[Парная] корреляция и регрессия

Типы статистических задач

Задачи	Инструменты
Описание совокупностей объектов	Анализ одной выборки; расчет параметров распределений (положения, формы); проверка нормальности распределений;
Сравнение параметров	Парные и множественные сравнения средних; сравнение распределений; сравнение частот; t-критерий; тест Манна-Уитни или Краскела-Уоллеса; дисперсионный анализ;
Анализ зависимостей	**Установление взаимосвязи между двумя переменными или между многими переменными; установление силы влияния одной или многих переменных на одну результирующую; корреляционный анализ, парная и множественная регрессия, логит-регрессия;**
Снижение размерности,	Кластерный, факторный, дискриминантный анализ; анализ соответствий; многомерное шкалирование и

Выбор статистического теста при сравнении распределений (сравнении центральных тенденций и частот)

Задача	Количественная шкала, нормальное распределение	Порядковая шкала или отклонение от нормального распределения	Номинальная шкала
Сравнить одну группу с гипотетическим значением	t-тест Стьюдента для одной выборки	Тест Вилкоксона	Тест хи-квадрат
Сравнить две не связанные совокупности	t-тест Стьюдента для не связанных совокупностей	Тест Манна-Уитни	Тест Фишера (тест хи-квадрат)
Сравнить две связанные совокупности	t-тест Стьюдента для связанных совокупностей	Тест Вилкоксона	Тест Мак-Неймера
Сравнить более двух не связанных совокупностей	Однофакторный дисперсионный анализ	Тест Краскела-Уоллиса	Тест хи-квадрат
Сравнить более двух связанных совокупностей	Дисперсионный анализ с повторными	Тест Фридмана	Тест Кохрана

Задачи оценки взаимосвязи между переменными или прогноза

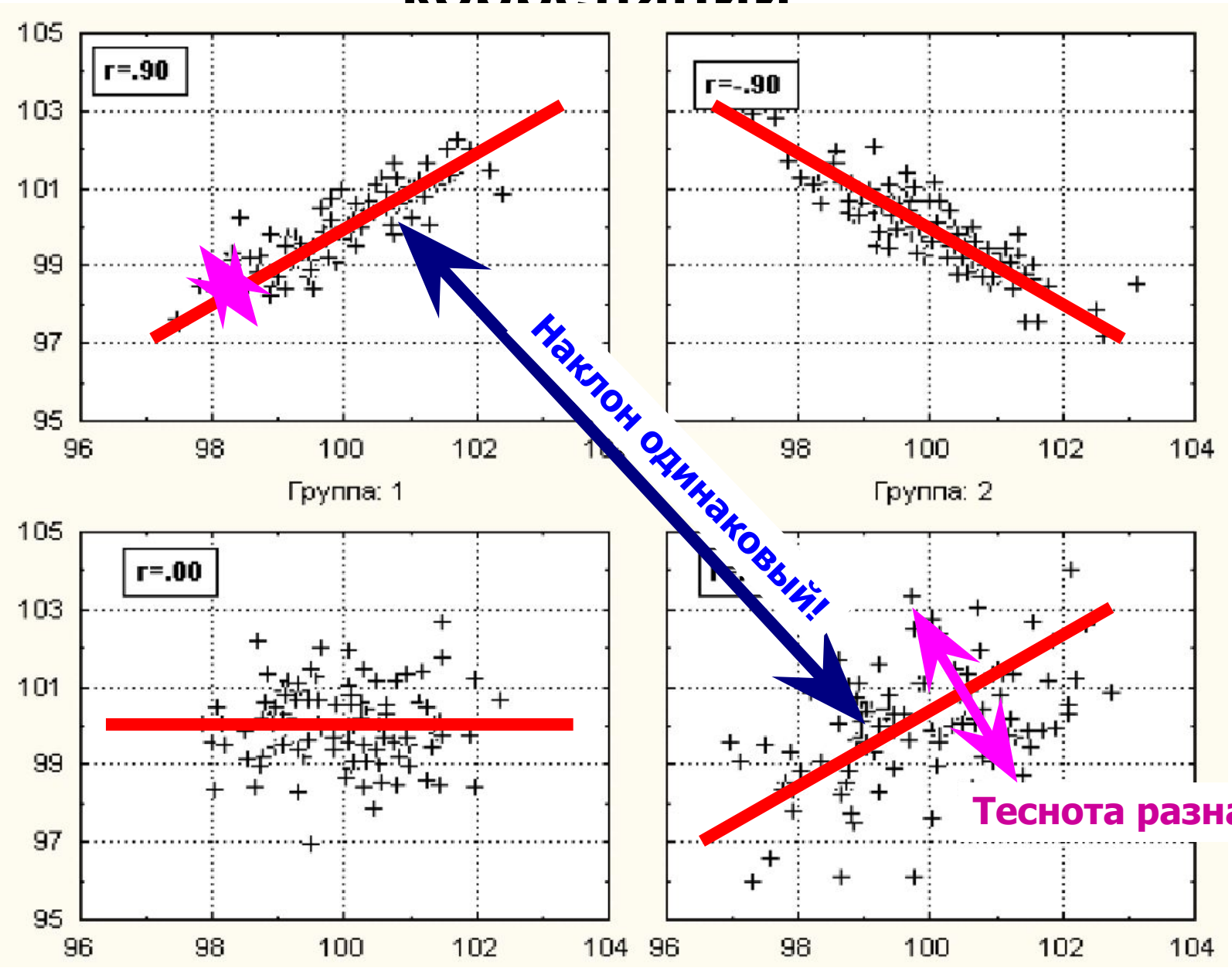
Задача	Количественные нормально распределенные переменные	Количественные ненормально распределенные переменные или ранги	Биноминальные данные (два возможных результата)
Оценить взаимосвязь между двумя	Коэффициент парной корреляции Пирсона	Коэффициенты ранговых корреляций (Спирмена,	Коэффициенты связи
Предсказать изменение одной переменной, если была измерена другая переменная	Простая линейная регрессия или нелинейная регрессия	Непараметрическая (ранговая) регрессия	Простая логистическая регрессия
Предсказать значение, базируясь на нескольких переменных	Множественная линейная (нелинейная) регрессия	Множественная линейная ранговая (медианная) регрессия	Множественная логистическая регрессия

