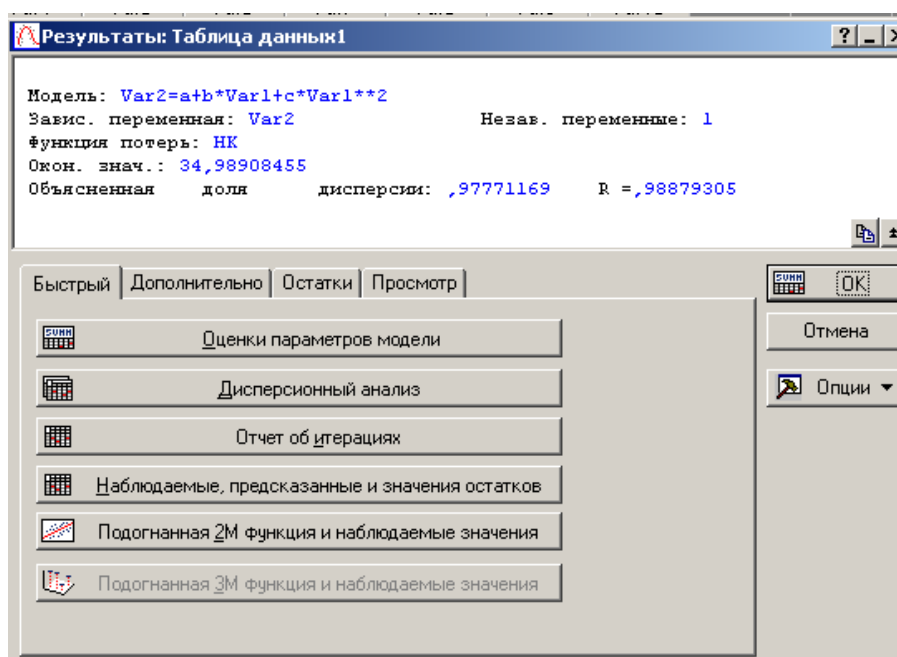


Нажмите «ОК».



Верхняя часть окна информирует о модели, методе, количестве взятых в анализ пар. В середине окна выберите метод аппроксимации. Например: «Gauss–Newton» (Гаусс-Ньютон). Нажмите «ОК».

5. В верхней части появившегося окна результатов показаны значения корреляционного отношения и его квадрата, 0,988 и 0,977. Это указывает на сильную корреляционную связь между переменными.



Окно результатов

6. В окне результатов щелкните мышью по кнопке «**Summary Parameters & standard errors**» (**Оценки параметров модели**). Полученные результаты подкрашены красным цветом, что свидетельствует о достоверности аппроксимации функцией:

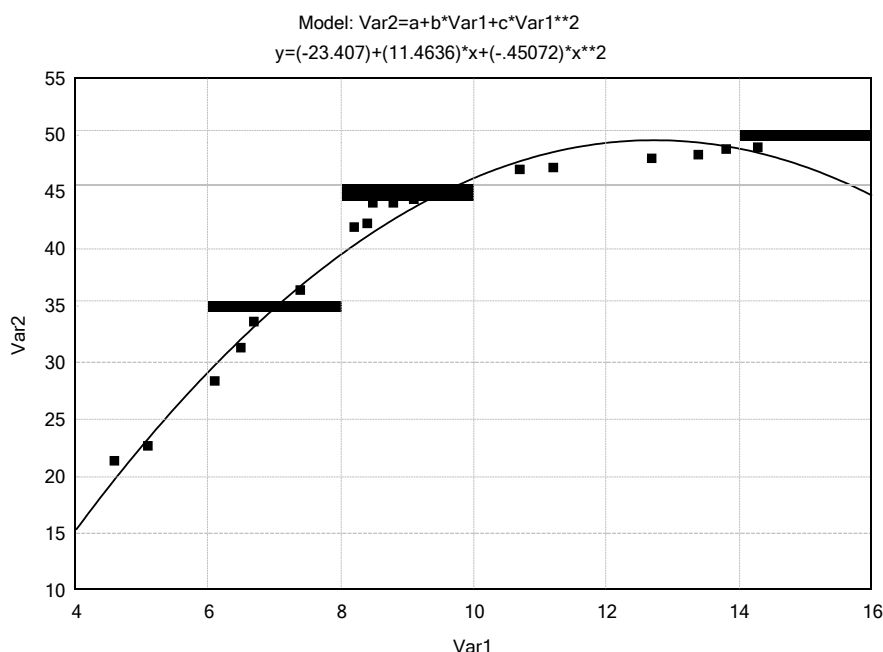
$$Y = -0,45 \cdot X^2 + 11,46 \cdot X - 23,41$$

Модель: Var2 = a+b*Var1+c*Var1**2 (Таблица данных1) Зав. Пер.: Var2 Уров. значимости: 95.0 % (альфа = 0.050)						
	Оценка	Стандарт	t-знач.	p-уров.	Ниж. Дов	Вер. Дов
a	-23,4071	3,010054	-7,7763	0,000000	-29,6496	-17,1646
b	11,4636	0,641511	17,8697	0,000000	10,1332	12,7940
c	-0,4507	0,032066	-14,0558	0,000000	-0,5172	-0,3842

В столбце «**Estimate**» (Оценка) показаны значения коэффициентов: a , b , c . Далее указаны стандартные ошибки, **t-критерий** при 22 степенях свободы, уровень значимости меньше 0,05, верхний и нижний пределы достоверности.

7. Вернитесь в окно «**Результаты**» (кнопка внизу слева). Щелкните мышью по кнопке «**Fitted 2D function & observed vals**» (**Подогнанная 2D функция**). На рисунке вы увидите графическую интерпретацию корреляционной связи исходных массивов в виде заданной функции:

$$Y = -0,45 \cdot X^2 + 11,46 \cdot X - 23,41$$



8. В окне результатов в режиме «**Quick**» (**Быстрый**) нажмите кнопку «**Analysis of Variance**» (**Дисперсионный анализ**). Результат выполненной операции представлен в виде таблицы и свидетельствует о достоверности регрессии ($F = 8806,2$ при $p < 0,00..$).

Модель: Var2=a+b*Var1+c*Var1**2 (Таблица данных1) Зав. Пер. : Var2					
	Сум. квадратов	СС	Сред. квадраты	F-Значение	P-знач.
Регрессия	42016,29	3,00000	14005,43	8806,160	0,00
Остатки	34,99	22,00000	1,59		
Сумма	42051,28	25,00000			
Привед. сумма	1569,84	24,00000			
Регрессия с Приведенной Суммой	42016,29	3,00000	14005,43	214,118	0,00

Модель: Var2=a+b*Var1+c*Var1**2 (Таблица данных1) Зав. Пер. : Var2					
Эффект	1 Сум. квадратов	2 СС	3 Сред. квадраты	4 F-знач.	5 p-знач.
Регрессия	42016,29	3,00000	14005,43	8806,160	0,00
Остатки	34,99	22,00000	1,59		
Сумма	42051,28	25,00000			
Привед. сумма	1569,84	24,00000			
Регрессия с Приведенной Суммой	42016,29	3,00000	14005,43	214,118	0,00

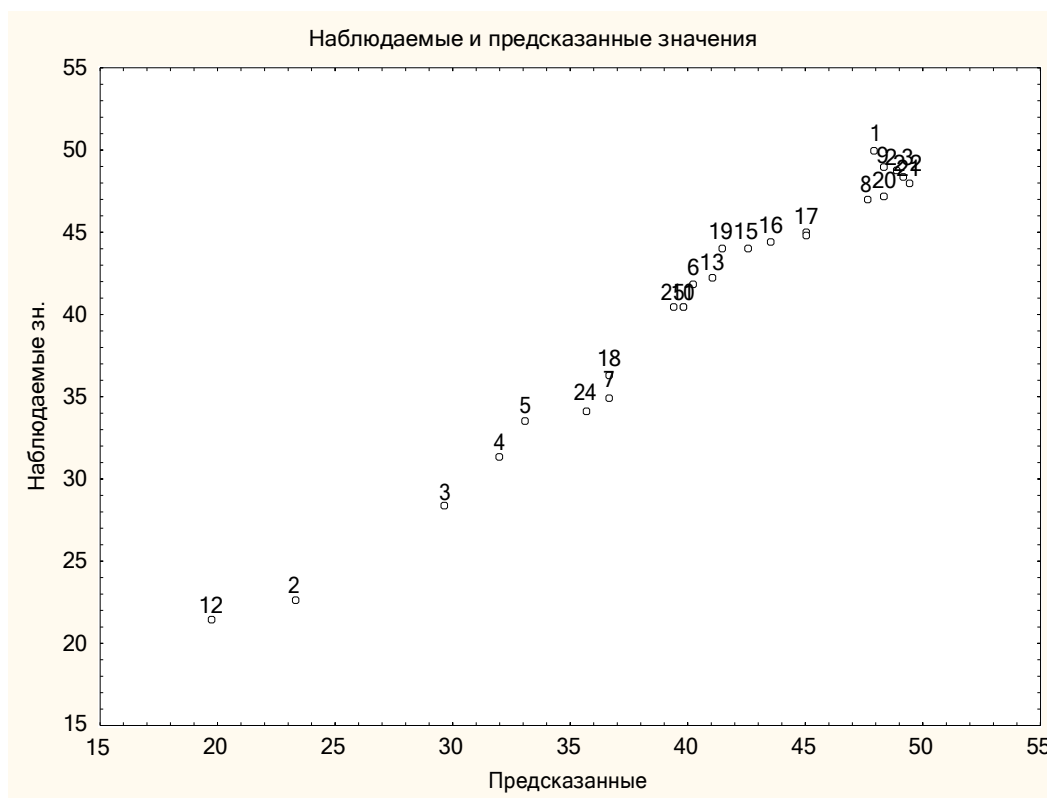
Результат анализа вариантов

9. В окне «Результаты...» перейдите в режим просмотра остатков «Residuals» (Остатки). Щелкните мышью по кнопке «Observed, predicted, residual vals» (Наблюдаемые, предсказанные, остатки). Результаты выполненной операции представлены на рисунке:

Модель: Var2=a+b*Var1+c*Var1**2 (Таблица данных1) Зав. Пер. : Var2					
	Наблюд.	Предсказанные	Остатки		
1	49,80000	48,05252	1,74748		
2	22,50000	23,33429	-0,83429		
3	28,20000	29,74990	-1,54990		
4	31,20000	32,06375	-0,86375		
5	33,50000	33,16658	0,33342		
6	41,80000	40,28853	1,51147		
7	34,90000	36,74256	-1,84256		
8	46,90000	47,65128	-0,75128		
9	48,80000	48,35592	0,44408		
10	40,30000	39,87683	0,42317		
11	40,30000	39,87683	0,42317		
12	21,30000	19,78845	1,51155		
13	42,20000	41,08488	1,11512		
14	44,90000	45,10577	-0,20577		
15	43,90000	42,56941	1,33059		
16	44,30000	43,58815	0,71185		
17	44,80000	45,10577	-0,30577		
18	36,20000	36,74256	-0,54256		
19	43,90000	41,46953	2,43047		
20	47,10000	48,44776	-1,34776		
21	47,90000	49,48504	-1,58504		

10. В окне «Результаты ...» для оценки адекватности модели щелкните мышью по кнопке «Observed vs. Predicted» (Предсказанные и наблюдаемые значения).

Из рисунка видно, массивы наблюдаемых и предсказанных значений описываются линейной функцией $Y = k \cdot X$. При этом $k = 1$ и коэффициент парной корреляции близок к 1.



Визуализация результатов анализа

Вывод по общему анализу предполагаемой зависимости: корреляция и регрессия достоверны, так как $F = 8806,2$ и $t_a = 7,8$, $t_b = 17,9$ и $t_c = 14,1$ (**t-статистика для каждого коэффициента**), что существенно выше критических значений при $p < 0,00..$

В результате эксперимента исследователь формирует **модель** (тип зависимости, приблизительную математическую формулу, степень зависимости), по которой он может с определенной вероятностью предсказывать поведение зависимого объекта, при изменении внешнего воздействия.

Ни в коем случае нельзя строить предположения, используя значения x , которых нет в пределах собранных вами данных.

✓ **Задание для выполнения**

Исходные данные представлены в виде таблицы:

X	Y
33,3	0,03
7,1	0,2
11,2	0,13
14,3	0,08
17,2	0,06
22,2	0,05

34,7	0,03
11,2	0,1
3,7	0,37
33,8	0,03
10,9	0,12
13,1	0,08
18,1	0,07
20,1	0,06
23,4	0,05
30,1	0,04
27	0,04
25,4	0,04
25,6	0,04
30,1	0,04
27,7	0,04
26	0,05
23,2	0,05
19	0,06
32,1	0,03
16,8	0,07
12,1	0,09
9,3	0,17
6,7	0,26
21,5	0,05

1. Необходимо произвести анализ согласно вышеописанному алгоритму и сделать вывод о предполагаемой зависимости между переменными.
2. Результаты скопировать в отдельный документ Word.

Контрольные вопросы

1. Чем отличается анализ криволинейной зависимости от линейной?
2. Что означает криволинейная связь между признаками?
3. Какая величина характеризует нелинейную связь?
4. Можно ли для характеристики нелинейной связи воспользоваться коэффициентом корреляции Пирсона?

Лабораторная работа № 7

Сравнение групп

Непараметрические критерии для анализа количественных признаков

Краткие сведения из теории

Обычно для сравнения групп между собой используют так называемые **параметрические** критерии: критерий Фишера при дисперсионном анализе или критерий Стьюдента при сравнении двух групп. Эти критерии основаны на допущении, что наблюдаемый признак в обеих группах **подчиняется нормальному закону распределения**. Еще одним важным условием для возможности использовать эти методы является **приблизительная равенность дисперсий** в обеих группах. Различаться могут только **средние значения**. По их различию и судят о различии совокупностей. Применяя **параметрические критерии**, необходимо быть уверенным, что условия их применения выполняются хотя бы приблизительно. Иначе велик риск, что, выполнив, казалось бы, правильную последовательность действий, исследователь придет к ошибочным выводам.

✓ Вспомните, какое распределение считается нормальным.

✓ Что такое вариационный ряд?

✓ Что такое среднее значение?

✓ Что характеризует дисперсия?

Необходимые условия применимости критерия Фишера и критерия Стьюдента выполняются далеко не всегда: в одних случаях слишком велика разница дисперсий, в других распределение далеко от нормального, измеряемый признак может оказаться качественным или порядковым.

Природа **порядковых признаков** такова, что о двух значениях можно сказать лишь, какое больше или меньше, но в принципе нельзя — на сколько больше или во сколько раз. (Любой количественный признак можно рассматривать как порядковый, но не наоборот).

Поэтому зачастую (а в медицинских исследованиях даже очень часто) исследователю приходится пользоваться методами, которые не столь требовательны к типу распределения. Такие методы называются **непараметрическими**.

Непараметрические методы заменяют реальные значения признака рангами.

Каждому значению признака в группе присваивается свой ранг в зависимости от величины значения признака.

Вариационный ряд	5	10	17,2	17	18	21	21,8	25	24,6
Ранги	1	2	4	3	5	6	7	9	8

Критерии, основанные на рангах, **не нуждаются** в предположениях о типе распределения. Единственное требование состоит в том, чтобы тип распределения в сравниваемых совокупностях был одинаковым. При этом не нужно знать, что это за распределение и каковы его параметры.

При использовании непараметрических критериев большая часть информации о распределении сохраняется, но нет необходимости знать, что это за распределение. Исследователя не интересуют более параметры распределения, отпадает и необходимость равенства дисперсий.

Если выполняется условие **нормальности** распределения, параметрические критерии обеспечивают наибольшую чувствительность. Если же это условие не выполняется хотя бы приблизительно, их чувствительность существенно снижается и непараметрические критерии дают больше шансов выявить реально существующие различия.

Как выяснить, согласуются ли данные с предположением о нормальности распределения? Простейший способ состоит в том, чтобы нанести их на график. Нарисовав график, посмотрите, похож ли он на нормальное распределение:

- ✓ Похожа ли его форма на график нормального распределения.
- ✓ Важным моментом является достаточная симметричность относительно среднего.
- ✓ Покрывают ли интервал (равный плюс-минус двум стандартным отклонениям от среднего) практически все наблюдения?
- ✓ Сравните графики для разных групп.
- ✓ Близок ли разброс значений? Выясните, насколько близки по величине значения дисперсий в обеих группах.

Если ответы на все вопросы утвердительны (допускаются небольшие отклонения), воспользуйтесь *параметрическим критерием*. В противном случае следует использовать *непараметрический критерий*. Изложенный прием почти наверняка поможет правильно выбрать тип критерия и, как следствие, правильный метод анализа.

Непараметрические методы и области их применения

Сравнение двух выборок:

Критерий Манна—Уитни (независимые (несвязанные) группы)

Критерий Манна-Уитни используется для сравнения двух независимых друг от друга выборок.

➤ Порядок вычисления Т-критерия Манна—Уитни:

1. Данные обеих групп объединяют и упорядочивают по возрастанию. Ранг 1 присваивают наименьшему из всех значений, ранг 2 — следующему и так далее. Наибольший ранг присваивают самому большому среди значений в обеих группах. Если значения совпадают, им присваивают один и тот же средний ранг (например, если два значения поделили 3-е и 4-е места, обоим присваивают ранг 3,5).

2. Для меньшей по количеству членов группы вычисляют T — сумму рангов ее членов. Если численность групп одинакова, T можно вычислить для любой из них.

3. Полученное значение T сравнивают с критическими значениями. Если T меньше или равно первому из них либо больше или равно второму, то нулевая гипотеза отвергается (различия статистически значимы).

Сравнение наблюдений до и после лечения: Критерий Уилкоксона (зависимые (связанные) группы)

Принцип критерия следующий. Для каждого больного вычисляют величину изменения признака. Все изменения упорядочивают по абсолютной величине (без учета знака). Затем рангам приписывают знак изменения и суммируют эти «знаковые ранги» — в результате получается значение **критерия Уилкоксона W** . Используется информация об абсолютной величине изменения и его знаке (т. е. уменьшении или увеличении наблюдаемого признака). Как в случае с критерием Манна — Уитни, здесь также можно перечислить все возможные величины W и найти критическое значение. Исходно ранги присваиваются в соответствии с абсолютной величиной изменения.

➤ Последовательность шагов:

1. Вычислите величины изменений наблюдаемого признака. Отбросьте пары наблюдений, которым соответствует нулевое изменение.

2. Упорядочьте изменения по возрастанию их абсолютной величины и присвойте соответствующие ранги. Рангами одинаковых величин назначьте средние тех мест, которые они делят в упорядоченном ряду.

3. Присвойте каждому рангу знак в соответствии с направлением изменения: если значение увеличилось — «+», если уменьшилось — «-».

4. Вычислите сумму знаковых рангов W (существует вариант критерия Уилкоксона, в котором суммируют только положительные или только отрицательные знаковые ранги. На выводе это никак не сказывается, однако значение W , естественно, получается другим. Поэтому важно знать, на какой вариант критерия рассчитана имеющаяся в вашем распоряжении таблица критических значений.).

5. Сравните полученную величину W с критическим значением. Если она больше критического значения, изменение показателя статистически значимо.

Сравнение нескольких групп:

Критерий Крускала — Уоллиса (независимые (несвязанные) группы)

Эта задача возникает, например, когда нужно определить, одинаково ли эффективны несколько методов лечения, каждый из которых испытывается на отдельной группе.

Критерий Крускала — Уоллиса (аналог дисперсионного анализа) представляет собой обобщение *критерия Манна — Уитни*. Сначала все значения, независимо от того, какой выборке они принадлежат, упорядочивают по возрастанию. Каждому значению присваивается ранг — номер его места в упорядоченном ряду. (Совпадающим значениям присваивают общий ранг, равный среднему тех мест, которые эти величины делят между собой в общем упорядоченном ряду.) Затем вычисляют суммы рангов, относящихся к каждой группе, и для каждой группы определяют средний ранг. При отсутствии межгрупповых различий средние ранги групп должны оказаться близки. Напротив, если существует значительное расхождение средних рангов, то гипотезу об отсутствии межгрупповых различий следует отвергнуть. Значение **критерия Крускала — Уоллиса H** и является мерой такого расхождения средних рангов.

➤ Последовательность действий:

1. Объединив все наблюдения, упорядочить их по возрастанию. Совпадающим значениям ранги присваиваются как среднее тех мест, которые делят между собой эти значения.
2. Вычислить критерий Крускала — Уоллиса H .
3. Сравнить вычисленное значение H с критическим значением χ^2 для числа степеней свободы, на единицу меньшего числа групп. Если вычисленное значение H окажется больше критического, различия групп статистически значимы.

Непараметрическое множественное сравнение

Потребность во множественном сравнении возникает всякий раз, когда с помощью дисперсионного анализа (или его непараметрического аналога — критерия Крускала — Уоллиса) обнаруживается различие нескольких выборок. В этом случае и требуется установить, в чем состоит это различие. Когда объемы выборок равны, для множественного сравнения используют непараметрические варианты критериев Ньюмена—Кейлса и Даннета. Когда же объемы выборок различны, применяется критерий Данна.

Повторные измерения:

КРИТЕРИЙ ФРИДМАНА (зависимые (связанные) группы)

Если одна и та же группа больных последовательно подвергается нескольким методам лечения или просто наблюдается в разные моменты времени, применяют дисперсионный анализ повторных измерений. Но чтобы использование дисперсионного анализа было правомерно, данные должны подчиняться *нормальному распределению*. Если вы в этом не уверены, лучше воспользоваться критерием Фридмана — непараметрическим аналогом дисперсионного анализа повторных измерений.

Логика критерия Фридмана очень проста. Каждый больной ровно один раз подвергается каждому методу лечения (или наблюдается в фиксированные моменты времени). Результаты наблюдений у каждого больного упорядочиваются. Отдельно упорядочиваются значения у каждого больного независимо от всех остальных. Таким образом, получается столько упорядоченных рядов, сколько больных участвует в исследовании. Далее, для каждого метода лечения (или момента наблюдения) вычислим сумму рангов. *Если разброс сумм велик — различия статистически значимы.*

► Порядок расчета критерия Фридмана:

1. Расположите значения признака по возрастанию, каждому значению присвойте ранг.
2. Для каждого из методов лечения (группы) подсчитайте сумму присвоенных ему рангов.
3. Вычислите значение χ^2 .
4. Если число методов лечения и число больных присутствует в таблице критических значений, определите критическое значение χ^2 по этой таблице. Если число методов лечения и число больных достаточно велико (отсутствует в таблице), воспользуйтесь критическим значением χ^2 с числом степеней свободы $v = k - 1$.
5. Если рассчитанное значение χ^2 превышает критическое — различия статистически значимы.

Анализ данных с помощью непараметрических критериев

Сравнение двух выборок:

Критерий Манна — Уитни (независимые (несвязанные) группы)

✓ Задача

Результат решения представить в отдельном файле Word, содержащем таблицы с данными, формулировку нулевой гипотезы, значение уровня значимости, таблицы результатов анализа, графики, выводы по каждой задаче.

Роды по Лебуайе

В последние десятилетия произошел коренной пересмотр взглядов на родовспоможение. Акушерская революция совершалась под лозунгом «Отец вместо седативных средств». Восторжествовала точка зрения, согласно которой при нормальных родах следует прибегать к помощи психологических, а не лекарственных средств. Что делать конкретно, мнения расходились. Масла в огонь подлила книга Лебуайе «Рождение без насилия». Французский врач предлагал комплекс мер, призванных свести к минимуму потрясение, которое испытывает новорожденный при появлении на свет. *Роды надлежит принимать в тихом затемненном помещении. Сразу после родов ребенка следует уложить на живот матери и не перерезать пуповину, пока та не перестанет пульсировать. Затем, успокаивая младенца легким поглаживанием, нужно поместить его в теплую ванну, чтобы «внушить, что разрыв с организмом матери — не шок, но удовольствие».* Лебуайе указывал, что дети, рожденные по его методике, здоровее и радостнее других. Многие врачи считали, что предложенная методика не только противоречит общепринятой практике, но и создает дополнительную опасность для матери и ребенка. Тем не менее у Лебуайе нашлись и сторонники.

Как часто бывает в медицине, отсутствие достоверных данных могло затянуть спор на многие годы. Пока Н. Нелсон и соавт. не провели клиническое испытание, материалы ограничивались «клиническим опытом» автора методики. В эксперименте Нелсон, проведенном в клинике канадского университета Макмастер, участвовали роженицы без показаний к искусственному родоразрешению, срок беременности которых составлял не менее 36 недель и которые были согласны рожать как по обычной методике, так и по Лебуайе. *Роженицы были случайным образом разделены на две группы. В контрольной роды проводились по общепринятой методике в нормально освещенном помещении с обычным уровнем шума; после рождения пуповина немедленно перерезалась, ребенка пеленали и отдавали матери. В экспериментальной группе роды принимались по методике Лебуайе.* В обеих группах при родах присутствовали мужья, применение обезболивающих средств было минимальным. Тем самым, группы различались только в том, в чем методика Лебуайе не совпадает с общепринятой. То, в какую группу попала роженица, было известно самой роженице и всем, кто присутствовал при родах. На этом этапе эффект плацебо исключить было невозможно. Однако уже на этапе послеродового наблюдения одна из сторон, а именно врачи, которые оценивали состояние ребенка, не знали, по какой методике происходили роды. Таким образом исследование Нелсон было **простым слепым**: условия знала только одна из сторон, наблюдателю же они были неизвестны.

Для оценки развития детей была разработана специальная шкала. Из числа детей, рожденных по обычной методике, оценку «отлично» по этой шкале получали примерно 30 %. Изучив труды Лебуайе, Нелсон и соавт.

пришли к выводу, что предлагаемый метод, судя по заявлениям автора, гарантирует оценку «отлично» у 90 % детей. Приняв уровень значимости $\alpha = 0,05$, исследователи рассчитали, что для обеспечения 90 % вероятность выявить такие различия в каждой из групп должно быть по 20 детей. Работа продолжалась целый год. За это время исследователи провели беседы с 187 потенциальными участницами, разъясняя им смысл предстоящего эксперимента. 34 женщины не подошли по состоянию здоровья, 97 отказались участвовать в эксперименте (из них 70 собирались рожать только по методике Лебуайе). Из оставшихся 56 женщин одна успела родить до рандомизации. В результате число участниц сократилось до 55. Их и разделили случайным образом на две группы. После того, как из исследования выбыла одна из попавших в контрольную группу, в этой группе оказалось 26, а в экспериментальной 28 рожениц. Однако у 6 женщин в контрольной группе и у 8 в экспериментальной возникли осложнения, и их пришлось исключить из участия в эксперименте. В итоге в каждой из групп оказалось по 20 женщин.

Обратите внимание, насколько трудно обеспечить достаточную численность групп даже в простом исследовании.

Оценка по шкале развития производилось сразу после родов, а также спустя несколько месяцев.

Остановимся на одном из показателей — *времени бодрствования в первый час жизни*. Предполагалось, что чем лучше состояние новорожденного, тем более он активен. Значит, у младенцев, рожденных по Лебуайе, время бодрствования должно быть продолжительнее, чем у рожденных по обычной методике.

Роды по обычной методике	Роды по Лебуайе
5	2
10,1	19
17,7	29,7
20,3	32,1
22	35,4
24,9	36,7
26,5	38,5
30,8	40,2
34,2	42,1
35	43
36,6	44,4
37,9	45,6
40,4	46,7
45,5	47,1
49,3	48
51,1	49
53,1	50,9
55	51,2
56,7	52,5
58	53,3

Полученные данные не подчиняются нормальному распределению. Особенно это заметно в экспериментальной группе. Поэтому параметрические методы, например *критерий Стьюдента*, к этим данным неприменимы. Воспользуемся непараметрическим **критерием Манна — Уитни**.

Порядок анализа в «Satstistica» 6

Нулевая гипотеза: различия между группами статистически незначимо.
Критерий значимости примем равным 0,05.

1. Скопируйте данные (числовые) на «Лист 1» табличного редактора MS Excel.

	А	В
1	Роды по обычной методике	Роды по Лебуайе
2	5	2
3	10,1	19
4	17,7	29,7
5	20,3	32,1
6	22	35,4
7	24,9	36,7
8	26,5	38,5
9	30,8	40,2
10	34,2	42,1
11	35	43
12	36,6	44,4
13	37,9	45,6
14	40,4	46,7
15	45,5	47,1
16	49,3	48
17	51,1	49
18	53,1	50,9
19	55	51,2
20	56,7	52,5
21	58	53,3

2. Представьте данные в виде:

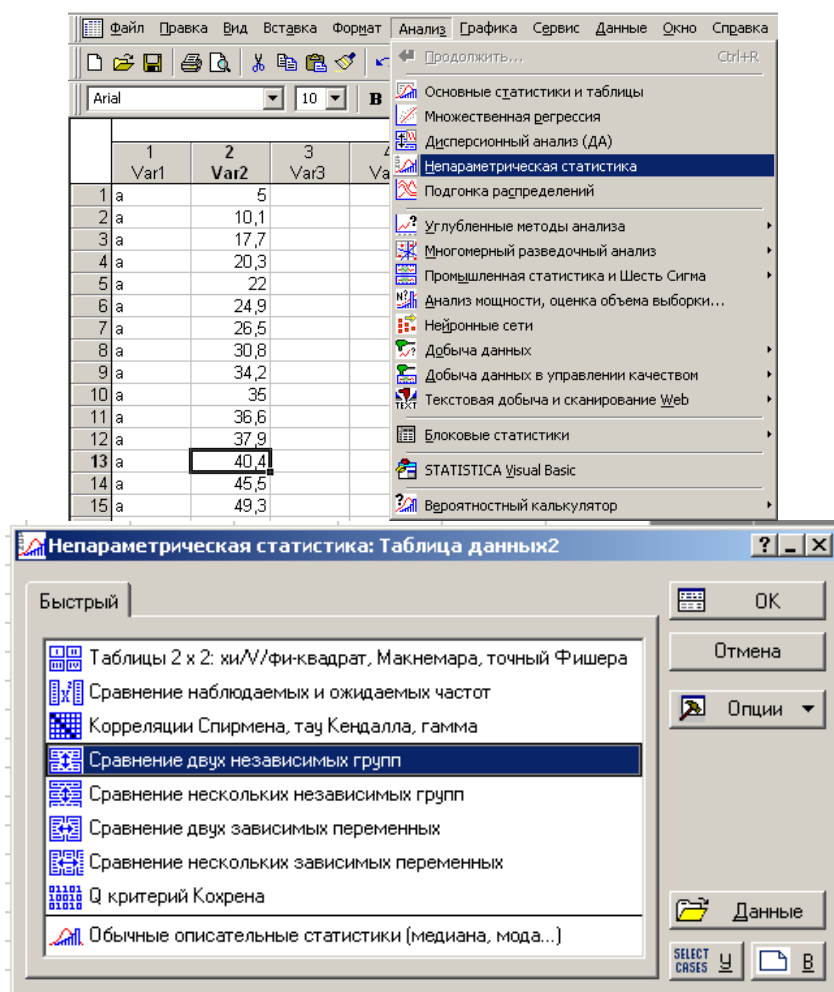
	А	В	С
10	a	35	
11	a	36,6	
12	a	37,9	
13	a	40,4	
14	a	45,5	
15	a	49,3	
16	a	51,1	
17	a	53,1	
18	a	55	
19	a	56,7	
20	a	58	
21	b	2	
22	b	19	
23	b	29,7	
24	b	32,1	
25	b	35,4	
26	b	36,7	
27	b	38,5	
28	b	40,2	

3. Скопируйте полученную таблицу в программу «Statistica» 6.

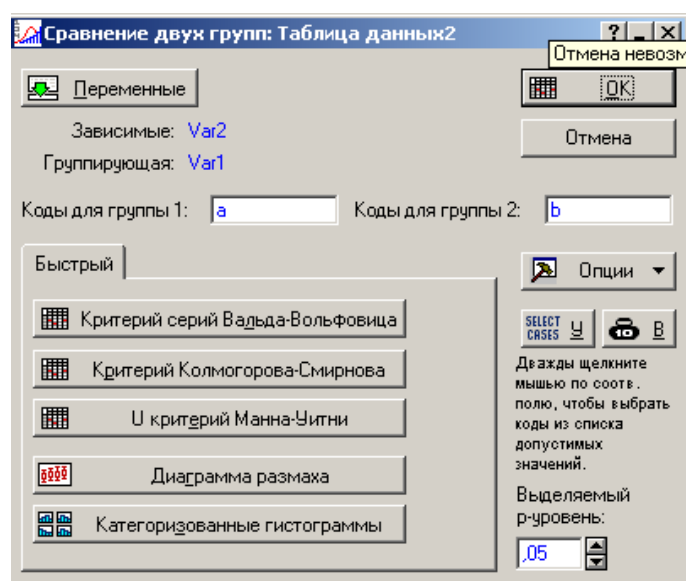
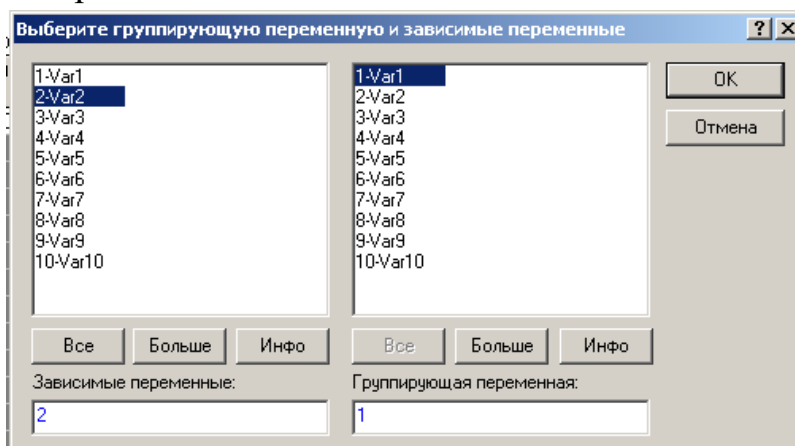
	1 Var1	2 Var2	3 Var3
4 a		20,3	
5 a		22	
6 a		24,9	
7 a		26,5	
8 a		30,8	
9 a		34,2	
10 a		35	
11 a		36,6	
12 a		37,9	
13 a		40,4	
14 a		45,5	
15 a		49,3	
16 a		51,1	
17 a		53,1	
18 a		55	
19 a		56,7	
20 a		58	
21 b		2	
22 b		19	
23 b		29,7	
24 b		32,1	
25 b		35,4	
26 b		36,7	

4. Далее: Анализ — Непараметрическая статистика — Сравнение двух независимых групп.

✓ Почему группы считаются независимыми?



5. Указать переменные:



6. Результат обработки:

Манна-Уитни U критерий (Таблица данных2)										
По перем. Var1										
Отмеченные критерии значимы на уровне $p < ,05000$										
Перем.	Сум. ранг a	Сум. ранг b	U	Z	p-уров.	Z скорр.	p-уров.	N набл. a	N набл. b	2-х стор точный p
Var2	374,0000	446,0000	164,0000	-0,973803	0,330155	-0,973803	0,330155	20	20	0,340785

В полученной таблице нас интересуют значения **критерия U** и величина **p-уровня**.

Значение **p-уровня** больше заданного значения **уровня значимости (0,05)**, следовательно мы остаемся в рамках нулевой гипотезы. Существенного преимущества в приеме родов по методике Лебуайе в сравнении с приемом родов по обычной методике нет.

Скопируйте результат в файл отчета Word. Выделить таблицу результатов, далее Правая кнопка мыши — Копировать с заголовками.

Workbook1* - Манна-Уитни U критерий (Таблица данных1)

По перем. Var1
Отмеченные критерии значимы на уровне $p < 0,05000$

Перем.	Сум. ранг a	Сум. ранг b	U	Z	p-уров.	Z скорр.	p-уров.	N набл. a	N набл. b	2-х стор точный p
Var2	374,0000	446,0000	164,0000	-0,973803	0,330155	-0,973803	0,330155	20	20	0,340785

Контекстное меню:

- Блочные статистики
- Графики блочных данных
- Графики исходных данных
- Вырезать Ctrl+X
- Копировать Ctrl+C
- Копировать с заголовками**
- Вставить Ctrl+V
- Специальная вставка...
- Заполнить/Стандартизовать блок
- Очистить
- Формат
- Выделение ячеек

7. Возвращаемся в окно «Сравнение двух групп».

Workbook3*

Манна-Уитни U критерий (Таблица данных2)

По перем. Var1
Отмеченные критерии значимы на уровне $p < 0,05000$

Перем.	Сум. ранг a	Сум. ранг b	U	Z	p-уров.	Z скорр.	p-уров.	N набл. a	N набл. b	2-х стор точный p
Var2	374,0000	446,0000	164,0000	-0,973803	0,330155	-0,973803	0,330155	20	20	0,340785

Окно «Сравнение двух групп»

8. Построение *диаграммы размаха*:

Сравнение двух групп: Таблица данных2

Переменные

Зависимые: Var2

Группирующая: Var1

Коды для группы 1: a Коды для группы 2: b

Быстрый

- Критерий серий Вальда-Вольфовица
- Критерий Колмогорова-Смирнова
- U критерий Манна-Уитни
- Диаграмма размаха**
- Категоризованные гистограммы

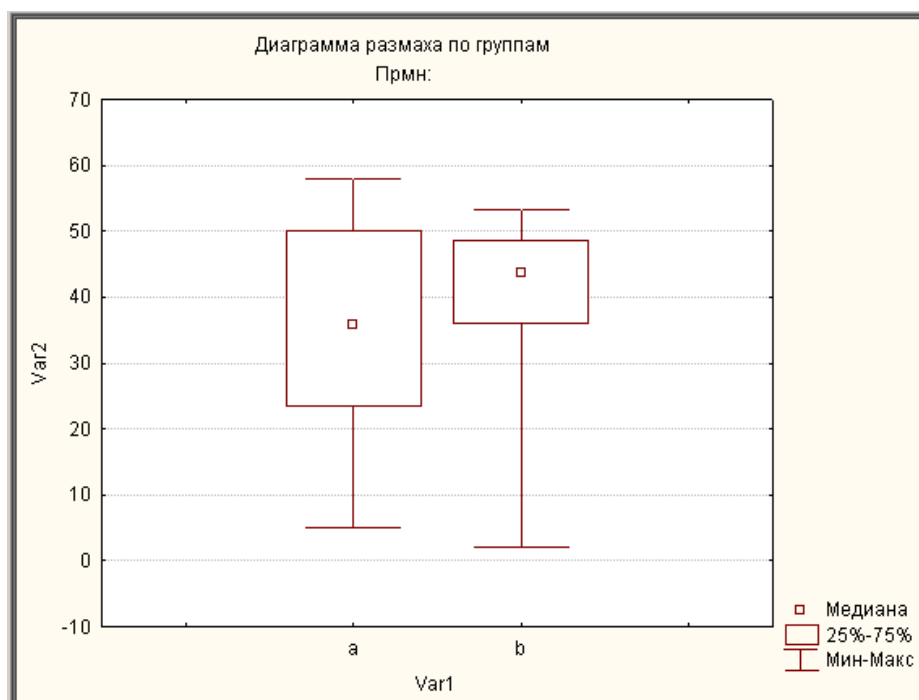
Выберите переменную для диаграммы размаха

- 1-Var1
- 2-Var2**
- 3-Var3
- 4-Var4
- 5-Var5
- 6-Var6
- 7-Var7
- 8-Var8
- 9-Var9
- 10-Var10

Выберите переменную: 2

ОК Отмена

Диаграмма размаха, с указанием медианы, процентилей, экстремумов.



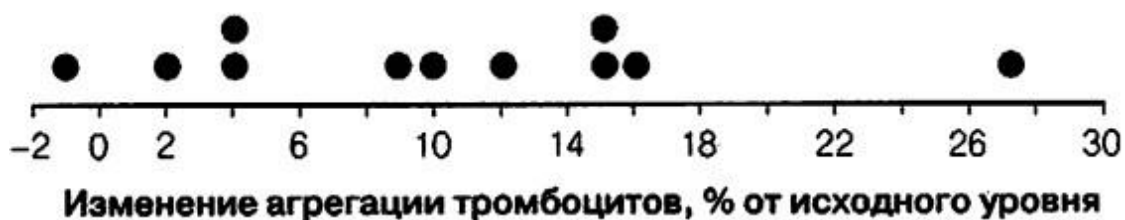
**Сравнение наблюдений до и после лечения:
Критерий Уилкоксона (зависимые (связанные) группы)**

✓ Задача

Курение и функция тромбоцитов

Агрегация тромбоцитов до и после выкуривания сигареты:

Участник	До курения	После курения
1	25	27
2	25	29
3	27	37
4	44	56
5	30	46
6	67	82
7	53	57
8	53	80
9	52	61
10	60	59
11	28	43



✓ Можно ли считать распределение изменения *нормальным*?

В данном случае для суждения о типе распределения данных слишком мало. Смущает и «выскакивающее» значение 27 % — оно говорит о возможной асимметрии распределения. В подобных случаях лучше не рисковать и воспользоваться непараметрическим критерием. Применим **критерий Уилкоксона**.

Нулевая гипотеза: различия между группами случайны и незначительны.

Уровень значимости: 0,05.

Порядок анализа в «Statistica» 6

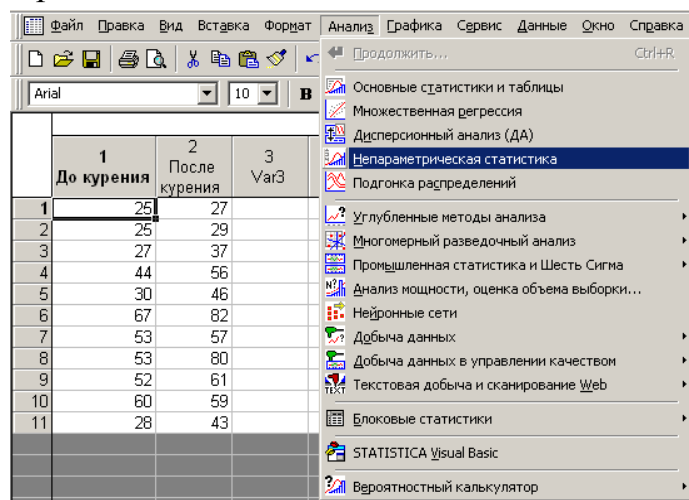
1. Скопировать данные на «Лист 1» табличного редактора Excel.

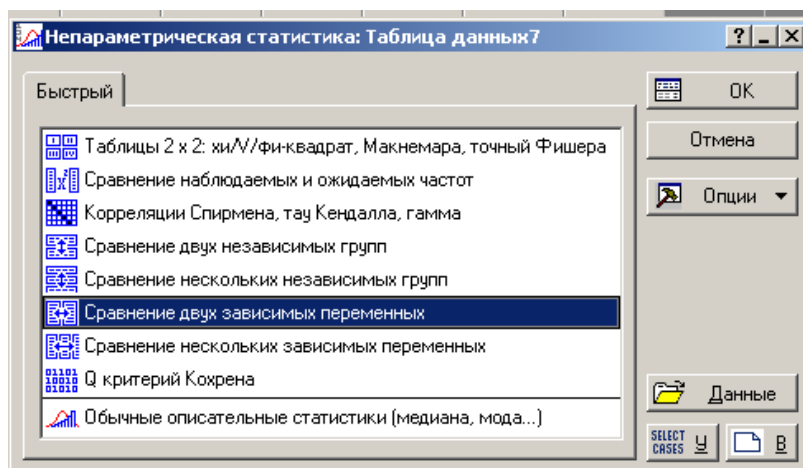
	А	В	С
	Участник	До курения	После курения
1			
2	1	25	27
3	2	25	29
4	3	27	37
5	4	44	56
6	5	30	46
7	6	67	82
8	7	53	57
9	8	53	80
10	9	52	61
11	10	60	59
12	11	28	43

2. Из табличного редактора данные по каждому этапу наблюдения скопировать в «Statistica», без номеров участников и заголовков таблицы. Переменные подписать как указано на иллюстрации.

	1 До курения	2 После курения
1	25	27
2	25	29
3	27	37
4	44	56
5	30	46
6	67	82
7	53	57
8	53	80
9	52	61
10	60	59
11	28	43

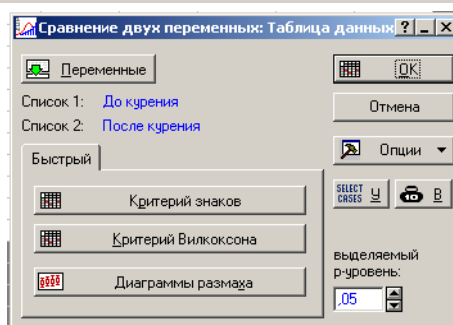
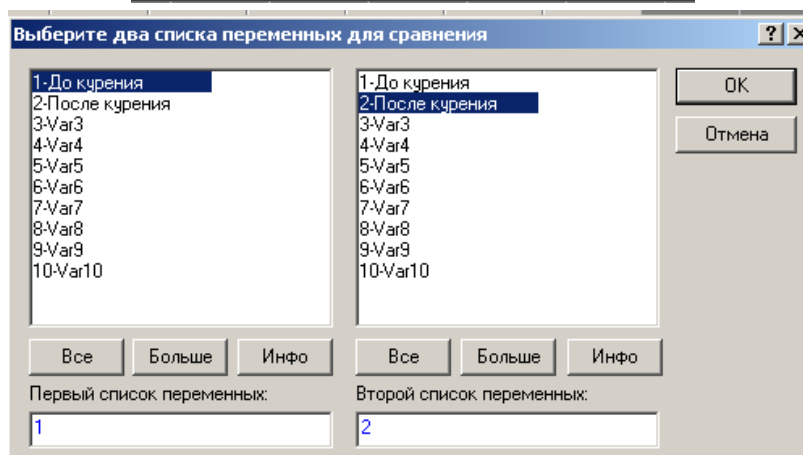
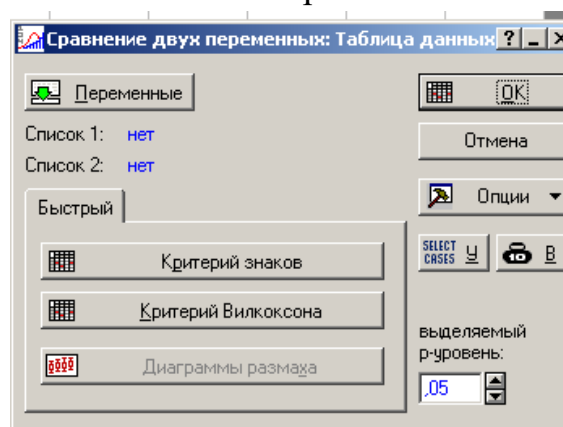
3. Далее: *Анализ — Непараметрическая статистика — Сравнение двух зависимых переменных.*





✓ Почему переменные считаются зависимыми?

4. Указать переменные: список переменных 1 и 2.

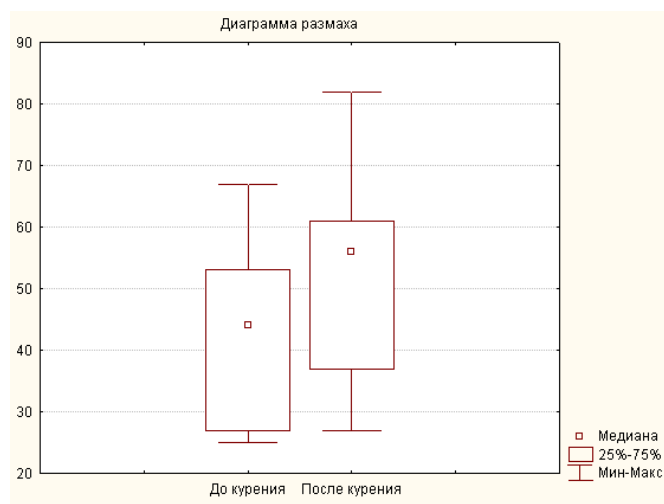
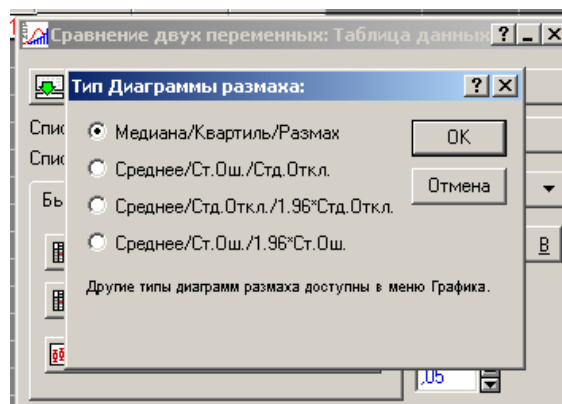
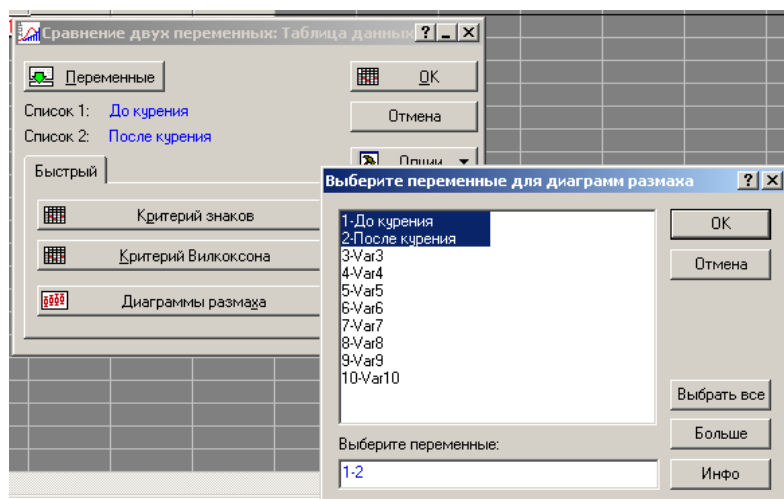


5. Таблица результатов:

		Критерий Вилкоксона (Таблица данных7)			
		Отмеченные критерии значимы на уровне $p < ,05000$			
Пара перем.	Число набл.	T	Z	p-уров.	
До курения & После курения	11	1,000000	2,845147	0,004439	

Значение **p-уровня** меньше значение выбранного **уровня значимости** (0,05). Следовательно нулевая гипотеза отклоняется.

6. Графическое представление. Диаграммы размаха.



Сравнение нескольких групп:

Критерий Крускала — Уоллиса (независимые (несвязанные) группы)

✓ Задача

Вводя изотоп внутривенно и наблюдая за его распространением с помощью гамма-камеры, можно определить кровенаполнение различных органов, в том числе легких. Р. Окада и соавт. решили использовать этот метод для локализации поражения коронарных артерий при ишемической болезни сердца. Правая коронарная артерия снабжает кровью главным образом правый желудочек, левая — главным образом левый. Левый желудочек перекачивает кровь, которая поступает в него из легких, по всему телу. При поражении левой коронарной артерии кровоснабжение левого желудочка ухудшается. В покое, когда объем перекачиваемой крови невелик, это никак не проявляется, однако при физической нагрузке это приводит к накоплению крови в легких. При поражении правой коронарной артерии этого не происходит.

Примерно так рассуждали авторы, приступая к работе. Было обследовано 33 человека: 9 здоровых (1-я группа) и 24 больных ишемической болезнью сердца, из них 5 с поражением только правой коронарной артерии (2-я группа) и 19 с поражением обеих коронарных артерий или только левой (3-я группа). Рассчитывали отношение кровенаполнения легких при физической нагрузке к кровенаполнению в покое: по мысли авторов, в 3-й группе этот показатель должен быть выше, чем в первых двух.

Результаты представлены в таблице:

1 группа	2 группа	3 группа
0,83	0,86	0,98
0,89	0,92	1,02
0,91	1	1,03
0,93	1,02	1,04
0,94	1,2	1,05
0,97		1,06
0,97		1,07
0,98		1,22
1,02		1,07
		1,23
		1,13
		1,08
		1,32
		1,1
		1,15
		1,37
		1,18
		1,12
		1,58

- ✓ Различаются ли группы между собой?
- ✓ Если да, то как именно и достаточно ли велико различие, чтобы исследуемый показатель можно было использовать для определения пораженной коронарной артерии?

Нулевая гипотеза: различие между группами статистически не значимо.

Порядок анализа в «Statistica» 6

1. Скопировать таблицу на «Лист 1» табличного редактора MS Excel.

	А	В	С
1	1 группа	2 группа	3 группа
2	0,83	0,86	0,98
3	0,89	0,92	1,02
4	0,91	1	1,03
5	0,93	1,02	1,04
6	0,94	1,2	1,05
7	0,97		1,06
8	0,97		1,07
9	0,98		1,22
10	1,02		1,07
11			1,23
12			1,13
13			1,08
14			1,32
15			1,1
16			1,15
17			1,37
18			1,18
19			1,12
20			1,58

2. Представить ее в виде как на иллюстрации. Группа 1 — а, группа 2 — b, группа 3 — с.

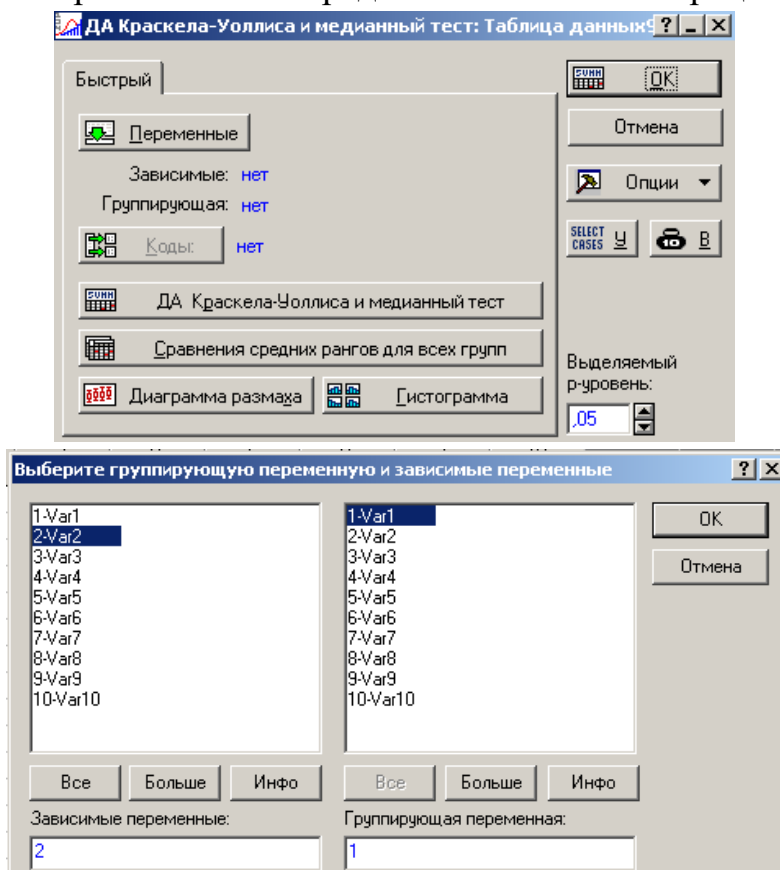
	А	В
4	а	0,93
5	а	0,94
6	а	0,97
7	а	0,97
8	а	0,98
9	а	1,02
10	b	0,86
11	b	0,92
12	b	1
13	b	1,02
14	b	1,2
15	с	0,98
16	с	1,02
17	с	1,03
18	с	1,04
19	с	1,05
20	с	1,06
21	с	1,07

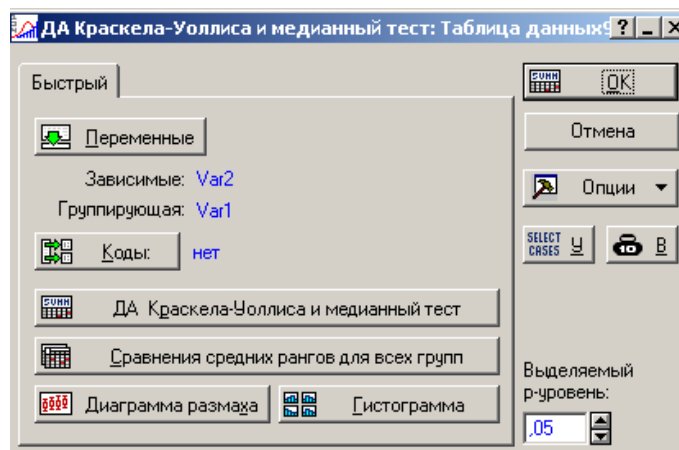
3. Скопировать полученную таблицу в программу «Statistica».

	1 Var1	2 Var2
1	a	0,83
2	a	0,89
3	a	0,91
4	a	0,93
5	a	0,94
6	a	0,97
7	a	0,97
8	a	0,98
9	a	1,02
10	b	0,86
11	b	0,92
12	b	1
13	b	1,02
14	b	1,2
15	c	0,98
16	c	1,02
17	c	1,03
18	c	1,04
19	c	1,05
20	c	1,06
21	c	1,07
22	c	1,22
23	c	1,07

4. Далее: *Анализ — Непараметрическая статистика — Сравнение нескольких независимых групп.*

5. Указать переменные как представлено на иллюстрациях.





6. Таблица результатов:

		Ранговый ДА Крассела-Уоллиса; Var2 (Таблица данных9) Группирующая переменная: Var1 Кр.Крассела-Уоллиса: $H(2, N = 33) = 18,36631$ $p = ,0001$					
Зависим.: Var2	Код	Допуст N	Сумма Ряды				
a	101	9	62,5000				
b	102	5	61,0000				
c	103	19	437,5000				

		Медианный тест, общ. медиана = 1,04000; Var2 (Таблица данных9) Группирующая переменная: Var1 Хи-квадрат = 17,15387 $ss = 2$ $p = ,0002$					
Зависимые: Var2		a	b	c	Всего		
<= Медиана: наблюд.		9,00000	4,00000	4,00000	17,00000		
ожидаемые		4,63636	2,57576	9,78788			
набл.-ожд.		4,36364	1,42424	-5,78788			
> Медиана: наблюд.		0,00000	1,00000	15,00000	16,00000		
ожидаемые		4,36364	2,42424	9,21212			
набл.-ожд.		-4,36364	-1,42424	5,78788			
Сумма: наблюд.		9,00000	5,00000	19,00000	33,00000		

Различия между группами статистически значимы ($p < 0,05$). Нулевая гипотеза отклоняется.

7. Сравнение каждой группы попарно: *Сравнение средних рангов для всех групп.*

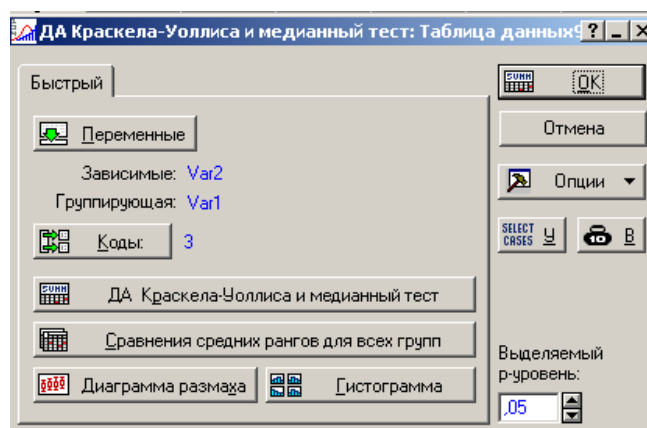


Таблица результатов сравнения групп:

Сравнения р значений (2-стороннее); Var2 (Таблица данных9) Группирующая переменная: Var1 Кр.Краскела-Уоллиса: $H(2, N=33) = 18,36631$ $p = ,0001$							
Зависим.: Var2	a	b	c				
	R:6,9444	R:12,200	R:23,026				
a		0,989514	0,000119				
b	0,989514		0,077728				
c	0,000119	0,077728					

В таблице приведены значения **p** (вероятность ошибочного результата). Исходя из полученных результатов можно заключить, что различия значимы между первой и третьей группами.

Повторные измерения:

Критерий Фридмана (зависимые (связанные) группы)

✓ Задача

Результаты измерения качества зрения у группы учащихся в разный период времени (1,5,11 классы):

	1		5		11	
ФИО	Левый	Правый	Левый	Правый	Левый	Правый
1	0,6	0,5	1	0,5	1	0,4
2	0,3	0,5	0,2	0,6	0,8	0,7
3	1	1	1	1	0,6	0,6
4	1	1	0,2	0,1	0,5	0,5
5	0,9	1	0,8	1	0,5	0,2
6	1	1	0,9	0,9	0,9	0,5
7	1	1	1	1	0,8	1
8	1	1	1	1	0,3	0,3
9	1	1	1	1	0,9	0,9
10	0,1	0,5	0,3	0,2	0,3	0,5
11	0,9	0,9	1	1	1	0,7
12	0,6	0,7	0,8	0,6	0,4	0,3
13	0,8	0,8	0,5	0,1	0,5	0,6
14	1	1	1	1	0,2	0,3

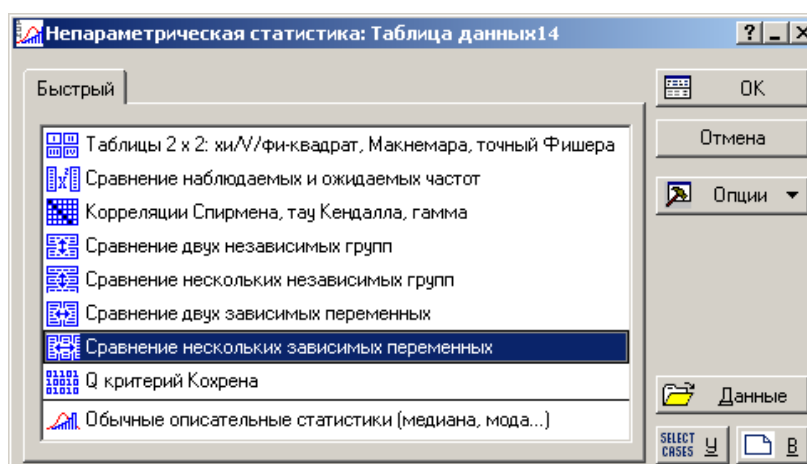
Определите есть ли различия между группами.

Порядок анализа в «Statistica» 6

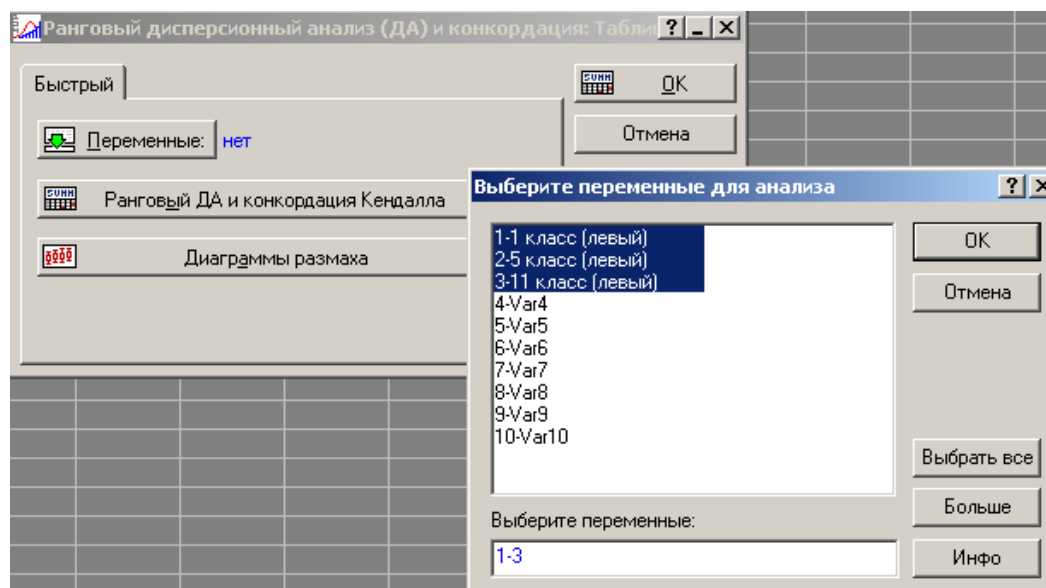
1. Скопировать таблицу в Excel.
2. Скопировать данные по левому глазу (для всех классов) в программу «Statistica», переменные обозначить.

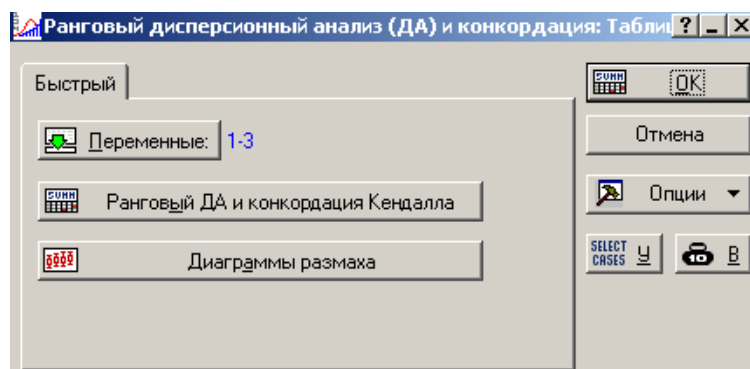
	1 1 класс (левый)	2 5 класс (левый)	3 11 класс (левый)
1	0,6	1	1
2	0,3	0,2	0,8
3	1	1	0,6
4	1	0,2	0,5
5	0,9	0,8	0,5
6	1	0,9	0,9
7	1	1	0,8
8	1	1	0,3
9	1	1	0,9
10	0,1	0,3	0,3
11	0,9	1	1
12	0,6	0,8	0,4
13	0,8	0,5	0,5
14	1	1	0,2

3. Анализ — Непараметрическая статистика — Сравнение нескольких зависимых переменных.



4. Указать переменные (группы).



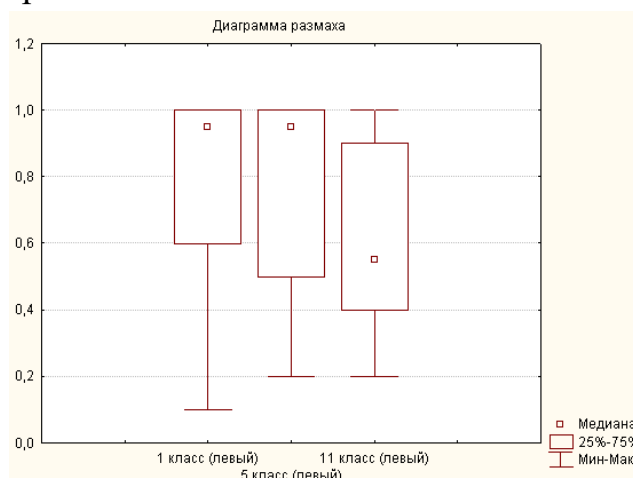


5. Таблица результатов:

Ранговый ДА и конкордация Кендалла (Таблица данных14)						
ДА хи-кв. (N = 14, ss = 2) = 4,043478 p < ,13243						
Козфф. конкордации = ,14441 Средн. ранг r = ,07860						
Перем.	Средний ранг	Сумма рангов	Среднее	Ст. откл		
1 класс (левый)	2,250000	31,50000	0,800000	0,293520		
5 класс (левый)	2,142857	30,00000	0,764286	0,320113		
11 класс (левый)	1,607143	22,50000	0,621429	0,275062		

Различие между группами статистически не значимо (величина p больше чем уровень значимости).

6. Диаграмма размаха.



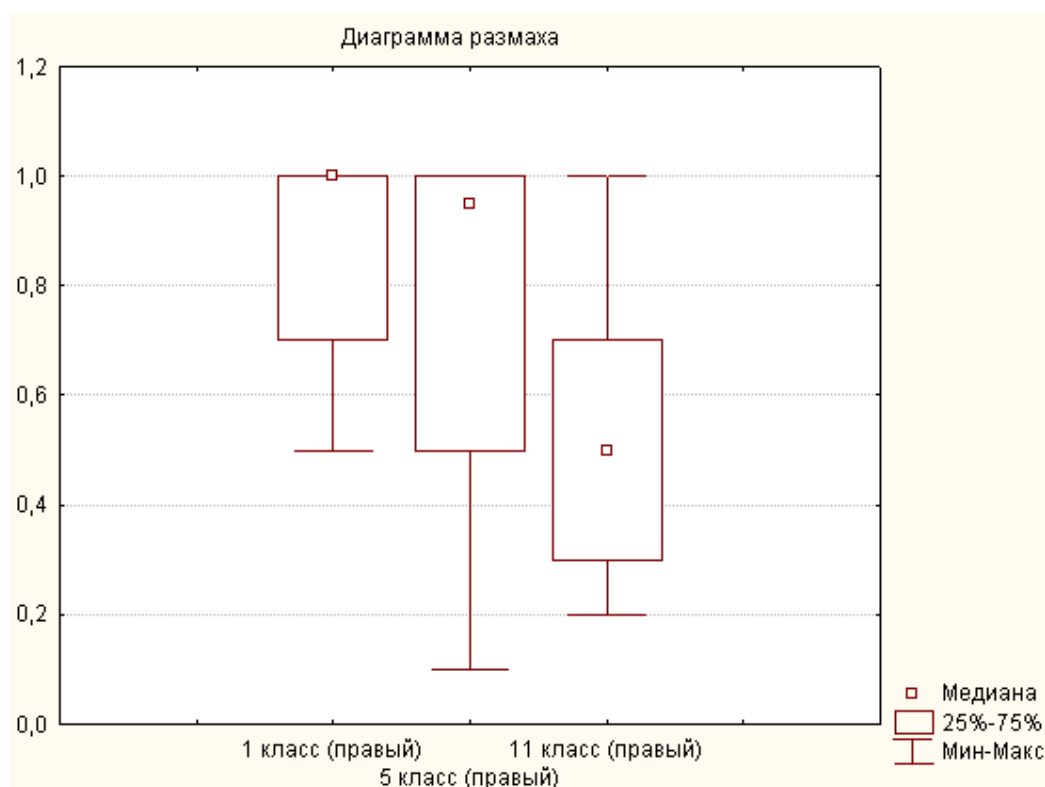
7. Тот же анализ для правого глаза.

	1	2	3
	1 класс (левый)	5 класс (левый)	11 класс (левый)
1	0,5	0,5	0,4
2	0,5	0,6	0,7
3	1	1	0,6
4	1	0,1	0,5
5	1	1	0,2
6	1	0,9	0,5
7	1	1	1
8	1	1	0,3
9	1	1	0,9
10	0,5	0,2	0,5
11	0,9	1	0,7
12	0,7	0,6	0,3
13	0,8	0,1	0,6
14	1	1	0,3

Ранговый ДА и конкордация Кендалла (Таблица данных 14)						
ДА хи-кв. (N = 14, ss = 2) = 8,844444 p < ,01201						
Козфф. конкордации = ,31587 Средн. ранг r = ,26325						
Перем.	Средний ранг	Сумма рангов	Среднее	Ст. откл		
1 класс (правый)	2,464286	34,50000	0,850000	0,210311		
5 класс (правый)	2,071429	29,00000	0,714286	0,361316		
11 класс (правый)	1,464286	20,50000	0,535714	0,234052		

Различие **значимо** на уровне значимости 0,05.

Диаграмма размаха:



✓ Сделайте вывод по задаче. Можно ли считать, что зрение учащихся ухудшилось?

Контрольные вопросы

1. Назовите условия использования параметрических критериев.
2. Какими должны быть дисперсии групп при использовании параметрических критериев?
3. Какой тип распределения позволяет использовать параметрические критерии?
4. В каком случае лучше воспользоваться непараметрическими критериями?
5. Какому условию должны удовлетворять исследуемые группы для того, чтобы была возможность использовать непараметрические критерии?

Лабораторная работа № 8

Анализ качественных признаков

Краткие сведения из теории

Качественные признаки

В предыдущих лабораторных работах мы производили анализ количественных признаков. Примером таких признаков служат артериальное давление, количество дней госпитализации, время послеродовой активности и т. д. Единицей их измерения могут быть миллиметры ртутного столба, часы или дни. Над значениями количественных признаков можно производить арифметические действия. Можно, например, сказать, что артериальное давление снизилось на какое-то количество единиц. Кроме того, их можно упорядочить: расположить в порядке возрастания или убывания.

! Однако очень многие признаки невозможно измерить числом.

Например, можно быть либо мужчиной, либо женщиной, либо, больным либо здоровым.

Это **качественные признаки**. Эти признаки *не связаны между собой никакими арифметическими соотношениями*, упорядочить их также нельзя.

Единственный способ описания качественных признаков состоит в том, чтобы подсчитать число объектов, **имеющих одно и то же значение**. Кроме того, можно подсчитать, **какая доля от общего числа объектов** приходится на то или иное значение.

Порядковые признаки

Существует еще один вид признаков. Это **порядковые признаки**. Их можно упорядочить, но производить над ними арифметические действия также нельзя. Пример порядкового признака — **состояние больного тяжелое, средней тяжести, удовлетворительное.**

Если часть объектов исследуемой группы характеризуется одним признаком, а вторая часть другим признаком, то можно подсчитать какую долю (p) или процент от общего количества объектов в группе составляют объекты той или иной группы.

Например, если в группе из 100 человек 30 человек — женщины, а 70 — мужчины, то доля p (процент) женщин в группе равен $30/100 = 0,3$ или 30 %, соответственно мужчин — 70 %. Разумеется, группы могут состоять и не из двух классов.

Для характеристики совокупности, которая состоит из двух классов, достаточно указать численность одного из них **если доля одного класса во всей совокупности равна p (вероятность), то доля другого равна $1 - p$.**

Или если известно общее число членов группы N с признаком M , то доля p этих членов можно выразить формулой:

$$p = M/N,$$

или в процентном соотношении

$$p = (M/N) \times 100 \, \%.$$

В некотором смысле доля p аналогична среднему μ по совокупности.

Сравнение долей

Довольно часто необходимо сравнить две группы, характеризующиеся качественным признаком между собой. Для проведения этой процедуры используется **z критерий**.

Критерий z , аналогичный критерию Стьюдента t :

$$z = \frac{\text{Разность выборочных долей}}{\text{Стандартная ошибка разности выборочных долей}}.$$

$$z = \frac{p_1 - p_2}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{p_1 - p_2}{\sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}},$$

где p_1 и p_2 — доля исследуемого признака в первой группе и во второй соответственно.

Если хотя бы для одной выборки условие значения np и $n(1-p)$ больше 5 (где n — объем выборки) не выполняется, то критерий z неприменим, и нужно воспользоваться **точным критерием Фишера**.

Если n_1 и n_2 — объемы двух выборок, то

$$s_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} \text{ и } s_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

О статистически значимом различии долей можно говорить, если значение z окажется «большим».

Критическое значение **z -критерия** находится по таблице критических значений в зависимости от количества членов выборки (вычисляется число степеней свободы) и выбранного уровня значимости. Однако если сравнивать с критическим значением критерия Стьюдента, то $t_{кр}$ подчиняется *распределению Стьюдента*, а $z_{кр}$ — *стандартному нормальному распре-*

делению. При увеличении числа степеней свободы распределение Стьюдента стремится к нормальному, критические значения z можно найти в последней строке таблицы критических значений для критерия Стьюдента.

Поправка Йейтса на непрерывность

Нормальное распределение служит лишь приближением для распределения z . При этом оценка P оказывается заниженной, и нулевая гипотеза может быть неправильно отвергнута. Причина состоит в том, что z принимает только дискретные значения, тогда как приближающее его нормальное распределение непрерывно. Для компенсации излишнего «оптимизма» критерия z введена поправка Йейтса на зываемая также поправкой на непрерывность. С учетом этой поправки выражение для z имеет следующий вид:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

Поправка Йейтса слегка уменьшает значение z , уменьшая тем самым расхождение с нормальным распределением.

Таблицы сопряженности: критерий χ^2 для таблицы 2×2

Довольно часто исследуемый объект может иметь *несколько* качественных признаков, например: человек болен и при этом принял или не принял соответствующее лекарство. В этом случае данные эксперимента записывают в так называемые **таблицы сопряженности**.

Пример таблицы сопряженности:

	Прививка	
Заболел	да	нет
да	2	12
нет	15	4

Таблица сопряженности — представляет собой таблицу (m на n , где m — значения первой переменной, n — значения второй переменной). В таблице выше $m = 2$ (заболел: да или нет), $n = 2$ (прививка: да или нет). Такие таблицы называются **таблицами сопряженности 2×2** .

Для анализа таблиц сопряженности используется критерий χ^2 .

Для понимания смысла анализа таблиц сопряженности необходимо учитывать, что если переменные между собой не связаны, то значения в таблице сопряженности (**ожидаемые значения**), при равной численности

в группах, будут примерно одинаковыми. Иными словами, исходя из вышеприведенного примера, количество *заболевших* не будет зависеть от наличия или отсутствия прививки.

Однако в нашей таблице значения в клетках (**наблюдаемые значения**) существенно отличаются, что наводит на мысль о возможной статистической значимости этих различий. *На сравнении ожидаемых и наблюдаемых значений и основан критерий χ^2 .*

Критерий χ^2 (читается «хи-квадрат») не требует никаких предположений относительно параметров совокупности, из которой извлечены выборки, — это непараметрический критерий.

Определяется критерий χ^2 следующим образом:

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

где O — наблюдаемое число в клетке таблицы сопряженности, E — ожидаемое число в той же клетке. Суммирование проводится по всем клеткам таблицы.

Нулевая гипотеза в подобных задачах звучит следующим образом: переменные не связаны между собой, т. е. являются независимыми, видимые различия в клетках таблицы сопряженности случайны и статистически незначимы.

✓Что такое уровень значимости?

Полученное значение критерия χ^2 характеризует значимость этих различий. Расчетное значение критерия сравнивается с **критическим значением критерия $\chi^2_{кр}$** (значение находится по таблице для заданного уровня значимости) и если полученное значение критерия χ^2 больше критического, то нулевая гипотеза об отсутствии связи между переменными отклоняется.

Применение критерия χ^2 правомерно, если ожидаемое число в любой из клеток больше или равно 5 (в противном случае мы вынуждены использовать **точный критерий Фишера**). Это условие аналогично условию применимости критерия z . Критическое значение χ^2 зависит от размеров таблицы сопряженности, то есть от числа строк таблицы и числа возможных столбцов таблицы. Размер таблицы выражается числом степеней свободы v :

$$v = (r - 1)(c - 1),$$

где r — число строк, а c — число столбцов.

Поправка Йейтса

Приведенная формула для χ^2 в случае таблицы 2×2 (то есть при 1 степени свободы) дает несколько завышенные значения, т. е. повышается вероятность совершить ошибку I рода. Это вызвано тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных значений χ^2 дискретен. На практике это приведет к тому, что нулевая гипотеза будет отвергаться слишком часто. Чтобы компенсировать этот эффект, в формулу вводят **поправку Йейтса**:

$$\chi^2 = \sum \frac{\left(|O - E| - \frac{1}{2}\right)^2}{E}.$$

Поправка Йейтса применяется только при $v = 1$, т. е. для таблиц 2×2 .

✓Что означает ошибка I рода?

✓Что означает ошибка II рода?

Для таблиц сопряженности размером 2×2 критерий χ^2 применим только в случае, когда все **ожидаемые числа больше 5**. С таблицами большего размера критерий χ^2 применим, если все ожидаемые числа не меньше 1 и доля клеток с ожидаемыми числами меньше 5 не превышает 20 %. При невыполнении этих условий критерий χ^2 может дать ложные результаты. В этой ситуации есть выход: можно собрать дополнительные данные, однако это не всегда осуществимо или объединить несколько строк или столбцов.

➤Порядок применения критерия χ^2

1. По имеющимся данным построить таблицу сопряженности.
2. Подсчитать число объектов в каждой строке и в каждом столбце, найти, какую долю от общего числа объектов составляют эти величины.
3. Зная эти доли, подсчитать с точностью до двух знаков после запятой ожидаемые числа — количество объектов, которое попало бы в каждую клетку таблицы, если бы связь между строками и столбцами отсутствовала.
4. Найти величину, характеризующую различия наблюдаемых и ожидаемых значений. Если таблица сопряженности имеет размер 2×2 , применить поправку Йейтса.
5. Вычислить число степеней свободы, и выбрать уровень значимости.
6. Определить критическое значение χ^2 . Сравнить полученное значение с критическим значением.
7. Если полученное значение критерия χ^2 больше критического, то нулевая гипотеза об отсутствии связи между переменными отклоняется, в противном случае мы остаемся в рамках нулевой гипотезы.

Точный критерий Фишера

Критерий χ^2 годится для анализа таблиц сопряженности 2×2 , если ожидаемые значения в любой из ее клеток не меньше 5. Когда число наблюдений невелико, это условие не выполняется и критерий χ^2 не применим. В этом случае используют точный критерий Фишера. Он основан на переборе всех возможных вариантов заполнения таблицы сопряженности при данной численности групп, поэтому, чем она меньше, тем проще его применить. Например, нулевая гипотеза состоит в том, что между лечением и исходом нет никакой связи. Тогда вероятность получить некоторую таблицу равна.

Обозначения, используемые в точном критерии Фишера

		Суммы по строкам	
Суммы по столбцам	O_{11}	O_{12}	R_1
	O_{21}	O_{22}	R_2
	C_1	C_2	N

$$P = \frac{R_1! R_2! C_1! C_2!}{N! O_{11}! O_{12}! O_{21}! O_{22}!},$$

где R_1 и R_2 — суммы по строкам (число больных, лечившихся первым и вторым способом), C_1 и C_2 — суммы по столбцам (число больных с первым и вторым исходом). O_{11} , O_{12} , O_{21} и O_{22} — числа в клетках, N — общее число наблюдений. Восклицательный знак обозначает факториал (Факториал числа — произведение всех целых чисел от этого числа до единицы $n! = n \times (n - 1) \times (n - 2) \times \dots \times 2 \times 1$. Например, $4! = 4 \times 3 \times 2 \times 1 = 24$. Факториал нуля равен единице.) Построив все остальные варианты заполнения таблицы, возможные при данных суммах по строкам и столбцам, по этой же формуле рассчитывают их вероятность. Вероятности, которые не превосходят вероятность исходной таблицы (включая саму эту вероятность), суммируют. Полученная сумма — это величина P для двустороннего варианта точного критерия Фишера.

В отличие от критерия χ^2 , существуют одно- и двусторонний варианты точного критерия Фишера.

✓ Задача

Тромбоз шунта у больных на гемодиализе

Гемодиализ позволяет сохранить жизнь людям, страдающим хронической почечной недостаточностью. При гемодиализе кровь больного пропускают через искусственную почку — аппарат, удаляющий из крови про-

дукты обмена веществ. Искусственная почка подсоединяется к артерии и вене больного: кровь из артерии поступает в аппарат и оттуда, уже очищенная — в вену. Так как гемодиализ проводится регулярно, больному устанавливают артериовенозный шунт. В артерию и вену на предплечье вводят тефлоновые трубки; их концы выводят наружу и соединяют друг с другом. При очередной процедуре гемодиализа трубки разъединяют между собой и присоединяют к аппарату. После диализа трубки вновь соединяют, и кровь течет по шунту из артерии в вену. Завихрения тока крови в местах соединения трубок и сосудов приводят к тому, что шунт часто тромбируется. Тромбы приходится регулярно удалять, а в тяжелых случаях даже менять шунт. Руководствуясь тем, что аспирин препятствует образованию тромбов, Г. Хартер и соавт. решили проверить, нельзя ли снизить риск тромбоза назначением небольших доз аспирина (160 мг/сут). Было проведено контролируемое испытание. Все больные, согласившиеся на участие в испытании и не имевшие противопоказаний к аспирину, были случайным образом разделены на две группы: 1-я получала плацебо, 2-я — аспирин. Ни врач, дававший больному препарат, ни больной не знали, был это аспирин или плацебо. Такой способ проведения испытания (он называется двойным слепым) исключает «подсуживание» со стороны врача или больного и, хотя технически сложен, дает наиболее надежные результаты. Исследование проводилось до тех пор, пока общее число больных с тромбозом шунта не достигло 25. Группы практически не различались по возрасту, полу и продолжительности лечения гемодиализом.

Можно ли говорить о статистически значимом различии доли больных с тромбозом, а тем самым об эффективности аспирина?

Таблица результатов исследования представлена в следующем виде (типичном для медицинских исследований):

№	Фамилия	Пол	Год рождения	Аспирин	Тромбоз	Дата исследования
1	Абрамов	муж.	1970	да	нет	11.02.2015
2	Адамова	жен.	1965	нет	да	12.02.2015
3	Алексеев	жен.	1960	да	нет	15.02.2015
4	Астафьев	муж.	1967	да	да	14.02.2015
5	Баринова	жен.	1965	нет	да	15.02.2015
6	Богданов	муж.	1961	нет	да	16.02.2015
7	Бочков	муж.	1967	да	нет	17.02.2015
8	Воробьев	муж.	1965	нет	да	19.02.2015
9	Герасимова	жен.	1975	да	да	19.02.2015
10	Громов	муж.	1967	да	нет	20.02.2015
11	Иванов	муж.	1961	нет	да	21.02.2015
12	Петров	муж.	1967	нет	да	22.02.2015
13	Сидоров	муж.	1965	нет	да	23.02.2015

№	Фамилия	Пол	Год рождения	Аспирин	Тромбоз	Дата исследования
14	Ивашкин	муж.	1975	нет	да	24.02.2015
15	Богданович	муж.	1967	нет	да	25.02.2015
16	Сердюков	муж.	1961	нет	да	26.02.2015
17	Иванова	жен.	1967	нет	да	27.02.2015
18	Петрова	жен.	1965	нет	да	23.02.2015
19	Сидорова	жен.	1975	нет	да	01.03.2015
20	Ивашкина	жен.	1967	нет	да	02.03.2015
21	Богдановича	жен.	1961	нет	да	03.03.2015
22	Сердюкова	жен.	1967	нет	да	04.03.2015
23	Козлов	муж.	1965	нет	да	07.03.2015
24	Лосев	муж.	1975	нет	да	06.03.2015
25	Лысенко	муж.	1967	нет	да	07.03.2015
26	Лысенков	муж.	1961	да	да	11.03.2015
27	Собакевич	муж.	1967	да	да	09.03.2015
28	Чичиков	муж.	1965	да	да	10.03.2015
29	Козлова	жен.	1975	да	да	11.03.2015
30	Лосева	жен.	1967	нет	нет	14.03.2015
31	Лысенко	жен.	1961	нет	нет	13.03.2015
32	Лысенкова	жен.	1967	нет	нет	14.03.2015
33	Собакевич	жен.	1965	нет	нет	15.03.2015
34	Чичикова	жен.	1975	нет	нет	18.03.2015
35	Петренко	муж.	1967	нет	нет	17.03.2015
36	Нестеров	муж.	1967	нет	нет	18.03.2015
37	Старовойтов	муж.	1965	да	нет	19.03.2015
38	Степанов	муж.	1975	да	нет	21.03.2015
39	Степкин	муж.	1967	да	нет	21.03.2015
40	Пупкин	муж.	1961	да	нет	22.03.2015
41	Нестерова	жен.	1967	да	нет	26.03.2015
42	Старовойтова	жен.	1965	да	нет	26.03.2015
43	Степанова	жен.	1975	да	нет	25.03.2015
44	Степкина	жен.	1967	да	нет	26.03.2015

Ход выполнения работы

Нулевая гипотеза: аспирин не влияет на возникновение тромбоза шунта.

Уровень значимости принимается 0,05.

1. Для решения поставленной задачи необходимо сформировать таблицу сопряженности (в нашем случае это таблица 2x2). Для этого воспользуемся возможностями **сводной таблицы MS Excel**.

2. Скопируем исходную таблицу на «Лист 1» программы MS Excel.

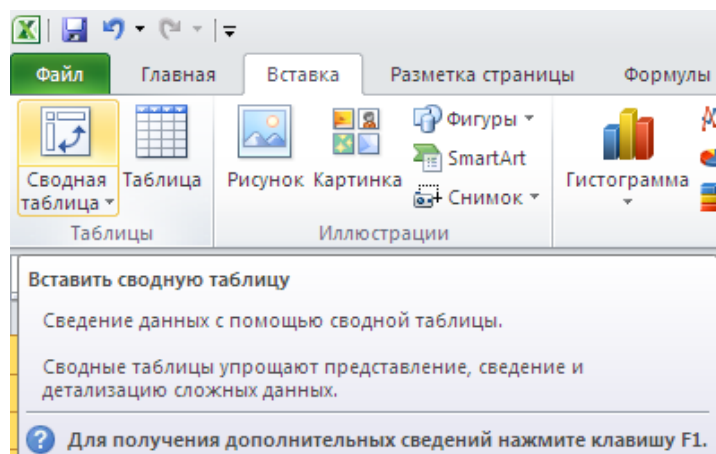
Основы статистики. Лабораторный практикум

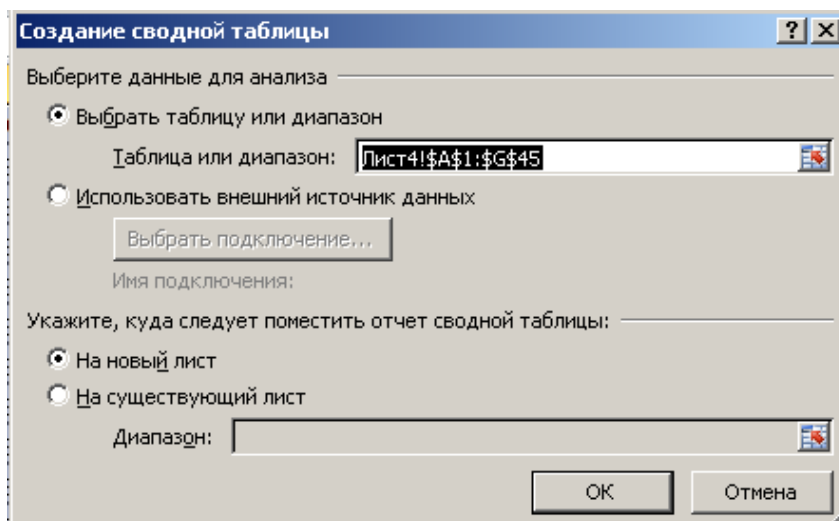
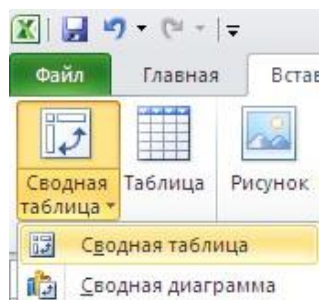
№	Фамилия	Пол	Год рождения	Аспирин	Тромбоз	Дата исследования
1	Абрамов	муж	1970	да	нет	11.02.2015
2	Адамова	жен	1965	нет	да	12.02.2015
3	Алексеев	жен	1960	да	нет	15.02.2015
4	Астафьев	муж	1967	да	да	14.02.2015
5	Баринова	жен	1965	нет	да	15.02.2015
6	Богданов	муж	1961	нет	да	16.02.2015
7	Бочков	муж	1967	да	нет	17.02.2015
8	Воробьев	муж	1965	нет	нет	19.02.2015
9	Герасимова	жен	1975	да	да	19.02.2015
10	Громов	муж	1967	да	нет	20.02.2015
11	Иванов	муж	1961	нет	да	21.02.2015
12	Петров	муж	1967	нет	да	22.02.2015
13	Сидоров	муж	1965	нет	да	23.02.2015
14	Ивашкин	муж	1975	нет	да	24.02.2015
15	Богданович	муж	1967	нет	да	25.02.2015
16	Сердюков	муж	1961	нет	да	26.02.2015
17	Иванова	жен	1967	нет	да	27.02.2015

3. Выделите всю таблицу.

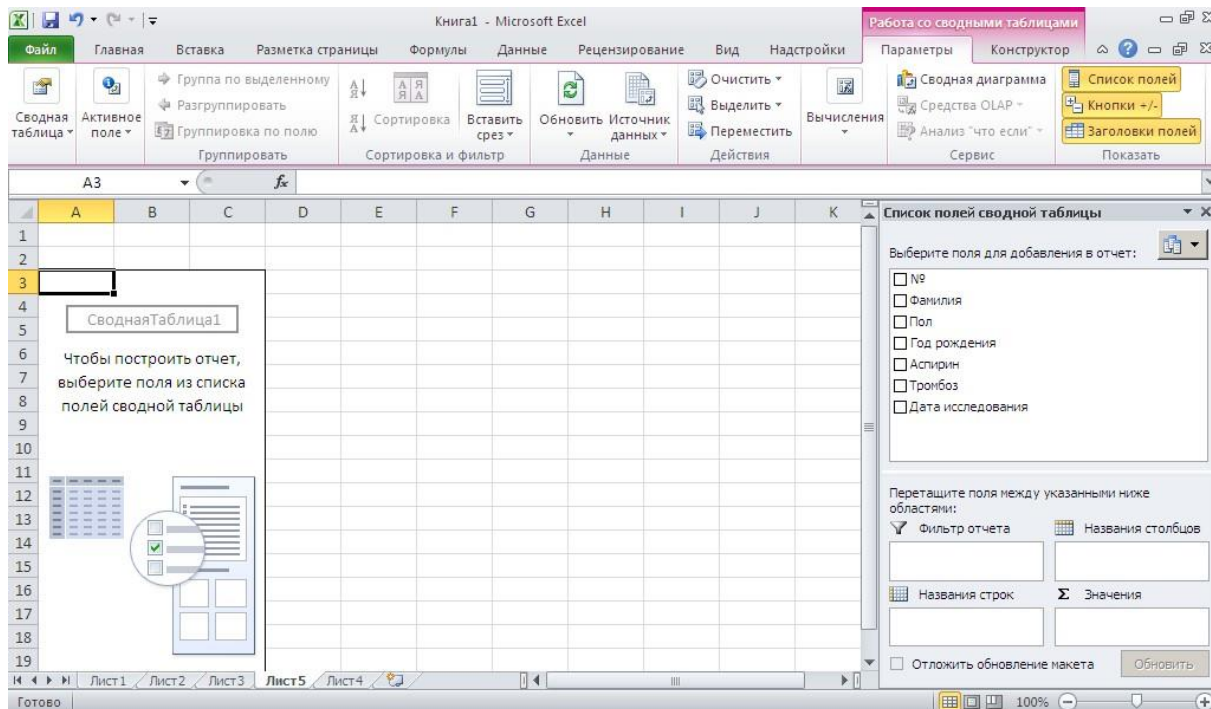
	A	B	C	D	E	F	G
33	32	Лысенкова	жен	1967	нет	нет	14.03.2015
34	33	Собакевич	жен	1965	нет	нет	15.03.2015
35	34	Чичикова	жен	1975	нет	нет	18.03.2015
36	35	Петренко	муж	1967	нет	нет	17.03.2015
37	36	Нестеров	муж	1967	нет	нет	18.03.2015
38	37	Старовойтов	муж	1965	да	нет	19.03.2015
39	38	Степанов	муж	1975	да	нет	21.03.2015
40	39	Степкин	муж	1967	да	нет	21.03.2015
41	40	Пупкин	муж	1961	да	нет	22.03.2015
42	41	Нестерова	жен	1967	да	нет	26.03.2015
43	42	Старовойтова	жен	1965	да	нет	26.03.2015
44	43	Степанова	жен	1975	да	нет	25.03.2015
45	44	Степкина	жен	1967	да	нет	26.03.2015

4. Далее: *Вставка* — *Сводная таблица* — *Сводная таблица* — *Ок*.





5. В появившемся макете сводной таблице укажите название строк, столбцов и значений так, как показано на иллюстрациях ниже.



При помощи сводной таблицы можно построить таблицу сопряженности (пример смотрите выше). Построение, внешний вид и содержимое сводной таблицы зависит от задания (или того, что вы желаете отобразить/проследить).

В нашем случае необходимо отобразить количество людей с тромбозом шунта и без него, при этом должна содержаться информация был или не был введен аспирин. Т. е. исследуемый объект имеет два признака: тромбоз есть или нет (тромбоз да/нет), аспирин введен или нет (аспирин да/нет).

6. Для лучшей наглядности сводной таблицы во вкладке «*Конструктор*» укажите *Макет отчета* как «*Показать в форме структуры*».

7. Получившуюся сводную таблицу (таблица сопряженности) скопируйте в отдельный документ Word и при необходимости дооформите ее.

Количество по полю Фамилия		Тромбоз		
Аспирин		да	нет	Общий итог
да		6	12	18
нет		19	7	26
Общий итог		25	19	44

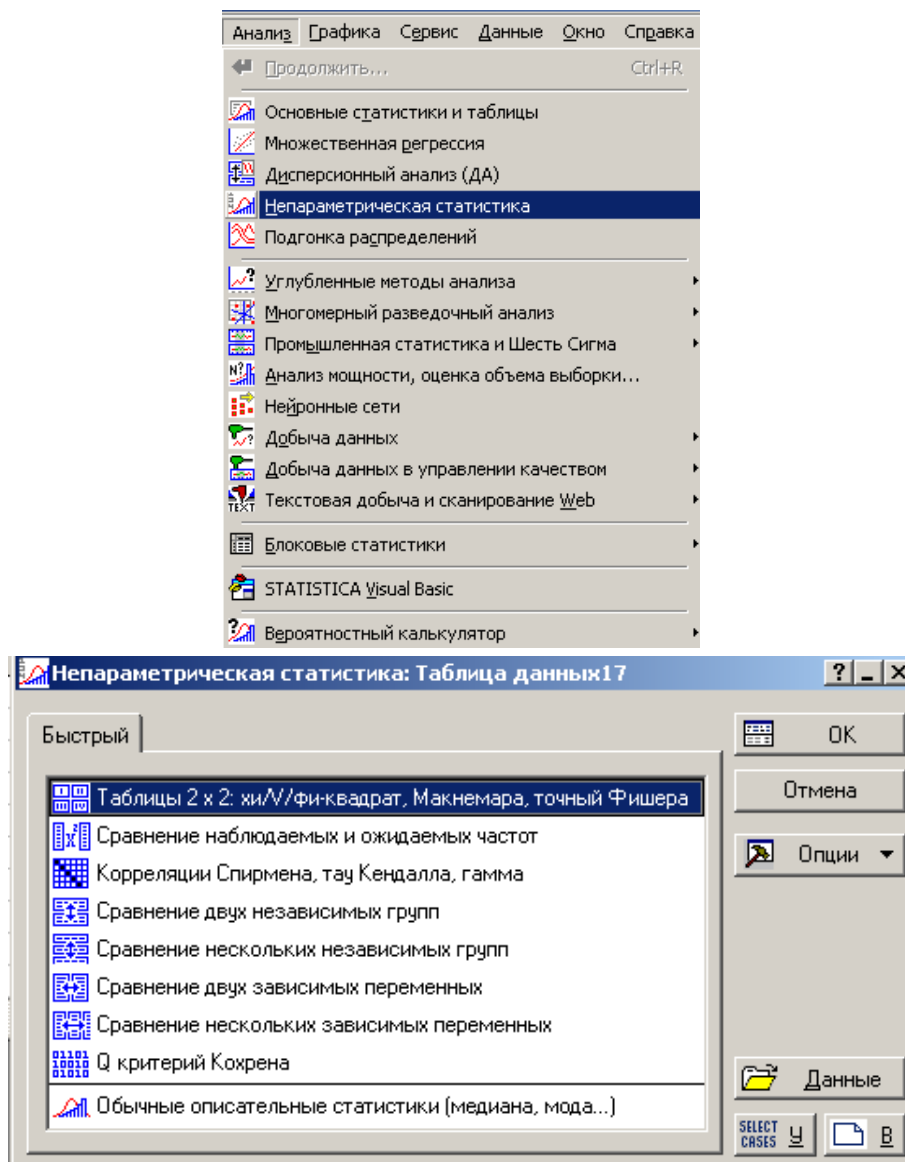
8. В результате предыдущих действий вы получили таблицу сопряженности 2x2, в принципе ее можно было бы получить и вручную, но возможности сводной таблицы Excel значительно упрощают процесс построения (особенно при большом количестве данных эксперимента).

Стоит заметить, что для подобных целей можно воспользоваться возможностями **фильтров MS Excel**.

Таблица сопряженности показывает количество человек с тромбозом, которым был введен аспирин и количество больных тромбозом, которым вместо аспирина давали плацебо. Как видно из таблицы налицо явные преимущества аспирина. Теперь видя только нужные для анализа цифры необходимо оценить статистическую значимость видимых различий. Воспользуемся критерием хи-квадрат.

9. Откройте программу «Statistica», создайте новый документ.

10. Далее: *Анализ — Непараметрическая статистика — Таблицы 2x2*



11. В появившемся окне введите значения из полученной таблицы сопряженности:

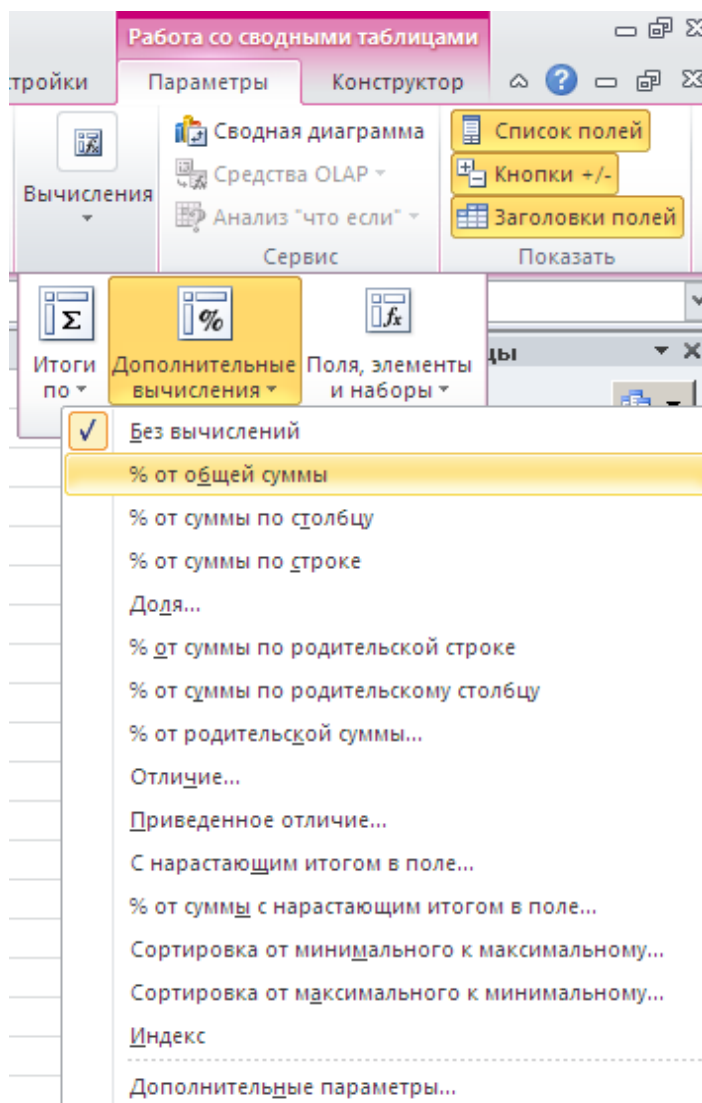
Диалоговое окно «Таблицы 2 x 2 : Таблица данных1» в SPSS. В первом снимке все значения частот равны 0. Во втором снимке значения частот установлены на 6, 12, 19 и 7.

12. Полученная таблица результата анализа. Так как в нашем случае анализ проводился таблицы сопряженности 2x2, то необходимо учитывать поправку Йейтса. Исходя из полученных значений **критерия хи-квадрат** и **вероятности p**, следует заключить, что видимые различия в клетках таблицы сопряженности значимы. Поэтому нулевая гипотеза отвергается. Аспирин действительно положительно влияет на снижение вероятности возникновения тромбоза шунта.

	Таблица 2x2 (Таблица данных1)		
	Столб. 1	Столб. 2	Сумма строк
Частоты, строка 1	6	12	18
Процент от общего	13,636%	27,273%	40,909%
Частоты, строка 2	19	7	26
Процент от общего	43,182%	15,909%	59,091%
Сумма по столбцам	25	19	44
Процент от общего	56,818%	43,182%	
Хи-квадрат (ст.св.=1)	6,85	p= ,0089	
V-квадрат (ст.св.=1)	6,69	p= ,0097	
Поправка Йетса	5,32	p= ,0210	
Фи коэффициент	,15563		
Фишера p, односторонний		p= ,0102	
двусторонний		p= ,0137	
Макнемара Хи-квадрат (A/D)	0,00	p=1,0000	
Хи-квадрат (B/C)	1,16	p= ,2812	

✓ **Результат скопируйте в документ Word. Сделайте ваш собственный вывод.**

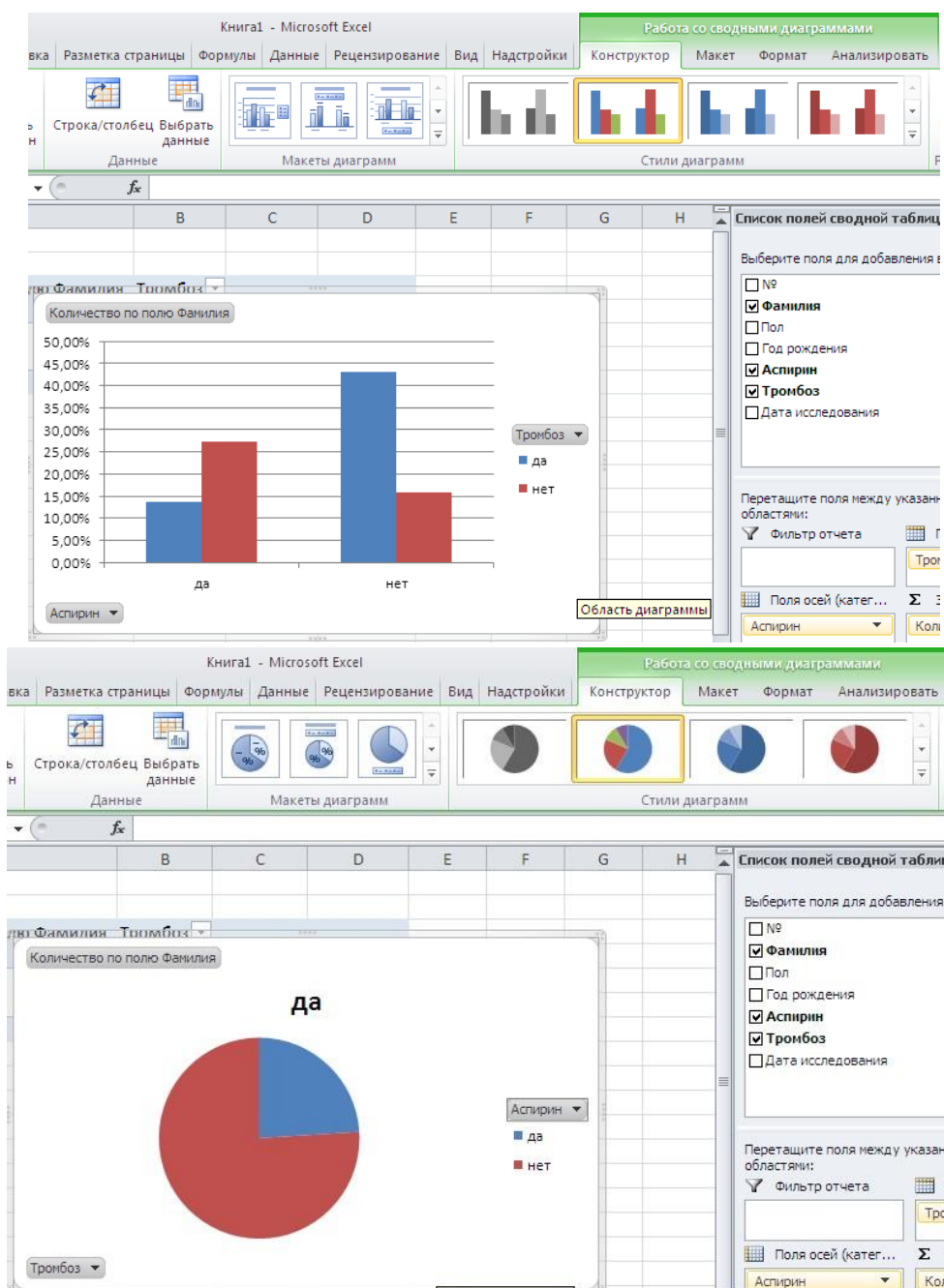
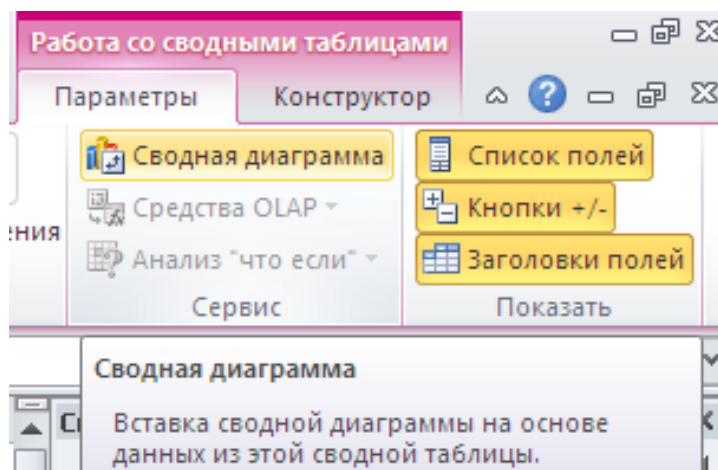
13. Так как в подобного рода анализе часто возникает потребность в указании процентного соотношения в таблице сопряженности, то вновь вернемся к сводной таблице Excel. И во вкладке «**Параметры**» выберите «**Вычисления**» — «**Дополнительные вычисления** — % от общей суммы».

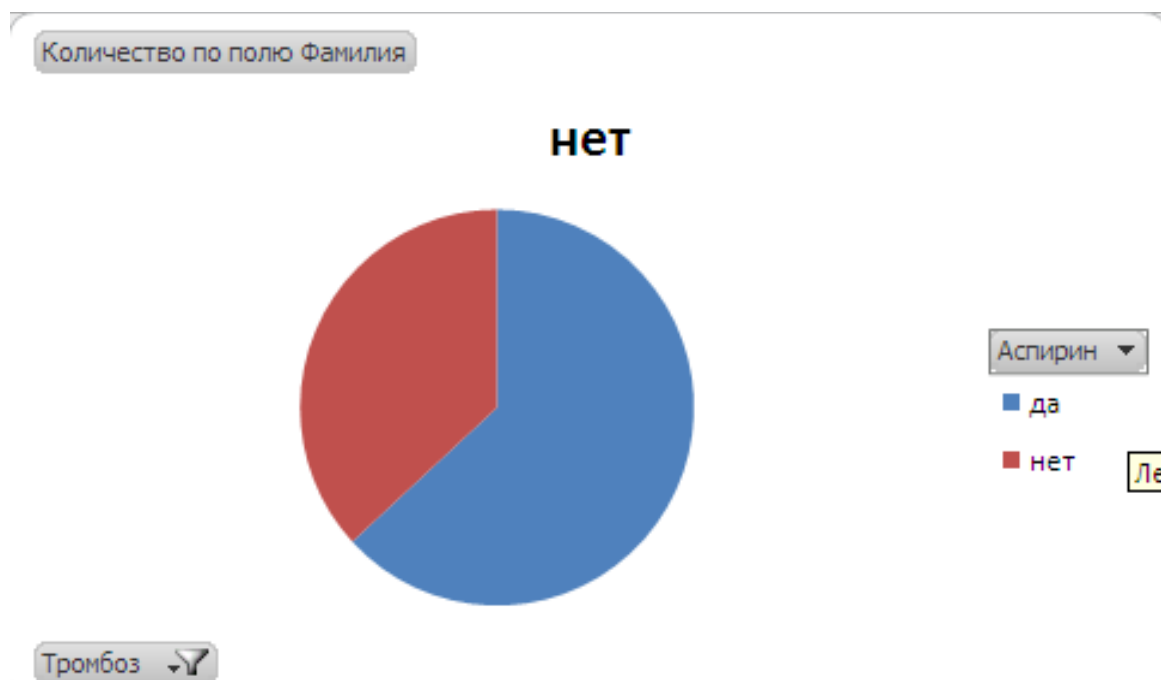


14. Полученный результат скопируйте в документ Word.

Количество по полю	Фамилия	Тромбоз	
Аспирин	да	нет	Общий итог
да	13,64%	27,27%	40,91%
нет	43,18%	15,91%	59,09%
Общий итог	56,82%	43,18%	100,00%

15. Постройте три типа сводных диаграмм как указано ниже: «**Параметры**» — «**Сводная диаграмма**». Скопируйте в отчет.





16. Сделайте выводы о проделанном анализе.

✓Задание

Т. Бишоп изучил эффективность высокочастотной стимуляции нерва в качестве обезболивающего средства при удалении зуба. Все больные подключились к прибору, но в одних случаях он работал, в других был выключен. Ни стоматолог, ни больной не знали, включен ли прибор. Позволяют ли следующие данные считать высокочастотную стимуляцию нерва действенным анальгезирующим средством?

	Прибор включен	Прибор выключен
Боли нет	24	3
Боль есть	6	17

Контрольные вопросы

1. Какие данные называются качественными?
2. Приведите примеры качественных данных.
3. Для чего используются сводные таблицы MS Excel?
4. Что такое таблица сопряженности?
5. Для чего используют критерий хи-квадрат?

Лабораторная работа № 9

Классификация. Кластерный и дискриминантный анализы

Краткие сведения из теории

Классификацией называют разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы.

Различают классификацию при наличии **обучающих выборок** (*дискриминантный анализ*) и классификацию **без обучения**. К классификации **без обучения** относят методы автоматической *классификации* (*кластерный анализ*).

Кластерный анализ

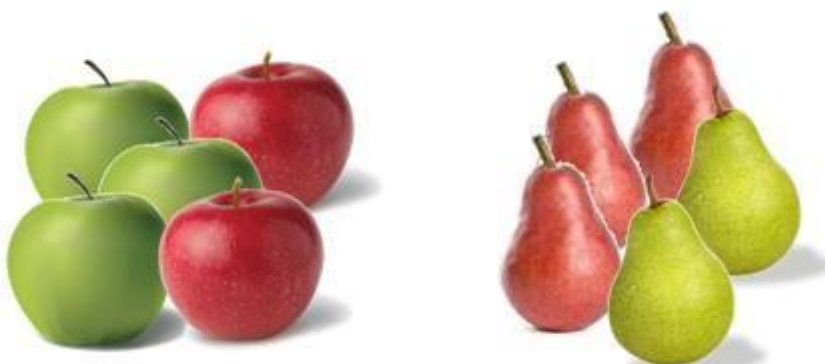
Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры.

Простейший пример

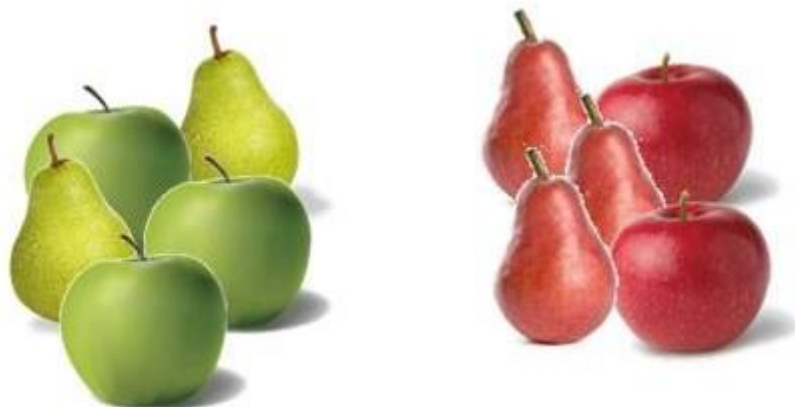
Представьте, что у нас есть несколько фруктов:



Как разбить их на группы, объединив похожие плоды? Самый очевидный способ — отделить груши от яблок:



Но, с другой стороны, можно сгруппировать фрукты по цветам:



Но можно сформировать больше групп, основываясь на цвете и на типе фрукта:



А если появится новый, неопознанный фрукт?



В какую группу его отнести? Или выделить под него новую группу?

В научных исследованиях задачи возникают куда более сложные и трудоемкие нежели, чем в выше приведенном примере. При наличии огромных массивов разнородных данных осуществить подобное разделение на группы (классифицировать объекты) — непростая задача.

Кластерным анализом называются разнообразные формализованные процедуры построения классификаций объектов. Лидирующей в развитии кластерного анализа наукой является биология.

Другими словами, задача кластерного анализа состоит в разбиении исходной совокупности объектов на группы схожих, близких между собой объектов. Эти группы называют кластерами.

Предмет **кластерного анализа** (от англ. «cluster» — гроздь, пучок, группа) был сформулирован в 1939 г. психологом Робертом Трионом. «Классиками» кластерного анализа являются американские систематики Роберт Сокэл и Питер Снит. Одно из важнейших их достижений в этой области — книга «Начала численной таксономии», выпущенная в 1963 году. В соответствии с основной идеей авторов, классификация должна строиться не на смешении плохо формализованных суждений о сходстве и родстве объектов, а на результатах формализованной обработки результатов математического вычисления сходства/отличий классифицируемых объектов. Для выполнения этой задачи нужны были соответствующие процедуры, разработкой которых и занялись авторы.

Еще пример, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (т. е. с обезьянами), чем с «отдаленными» членами семейства млекопитающих (например, собаками) и т. д.

Кластерный анализ позволяет из множества всех объектов выделять группы объектов, похожих по определенным признакам.

➤ Основные этапы кластерного анализа таковы:

1. Выбор сравниваемых друг с другом объектов.
2. Выбор множества признаков (характеристик), по которому будет проводиться сравнение и описание объектов по этим признакам.
3. Вычисление меры сходства между объектами (или меры различия объектов) в соответствии с избранной метрикой.
4. Группировка объектов в кластеры с помощью той или иной процедуры объединения.
5. Проверка применимости полученного кластерного решения (проверка построенной модели).

Кластер — это тип объектов, схожих по определенному признаку.

Если вы взглянете на географическую карту и увидите на ней горы или посмотрите на звездное небо и увидите там созвездия, то поймете, что такое кластеры.

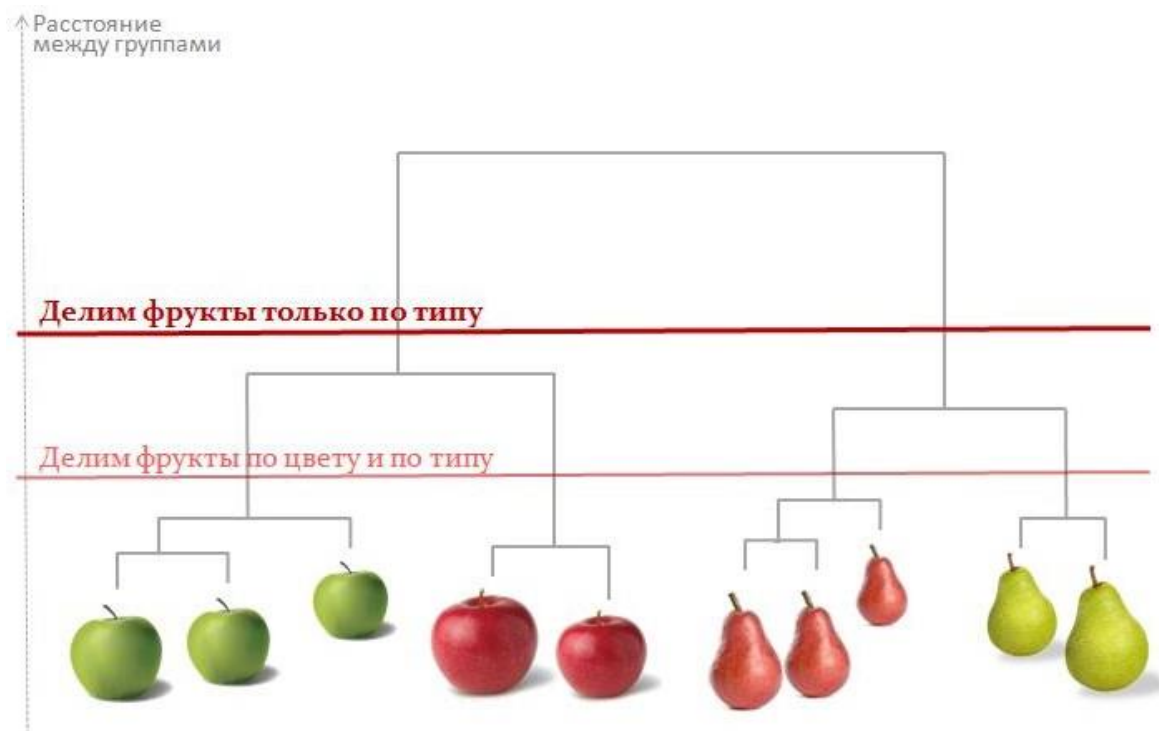
Важно еще раз отметить, что задача кластеризации не является тривиальной. Сложность задач кластерного анализа состоит в том, что *реальные объекты являются многомерными*, т. е. описываются не одним, а несколькими параметрами, и объединение объектов в группы проводится в пространстве многих измерений, что весьма непросто. Кроме того, данные могут носить нечисловой характер.

Методы кластеризации

В целом методы кластеризации делятся на **агломеративные** (от слова агломерат — скопление) и итеративные **дивизивные** (от слова division — деление, разделение).

В **агломеративных**, или объединительных, методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде **дендрограммы**, или дерева объединения. Это удобное представление позволяет наглядно представить кластеризацию агломеративными алгоритмами.

На каждом шаге ее составления алгоритм находит два самых близких объекта по **расстоянию**, подсчитанному специальным методом. Это расстояние откладывается по оси y . В итоге, исходя из расстояний на дендрограмме, можно определить необходимое количество групп.



Дендрограмма

Исходными данными для анализа могут быть собственно объекты и их параметры. Данные для анализа могут быть также представлены матрицей расстояний между объектами.

Расстояние между объектами — одна из мер сходства, чем меньше расстояние между объектами, тем они более схожи.

В биологических науках кластеризация имеет множество приложений в самых разных областях. Например, в биоинформатике с помощью неё анализируются сложные сети взаимодействующих генов, состоящие порой из сотен или даже тысяч элементов. Кластерный анализ позволяет выделить подсети, узкие места, концентраторы и другие скрытые свойства изучаемой системы, что позволяет в конечном счете узнать вклад каждого гена в формирование изучаемого феномена.

Дискриминантный анализ

Дискриминантный анализ является одним из методов многомерного статистического анализа.

Цель дискриминантного анализа состоит в том, чтобы на основе измерения различных характеристик (признаков, параметров) объекта **классифицировать** его, т. е. отнести к одной из нескольких групп (классов) некоторым оптимальным способом.

Под оптимальным способом понимается либо минимум средних потерь, либо минимум вероятности ложной классификации.

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

Типичные **области применения** дискриминантного анализа — биология, медицина, управление производством, экономика, геология, контроль качества. В медицине объектом исследования является пациент, когда по результатам измерений различных параметров, проведения диагностических тестов врач определяет, например, необходимо ли хирургическое вмешательство при лечении. Медик может регистрировать различные переменные, относящиеся к состоянию больного, чтобы выяснить, какие переменные лучше предсказывают, что пациент, вероятно, **выздоровел полностью** (группа 1), **частично** (группа 2) или **совсем не выздоровел** (группа 3). Биолог может записать различные характеристики сходных типов (групп) цветов, чтобы затем провести анализ дискриминантной функции, наилучшим образом разделяющей типы или группы.

Задача дискриминантного анализа

Предположим, имеется n объектов с m характеристиками. В результате измерений каждый объект характеризуется вектором из m характеристик: $x_1 \dots x_m$, $m > 1$. Задача состоит в том, чтобы по результатам измерений отнести объект к одной из нескольких ранее определенных групп (классов) G_1, \dots, G_k , $k \geq 2$.

Иными словами, нужно построить **решающее правило**, позволяющее по результатам измерений параметров объекта указать группу, к которой он принадлежит. Число групп заранее известно, также известно, что объект заведомо принадлежит к определенной группе.

Рассмотрим простой пример

Предположим, что вы измеряете рост в случайной выборке из 50 мужчин и 50 женщин. Женщины в среднем не так высоки, как мужчины, и эта разница должна найти отражение для каждой группы средних (для переменной *Рост*). Поэтому переменная «Рост» позволяет вам провести **дискриминацию** между мужчинами и женщинами лучше, чем, например, вероятность, выраженная следующими словами: «Если человек большой, то это, скорее всего, мужчина, а если маленький, то это вероятно женщина».

Основная идея дискриминантного анализа заключается в том, чтобы *определить, отличаются ли совокупности по среднему значению какой-либо переменной (или линейной комбинации переменных), и затем использовать эту переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе*. Другими словами, вы хотите построить «модель», позволяющую лучше всего предсказать, к какой совокупности будет принадлежать тот или иной образец.

➤ Алгоритм дискриминантного анализа

Решение задач дискриминации (дискриминантный анализ) состоит в разбиении всего выборочного пространства (множества реализации всех рассматриваемых многомерных случайных величин) на некоторое число областей.

Пусть имеются две генеральные совокупности X и Y , имеющие многомерный (трехмерный) нормальный закон распределения с неизвестными, но равными ковариационными матрицами.

Из этих совокупностей взяты **обучающие выборки** объемами n_1 и n_2 соответственно:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots \\ x_{n_1 1} & x_{n_1 2} & x_{n_1 3} \end{pmatrix}; Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ \dots & \dots & \dots \\ y_{n_2 1} & y_{n_2 2} & y_{n_2 3} \end{pmatrix}$$

Целью дискриминантного анализа в этом случае является отнесение нового наблюдения (строки) из матрицы:

$$Z = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ \dots & \dots & \dots \\ z_{l1} & z_{l2} & z_{l3} \end{pmatrix} \text{ либо к } X, \text{ либо к } Y.$$

✓ Кластерный анализ. Задача

Классификация населенных пунктов, расположенных в зоне радиоактивного загрязнения

Для классификации населенных пунктов по степени первоочередности проведения мероприятий радиационного или социального характера были учтены следующие факторы:

- демографический (численность населения, возрастная структура населения i -го населенного пункта);
- хозяйственный (отношение числа жителей к числу коров);
- радиационный (средние значения суммарной годовой эффективной индивидуальной дозы и удельной активности молока по i -му населенному пункту).

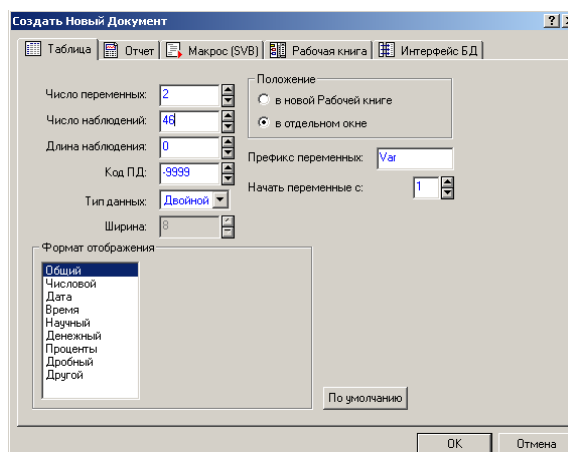
На основе этой информации для каждого населенного пункта были рассчитаны **социально-экономические и радиологические показатели**, которые затем были *отнормированы* на максимальное значение. Соответствующие значения в баллах, присваиваемые i -му населенному пункту приведены в таблице. Максимальное значение — 1 балл.

№ НП	Соц.-экон.	Радиолог.
1	0,24719329	0,29540061
2	0,49097333	0,49549701
3	0,76815313	0,24014938
4	0,83789641	0,35430514
5	0,92087343	0,31977715
6	0,83693199	0,40867036
7	0,64208613	0,30919644
8	0,75447239	0,28769678
9	0,84431659	0,33989095
10	0,43312923	0,24221189
11	0,92254975	0,28911137
12	0,823976	0,24254329
13	0,96296219	0,33465852
14	0,80014316	0,22299802
15	0,82842084	0,69562283
16	1	0,36552184
17	0,71545294	0,35440414

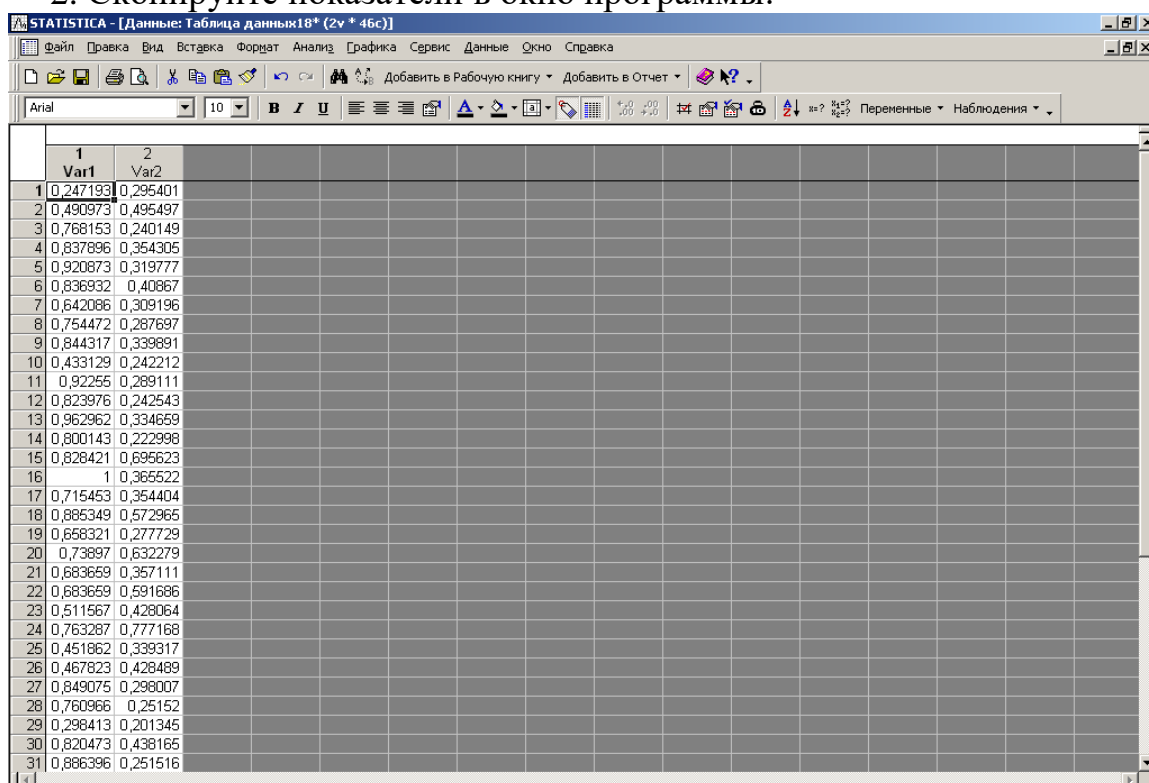
№ НП	Соц.-экон.	Радиолог.
18	0,8853492	0,57296466
19	0,65832115	0,27772873
20	0,73896951	0,63227918
21	0,68365851	0,35711119
22	0,68365851	0,59168568
23	0,51156682	0,42806439
24	0,76328687	0,77716795
25	0,45186182	0,33931741
26	0,46782343	0,42848911
27	0,84907526	0,29800686
28	0,76096616	0,25151952
29	0,29841345	0,20134517
30	0,82047266	0,43816498
31	0,88639612	0,2515159
32	0,76052719	0,53358935
33	0,8274442	0,36113046
34	0,89421902	1
35	0,38097797	0,45873837
36	0,7502056	0,30138691
37	0,61300179	0,51226607
38	0,44290376	0,36935574
39	0,85262386	0,23098893
40	0,78977027	0,39782878
41	0,74822401	0,52675873
42	0,89249236	0,36927656
43	0,73758751	0,65779693
44	0,61224609	0,24290559
45	0,77375606	0,23284717
46	0,93288537	0,2708384

Необходимо разделить населенные пункты на соответствующие классы при помощи процедуры кластерного анализа в программе «Statistica».

1. Запустите программу и укажите исходные настройки как показано на рисунке:

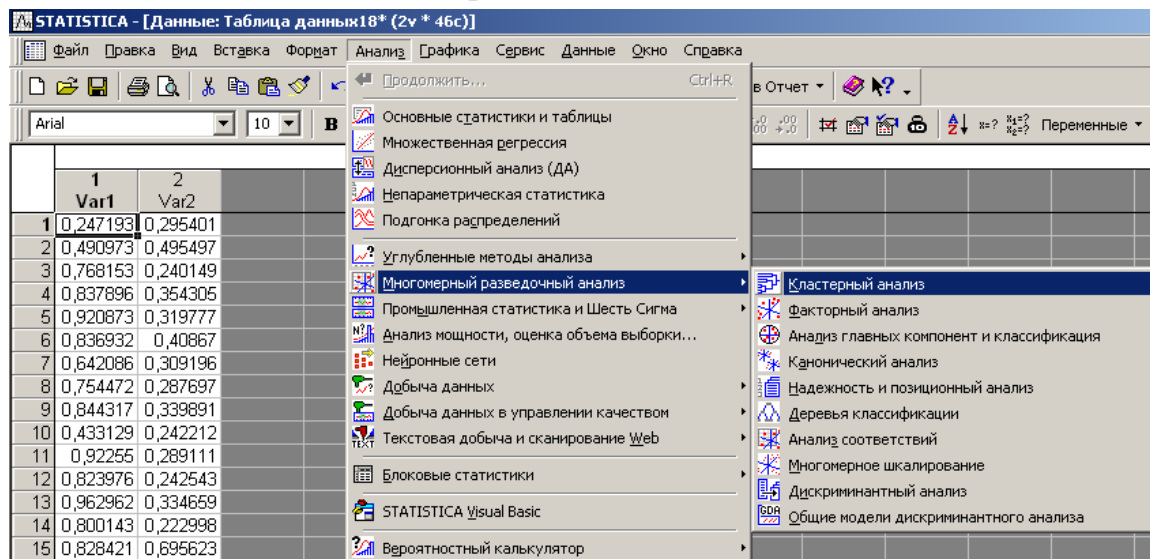


2. Скопируйте показатели в окно программы:

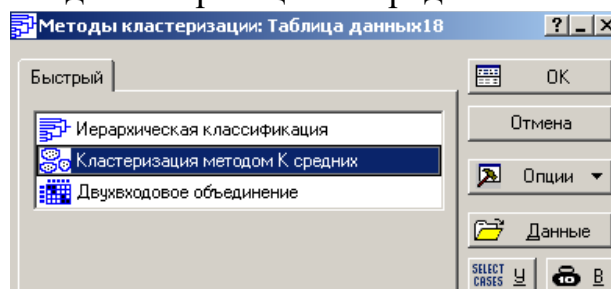


	1 Var1	2 Var2
1	0,247193	0,295401
2	0,490973	0,495497
3	0,768153	0,240149
4	0,837896	0,354305
5	0,920873	0,319777
6	0,836932	0,40867
7	0,642086	0,309196
8	0,754472	0,287697
9	0,844317	0,339891
10	0,433129	0,242212
11	0,92255	0,289111
12	0,823976	0,242543
13	0,962962	0,334659
14	0,800143	0,222998
15	0,828421	0,695623
16	1	0,365522
17	0,715453	0,354404
18	0,885349	0,572965
19	0,658321	0,277729
20	0,73897	0,632279
21	0,683659	0,357111
22	0,683659	0,591686
23	0,511567	0,428064
24	0,763287	0,777168
25	0,451862	0,339317
26	0,467823	0,428489
27	0,849075	0,298007
28	0,760966	0,25152
29	0,298413	0,201345
30	0,820473	0,438165
31	0,886396	0,251516

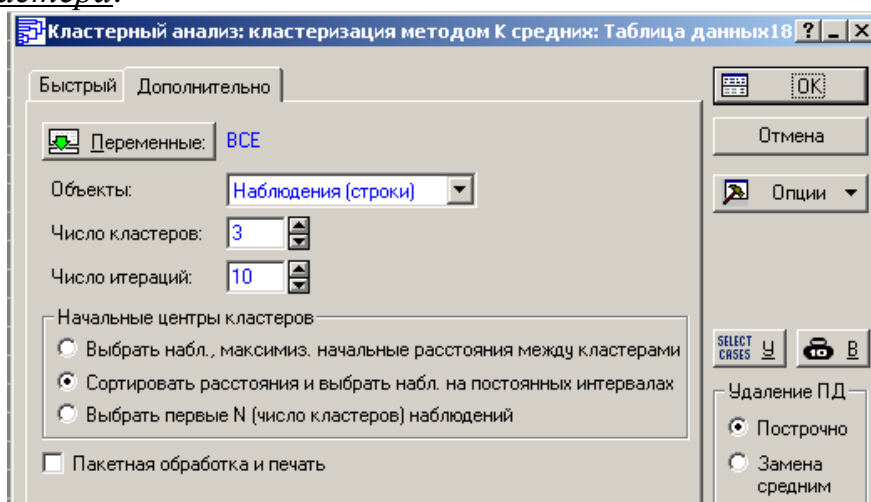
3. Запустите процедуру кластерного анализа: Анализ — Многомерный разведочный анализ — Кластерный анализ.



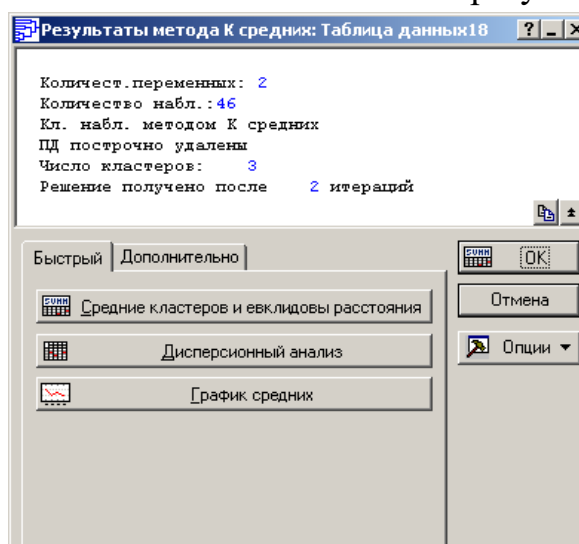
4. Выберите метод кластеризации К-средних.



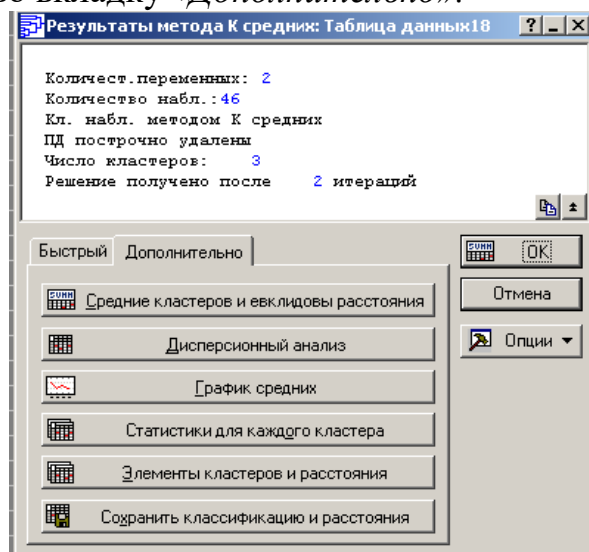
5. В появившемся окне укажите значения *переменных* (выберите все), *объекты установите* — *наблюдения (строки)*. Попытаемся разбить объекты на 3 кластера.



6. После нажатия кнопки «ОК» появятся результаты обработки:



7. Перейдите во вкладку «Дополнительно».



8. Нажмите кнопку «Элементы кластеров и расстояния». В результате вы получите 3 таблицы, показывающие, какие объекты относятся к одному из **трех кластеров**.

Элементы кластера номер 1 (Таблица данных18) и расстояния до центра кластера. Кластер содержит 9 набл.									
Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.
С 1	С 2	С 10	С 23	С 25	С 26	С 29	С 35	С 38	
Расст.	0,126932	0,108981	0,085823	0,083375	0,031304	0,060520	0,139921	0,072219	0,021170

Первый кластер

Элементы кластера номер 2 (Таблица данных18) и расстояния до центра кластера. Кластер содержит 27 набл.													
Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.
С 3	С 4	С 5	С 6	С 7	С 8	С 9	С 11	С 12	С 13	С 14	С 16	С 17	С 18
Расст.	0,057823	0,036787	0,077444	0,072492	0,120046	0,043348	0,031556	0,079559	0,047952	0,108346	0,061560	0,138857	0,072219

Второй кластер

Элементы кластера номер 3 (Таблица данных18) и расстояния до центра кластера. Кластер содержит 10 набл.									
Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.	Набл.Но.
С 15	С 18	С 20	С 22	С 24	С 32	С 34	С 37	С 41	С 43
Расст.	0,055052	0,100852	0,022462	0,070963	0,089924	0,082394	0,263728	0,145218	0,087989

Третий кластер

В строках таблиц указано расстояние от каждого населенного пункта до центра кластера.

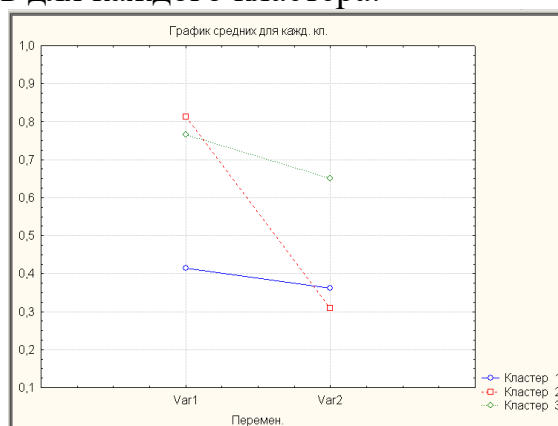
9. Вернитесь в окно анализа и нажмите кнопку «Средние кластеры и евклидовы расстояния».

Результат обработки появится на экране.

Евклидовы расст. между кластерами (Таблица данных18) Расстояния под диагональю Квадраты расстояний над диагональю									
Кластер Номер	Но. 1	Но. 2	Но. 3						
Но. 1	0,000000	0,080589	0,103222						
Но. 2	0,283882	0,000000	0,059137						
Но. 3	0,321282	0,243181	0,000000						

Над диагональю в таблице даны квадраты расстояний между кластерами.

10. Вернитесь в окно анализа и нажмите кнопку «График средних». В результате строятся следующие графики средних значений характеристик населенных пунктов для каждого кластера.

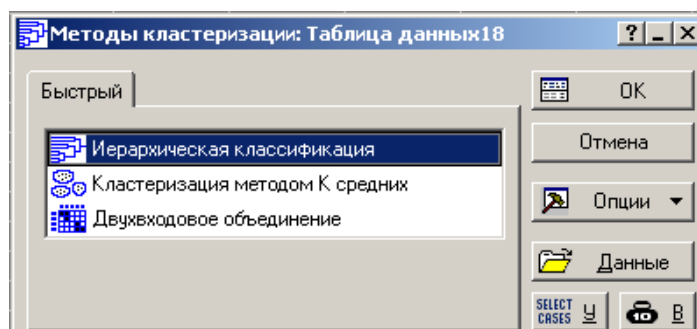


11. По результатам проведения анализа (пункт 8) создайте таблицу.

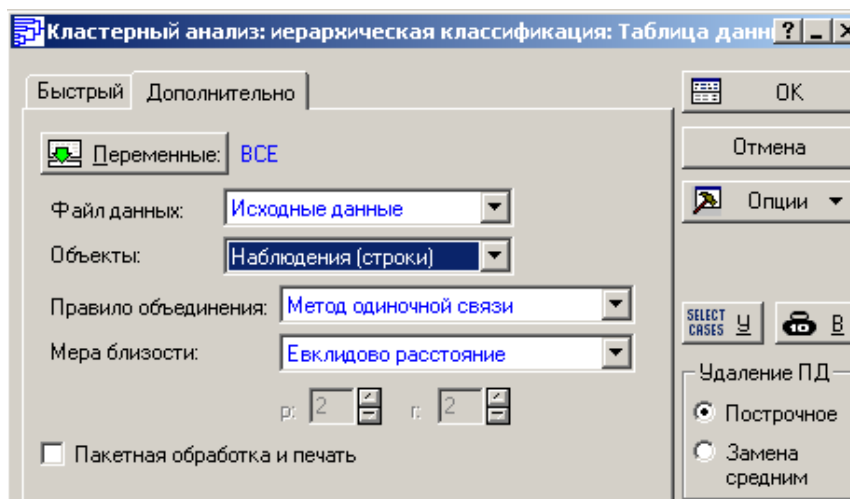
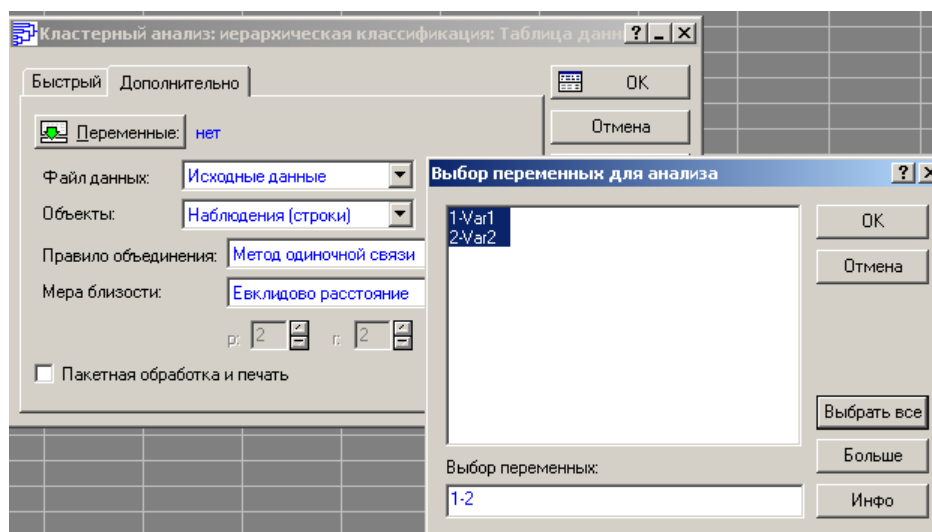
1 кластер			2 кластер			3 кластер		
№ НП	Соц-экон.	Радио-лог.	№ НП	Соц-экон.	Радио-лог.	№ НП	Соц-экон.	Радио-лог.
1	0,25	0,30	3	0,77	0,24	15	0,83	0,70
2	0,49	0,50	4	0,84	0,35	18	0,89	0,57
10	0,43	0,24	5	0,92	0,32	20	0,74	0,63
23	0,51	0,43	6	0,84	0,41	22	0,68	0,59
25	0,45	0,34	7	0,64	0,31	24	0,76	0,78
26	0,47	0,43	8	0,75	0,29	32	0,76	0,53
29	0,30	0,20	9	0,84	0,34	34	0,89	1,00
35	0,38	0,46	11	0,92	0,29	37	0,61	0,51
38	0,44	0,37	12	0,82	0,24	41	0,75	0,53
			13	0,96	0,33	43	0,74	0,66
			14	0,80	0,22			
			16	1,00	0,37			
			17	0,72	0,35			
			19	0,66	0,28			
			21	0,68	0,36			
			27	0,85	0,30			
			28	0,76	0,25			
			30	0,82	0,44			
			31	0,89	0,25			
			33	0,83	0,36			
			36	0,75	0,30			
			39	0,85	0,23			
			40	0,79	0,40			
			42	0,89	0,37			
			44	0,61	0,24			
			45	0,77	0,23			
			46	0,93	0,27			

✓ Сколько кластеров вы получили?
✓ Можно ли было сделать больше кластеров или меньшее их количество?

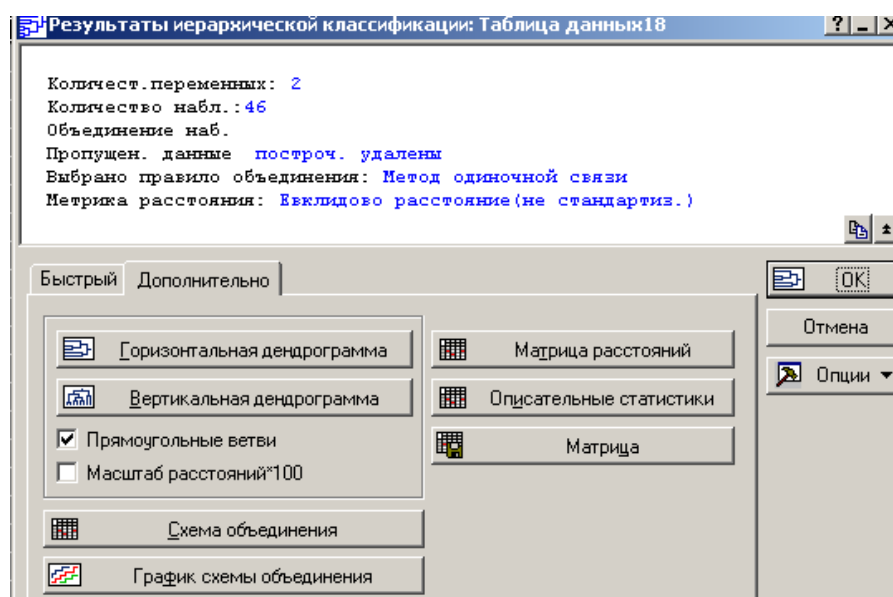
12. Построение дендрограммы. Вернитесь в окно анализа и закройте его. В появившемся окне выберите «Иерархическая классификация».



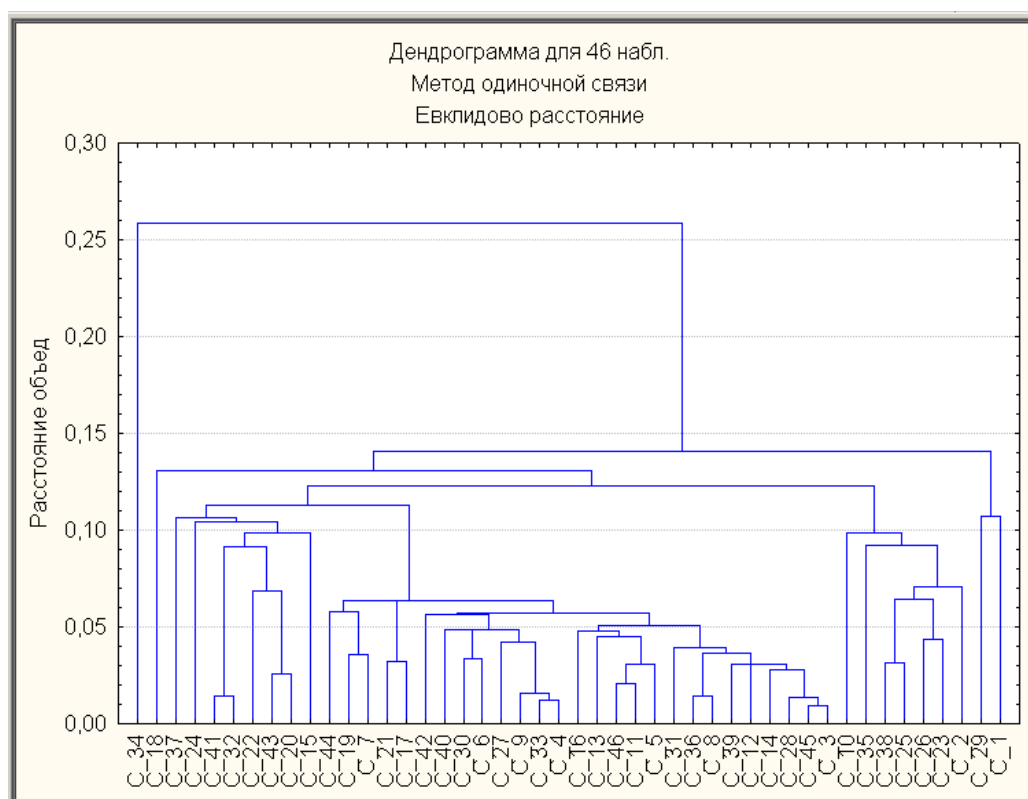
13. Задайте установки как показано на следующих рисунках:



14. После нажатия кнопки «ОК» появится окно результатов вычисления. В котором необходимо нажать «Вертикальная дендрограмма».



Результат отобразится на экране:



✓ Дискриминантный анализ. Задача

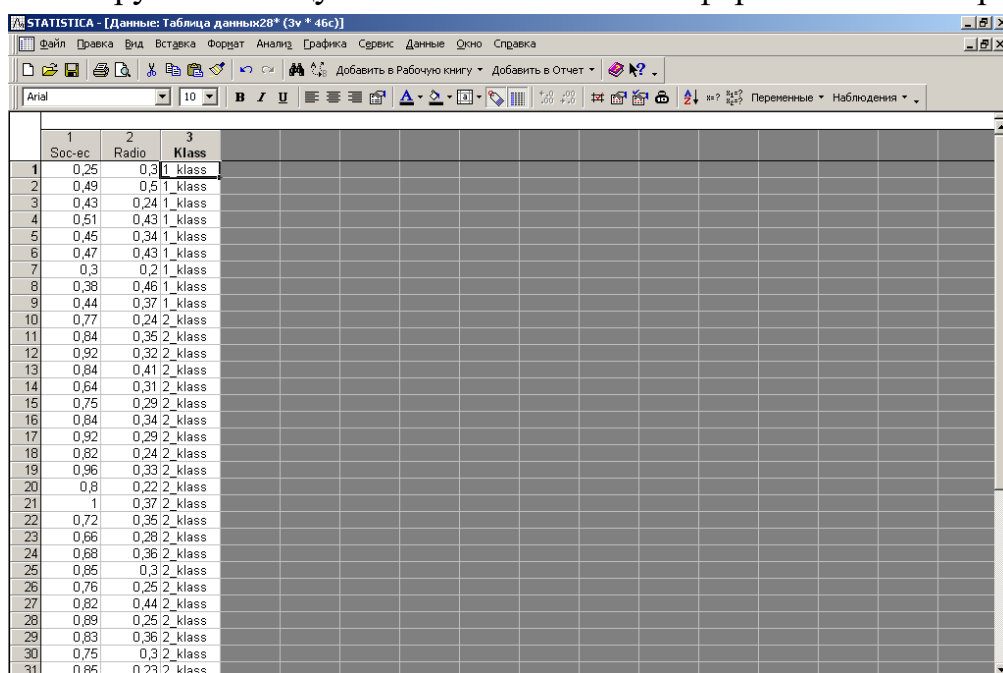
В предыдущем анализе мы получили таблицу, в которой показано то как объекты распределены по трем классам в зависимости от двух параметров. Преобразовав эту таблицу, выполним дискриминантный анализ данных и проверим, к какому классу можно отнести новый населенный пункт (социально-экономические показатели — 0,78; радиологические — 0,61).

№ НП	Соц-экон.	Радиолог.	Класс
1	0,25	0,3	1_klass
2	0,49	0,5	1_klass
10	0,43	0,24	1_klass
23	0,51	0,43	1_klass
25	0,45	0,34	1_klass
26	0,47	0,43	1_klass
29	0,3	0,2	1_klass
35	0,38	0,46	1_klass
38	0,44	0,37	1_klass
3	0,77	0,24	2_klass
4	0,84	0,35	2_klass
5	0,92	0,32	2_klass
6	0,84	0,41	2_klass
7	0,64	0,31	2_klass
8	0,75	0,29	2_klass
9	0,84	0,34	2_klass

№ НП	Соц-экон.	Радиолог.	Класс
11	0,92	0,29	2_klass
12	0,82	0,24	2_klass
13	0,96	0,33	2_klass
14	0,8	0,22	2_klass
16	1	0,37	2_klass
17	0,72	0,35	2_klass
19	0,66	0,28	2_klass
21	0,68	0,36	2_klass
27	0,85	0,3	2_klass
28	0,76	0,25	2_klass
30	0,82	0,44	2_klass
31	0,89	0,25	2_klass
33	0,83	0,36	2_klass
36	0,75	0,3	2_klass
39	0,85	0,23	2_klass
40	0,79	0,4	2_klass
42	0,89	0,37	2_klass
44	0,61	0,24	2_klass
45	0,77	0,23	2_klass
46	0,93	0,27	2_klass
15	0,83	0,7	3_klass
18	0,89	0,57	3_klass
20	0,74	0,63	3_klass
22	0,68	0,59	3_klass
24	0,76	0,78	3_klass
32	0,76	0,53	3_klass
34	0,89	1	3_klass
37	0,61	0,51	3_klass
41	0,75	0,53	3_klass
43	0,74	0,66	3_klass

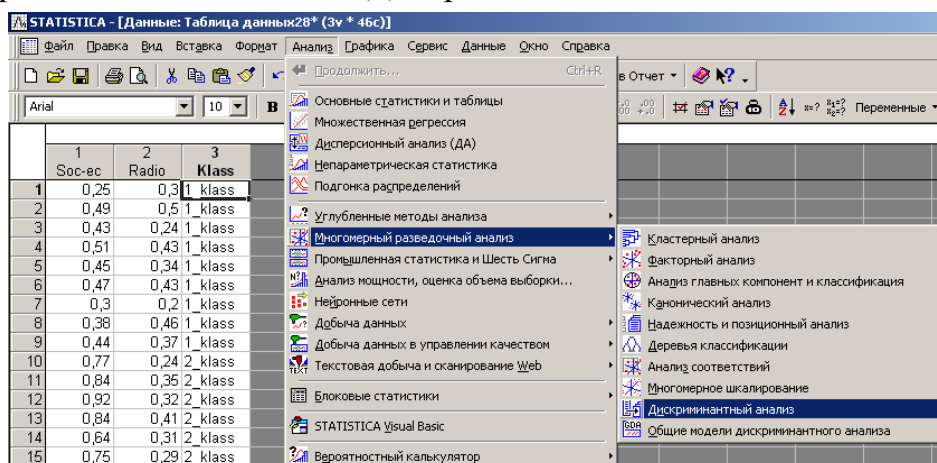
1. Задайте исходные настройки нового файла программы.

2. Скопируйте таблицу в появившееся поле и оформите согласно рисунку.

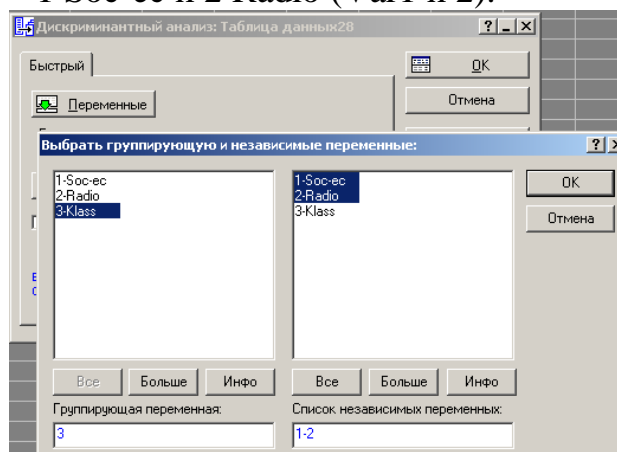


	1 Soc-ec	2 Radio	3 Klass
1	0,25	0,3	1_klass
2	0,49	0,5	1_klass
3	0,43	0,24	1_klass
4	0,51	0,43	1_klass
5	0,45	0,34	1_klass
6	0,47	0,43	1_klass
7	0,3	0,2	1_klass
8	0,38	0,46	1_klass
9	0,44	0,37	1_klass
10	0,77	0,24	2_klass
11	0,84	0,35	2_klass
12	0,92	0,32	2_klass
13	0,84	0,41	2_klass
14	0,64	0,31	2_klass
15	0,75	0,29	2_klass
16	0,84	0,34	2_klass
17	0,92	0,29	2_klass
18	0,82	0,24	2_klass
19	0,96	0,33	2_klass
20	0,8	0,22	2_klass
21	1	0,37	2_klass
22	0,72	0,35	2_klass
23	0,66	0,28	2_klass
24	0,68	0,36	2_klass
25	0,85	0,3	2_klass
26	0,76	0,25	2_klass
27	0,82	0,44	2_klass
28	0,89	0,25	2_klass
29	0,83	0,36	2_klass
30	0,75	0,3	2_klass
31	0,85	0,23	2_klass

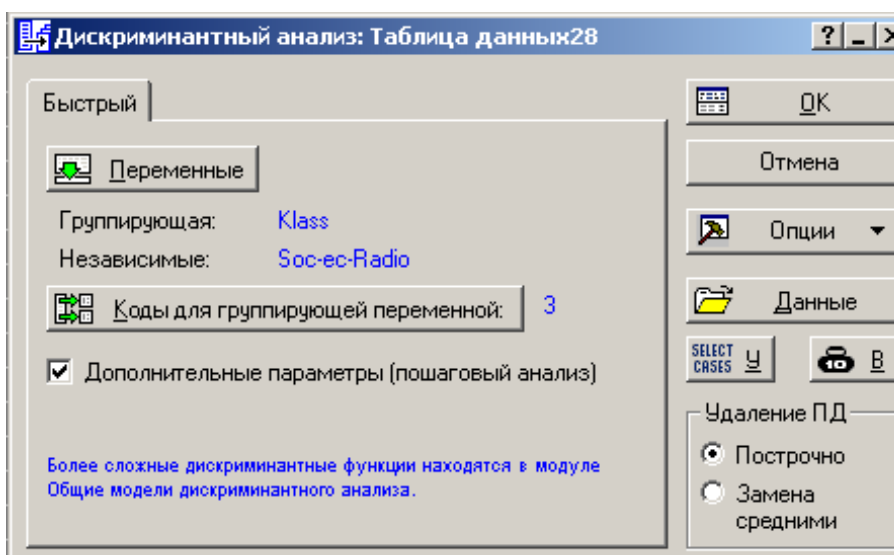
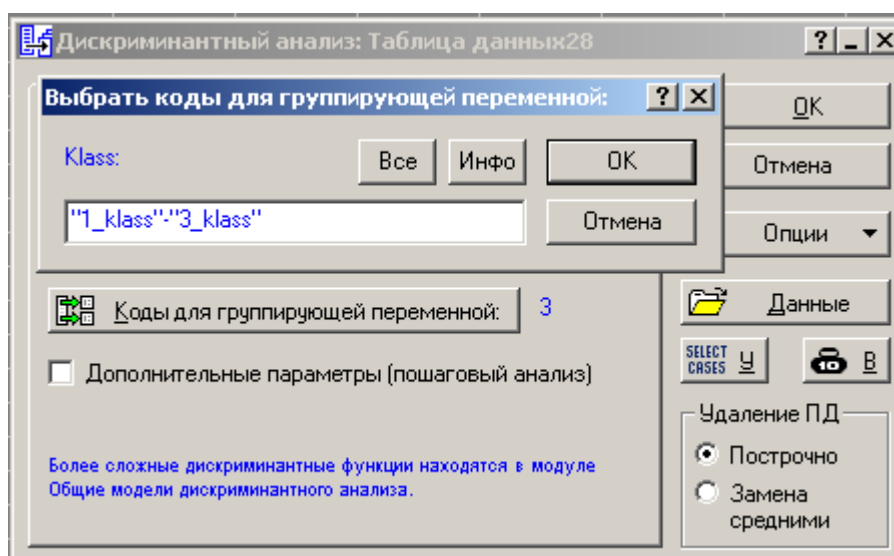
3. Запустите процедуру дискриминантного анализа: *Анализ — Многомерный разведочный анализ — Дискриминантный анализ.*



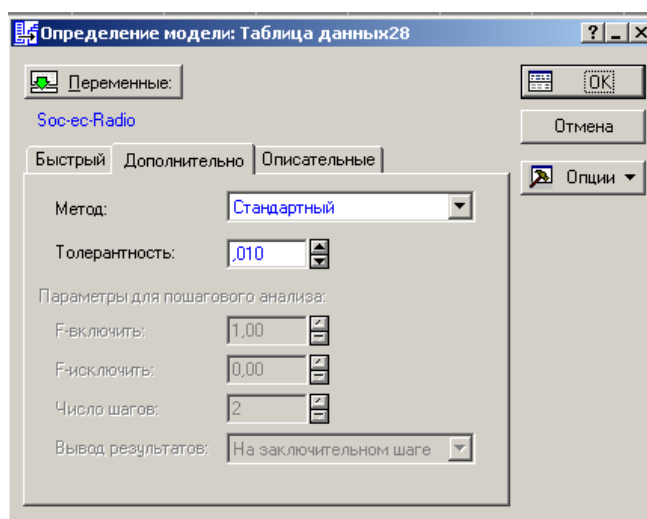
4. Укажите Группирующую переменную как 3-Klass (Var3), а независимые переменные — 1 Soc-ec и 2 Radio (Var1 и 2).



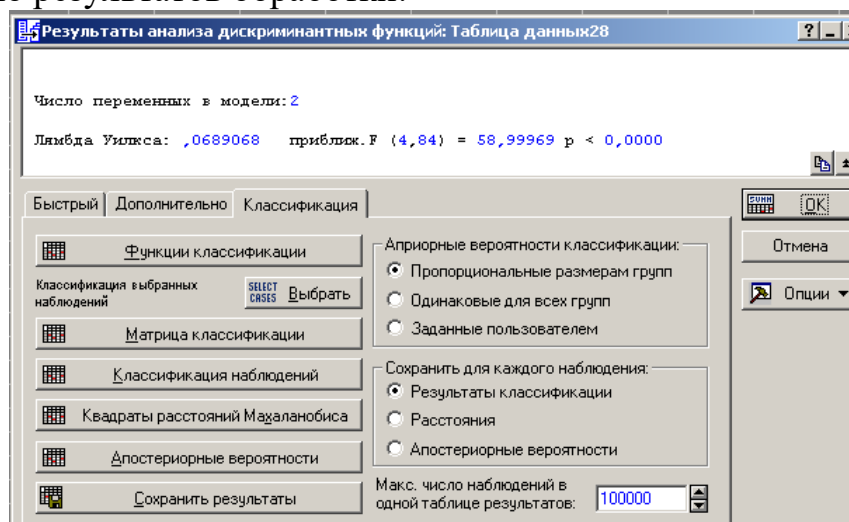
5. Нажмите на кнопку коды переменной, затем «ОК».



6. Нажмите «ОК» и в появившемся окне задайте установки как представлено на рисунке. Затем «ОК».



7. Окно результатов обработки:



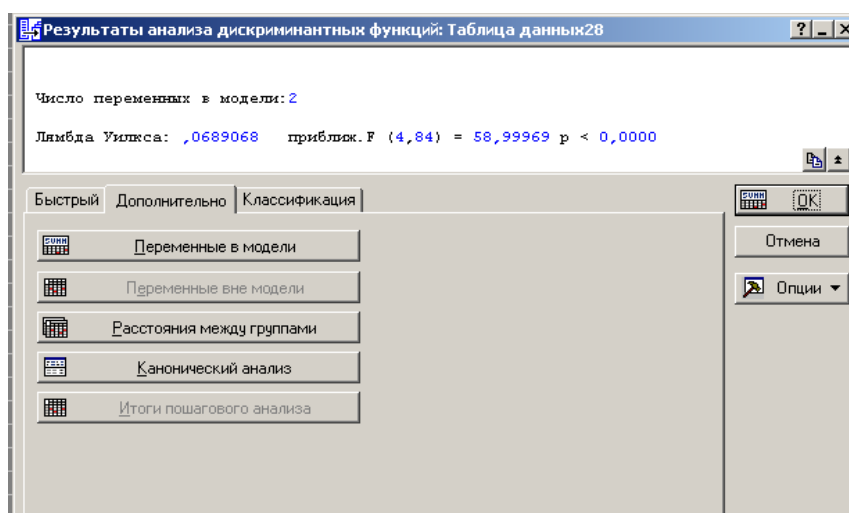
Информационная часть окна сообщает, что использовано:

- **Число переменных в модели: 2;**
- **Лямбда Уилкса: 0,0689068;**
- **приблизж. $F(4,84) = 58,999$** (Приближенное значение F -статистики), связанной с лямбдой Уилкса;
- **p — уровень значимости F -критерия для значения 58,999;**
- **значения статистики лямбда Уилкса** лежат в интервале 0–1.

Значения **статистики Уилкса**, лежащие около нуля, свидетельствуют о хорошей дискриминации. Значения **статистики Уилкса**, лежащие около единицы, свидетельствуют о плохой дискриминации.

Иными словами, это можно выразить следующим образом: если значения **лямбды Уилкса** близки к нулю, то мощность дискриминации (мощность = 1 — вероятность ошибки) близка к 1, если **лямбда Уилкса** близка к единице, то мощность близка к нулю.

8. Перейдите во вкладку *Дополнительно* и нажмите кнопку *Переменные в модели*.



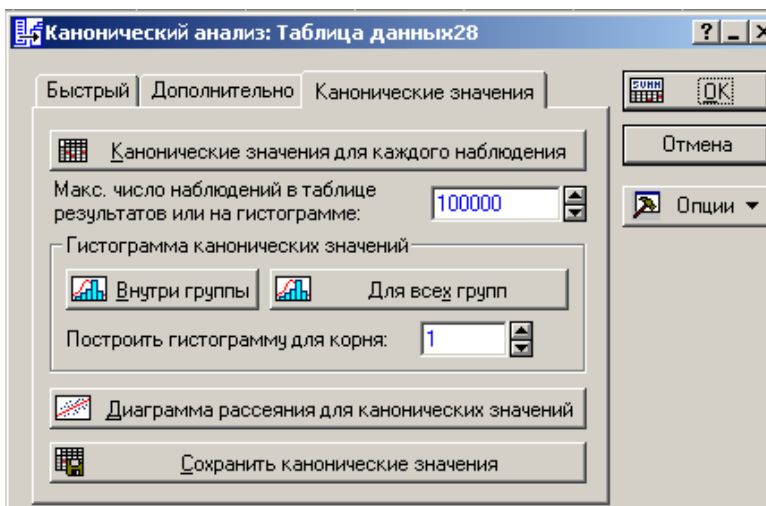
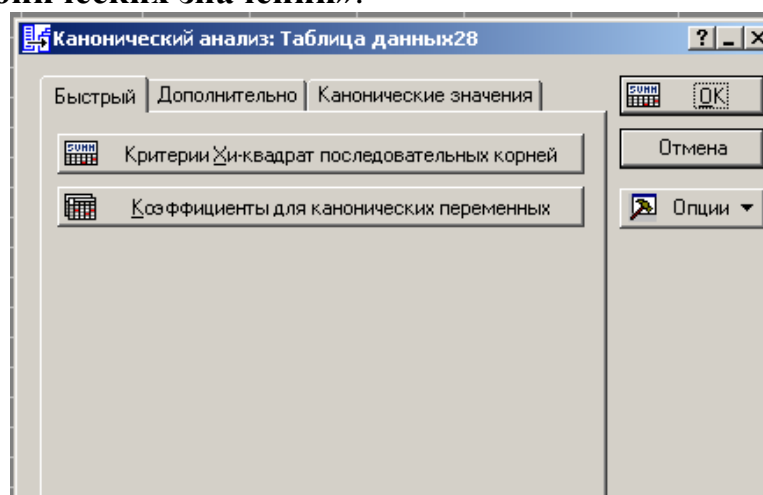
На экране появится итоговая таблица анализа.

Итоги анализа дискриминантн. функций (Таблица данных28) Переменных в модели: 2; Группир.: Klass (3 гр.) Лямбда Уилкса: ,06891 при бл. F (4,84)=59,000 p<0,0000						
N=46	Уилкса лямбда	Частная лямбда	F-исключ (2,42)	p-уров.	Толер.	1-толер. (R-кв.)
Soc-ec	0,307864	0,223822	72,82468	0,000000	0,864901	0,135099
Radio	0,256387	0,268761	57,13625	0,000000	0,864901	0,135099

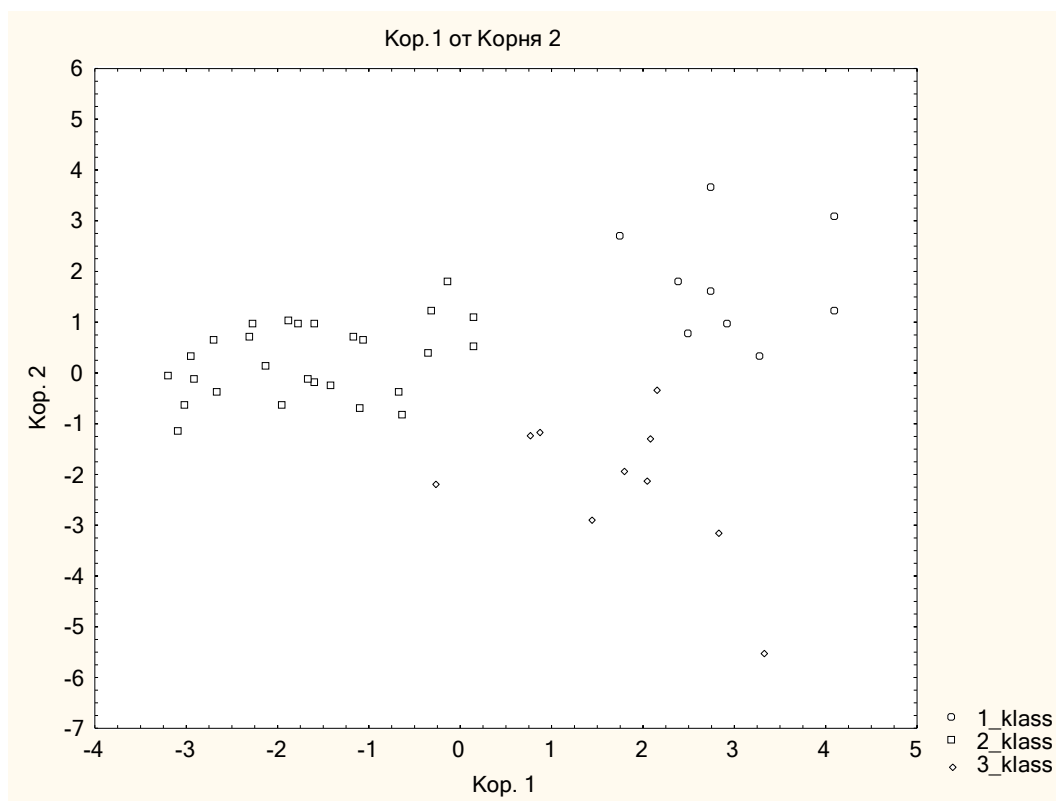
Итоги анализа дискриминантн. функций (Таблица данных28) Переменных в модели: 2; Группир.: Klass (3 гр.) Лямбда Уилкса: ,06891 при бл. F (4,84)=59,000 p<0,0000						
	Уилкса	Частная	F-исключ	p-уров.	Толер.	1-толер.
Soc-ec	0,307864	0,223822	72,82468	0,000000	0,864901	0,135099
Radio	0,256387	0,268761	57,13625	0,000000	0,864901	0,135099

Результаты свидетельствуют о хорошей дискриминации.

9. Просмотрите разделение групп на графике. Для этого иницилируйте кнопку «**Канонический анализ**». В появившемся диалоговом окне перейдите во вкладку *Канонические значения* и выберите «**Диаграмма рассеяния для канонических значений**».



На экране появится график:



Из рисунка видно, что населенные пункты четко разделены на три класса:

Населенные пункты, объединенные в первый кластер, имеют крайне низкий демографический потенциал, а значит и социально-экономический фактор, т.е. «умирающие» населенные пункты. В данных населенных пунктах нецелесообразно развивать социально-экономическую структуру.

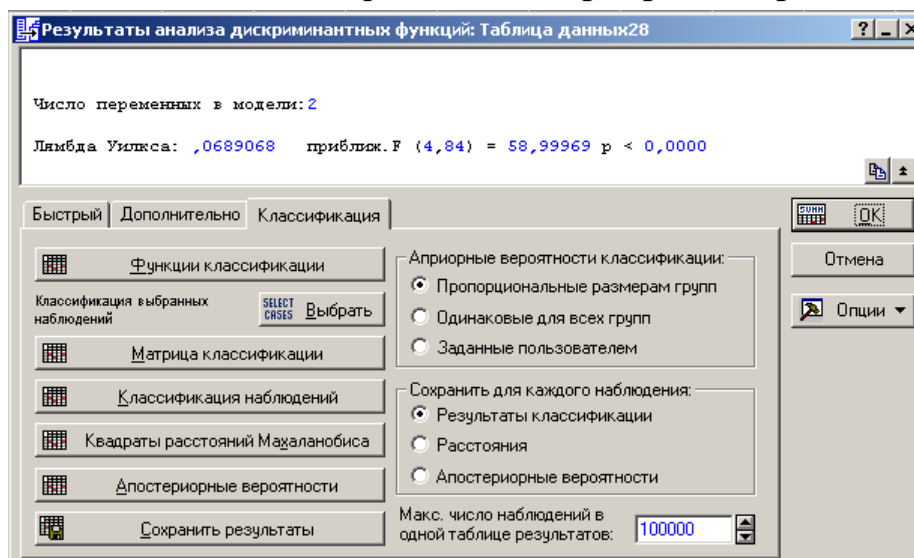
Населенные пункты, находящиеся во втором кластере приоритетны в распределении инвестиций, направленных на их социально-экономическое развитие. При этом наиболее перспективными являются населенные пункты, расположенные в правой части этого кластера. Эти пункты имеют стабильный социально-экономический фактор, характеризуются стабильной радиационной обстановкой и в них нецелесообразно проводить защитные мероприятия.

Населенные пункты третьего кластера характеризуются повышенными значениями активности ^{137}Cs в молоке и суммарной годовой дозой. Поэтому в этих населенных пунктах необходимо рекомендовать проведение контролер. При этом мероприятия должны быть первоочередными в населенных пунктах, расположенных в правой и верхней части первого кластера.

10. Апостериорные вероятности. Нажав кнопку «**Апостериорные вероятности**», вы увидите таблицу с апостериорными вероятностями принадлежности объекта к определенному классу.

12. Для этого не закрывая предыдущий анализ запустите новый, процедура та же, что и описанная выше.

13. В появившемся окне выберите «Апостериорные вероятности».



14. По результатам вычисления видно, что новый населенный пункт № 47 с вероятностью 99 % относится к 3 классу.

Апостериорные вероятности (Таблица данных28)					
Неправильные классификации отмечены *					
Наблюдение	Наблюд. Класс.	1_klass p=,19565	2_klass p=,58696	3_klass p=,21739	
34	2_klass	0,029322	0,970394	0,000285	
35	2_klass	0,000003	0,999992	0,000005	
36	2_klass	0,000000	0,999999	0,000001	
37	3_klass	0,000007	0,000240	0,999754	
38	3_klass	0,000010	0,283444	0,716546	
39	3_klass	0,000529	0,000821	0,998650	
40	3_klass	0,008406	0,001357	0,990237	
41	3_klass	0,000013	0,000001	0,999986	
42	3_klass	0,001848	0,119571	0,878582	
43	3_klass	0,000000	0,000000	1,000000	
44	3_klass	0,296116	0,008174	0,695710	
45	3_klass	0,002624	0,097728	0,899649	
46	3_klass	0,000288	0,000203	0,999509	
47	---	0,000214	0,005114	0,994673	

Контрольные вопросы

1. Что означает классифицировать объект?
2. Какие статистические методы классификации вам известны?
3. Для чего используется кластерный анализ?
4. Что такое кластер?
5. Результат кластерного анализа.
6. Для чего используется дискриминантный анализ?
7. Что такое дискриминирующая переменная?
8. Что такое обучающая выборка?
9. Цель дискриминантного анализа.

ЛИТЕРАТУРА

1. Гланц, С. Медико-биологическая статистика: пер. с англ. / Гланц С. — М., Практика, 1998. — 459 с.
2. Платонов, А. Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы / А. Е. Платонов. — М.: Издательство РАМН, 2000. — 52 с.
3. Жученко, Ю. М. Информационные технологии в биологии и химии: лабораторный практикум для студентов вузов по специальности 1-31 01 01 «Биология» / Ю. М. Жученко. М-во образования РБ, Гомельский гос. ун-т им. Ф. Скорины. — Гомель: ГГУ им. Ф. Скорины, 2010. — 148 с.
4. Реброва, О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA / О. Ю. Реброва. — 3-е изд. — М., МедиаСфера, 2006. — 312 с.
5. Халафян, А. А. STATISTICA 6. Статистический анализ данных / А. А. Халафян. — 3-е изд. — М.: ООО «Бином-Пресс», 2007. — 512 с.
6. Обработка экспериментальных данных в MS Excel: методические указания к выполнению лабораторных работ для студентов дневной формы обучения / сост. Е. Г. Агапова, Е. А. Битехтина. — Хабаровск: Изд-во Тихоокеан. гос. ун-та, 2012. — 32 с.
7. STATISTICA 6.0 — фирменное руководство. Компания StatSoft. Электронная публикация, 1995.
8. Максимов, С. И. Статистический анализ и обработка данных с применением MSExcel и SPSS: учеб.-метод. пособие / С. И. Максимов. — Минск: РИВШ, 2012. — 114 с.
9. Джелен, Б. Сводные таблицы в MSExcel 2010.: пер. с англ. / Б. Джелен, М. Александер. — М.: ООО «И. Д. Вильямс», 2011. — 464 с.
10. Лялин, В. С. Статистика: теория и практика в Excel: учеб. пособие / В. С. Лялин, И. Г. Зверева, Н. Г. Никифорова. — М.: Финансы и статистика; ИНФРА-М, 2010. — 448 с.
11. Mario, F. Triola. Elementary statistics / F. Triola Mario. — 10 th ed. — 770 с.
12. Боровиков, В. Statistica. Искусство анализа данных на компьютере: Для профессионалов / В. Боровиков. — 2-е изд. — Спб.: Питер, 2003. — 688 с.