Корреляция?

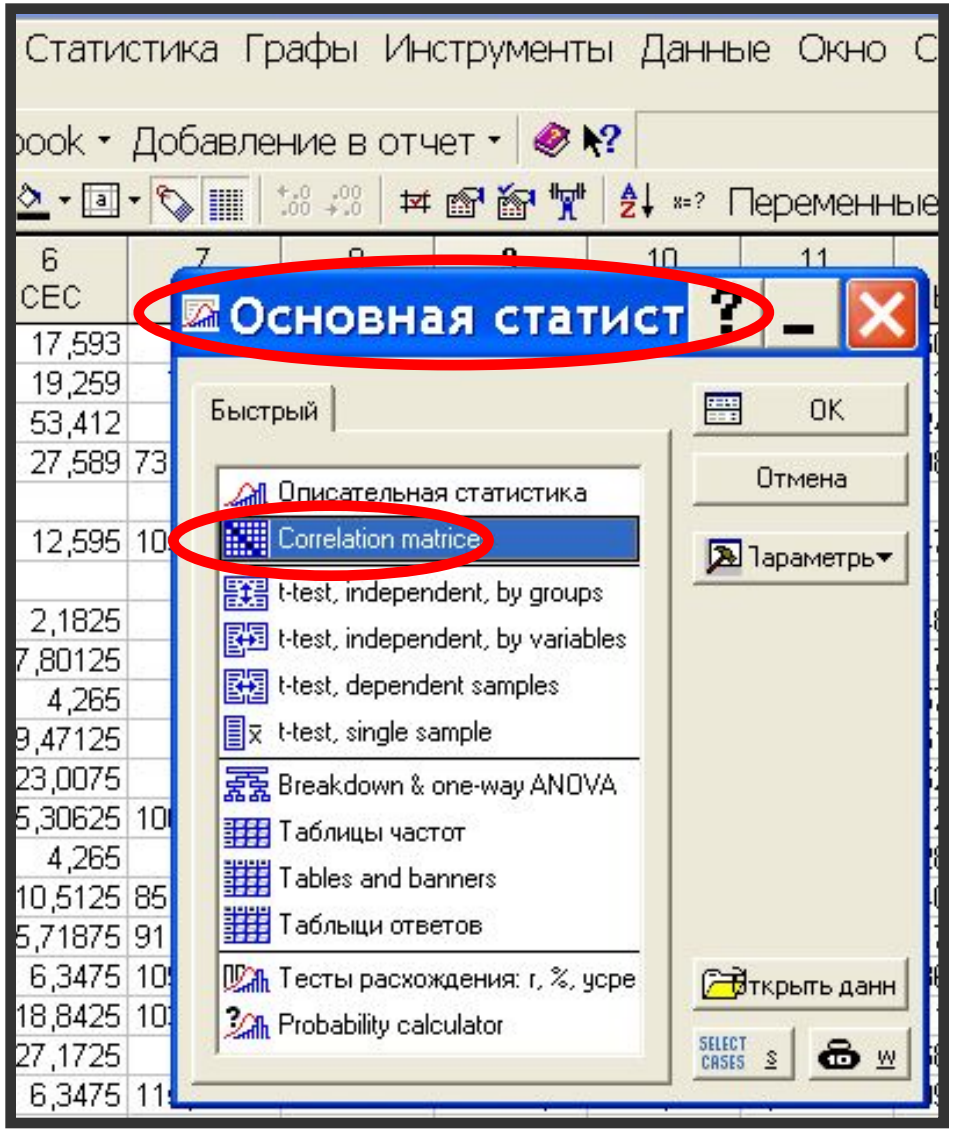
- Использование коэффициента корреляции позволяет оценить, в какой степени две переменные изменяются совместно – увеличивается ли или уменьшается одна переменная при изменении другой.
- Коэффициент корреляции – мера силы (тесноты и направления) связи между изменчивостью переменных.
- Интерпретация знака коэффициента корреляции – есть вопросы?
- Надежность коэффициент корреляции зависит от его величины и *n*.

- **Никаких причинных интерпретаций коэффициент корреляции сделать не позволяет!**
- **Коэффициент корреляции может быть использован только для прогноза направления (но не величины!) изменения одной переменной в связи с изменением другой**

Знаки и теснота коэффициента корреляции

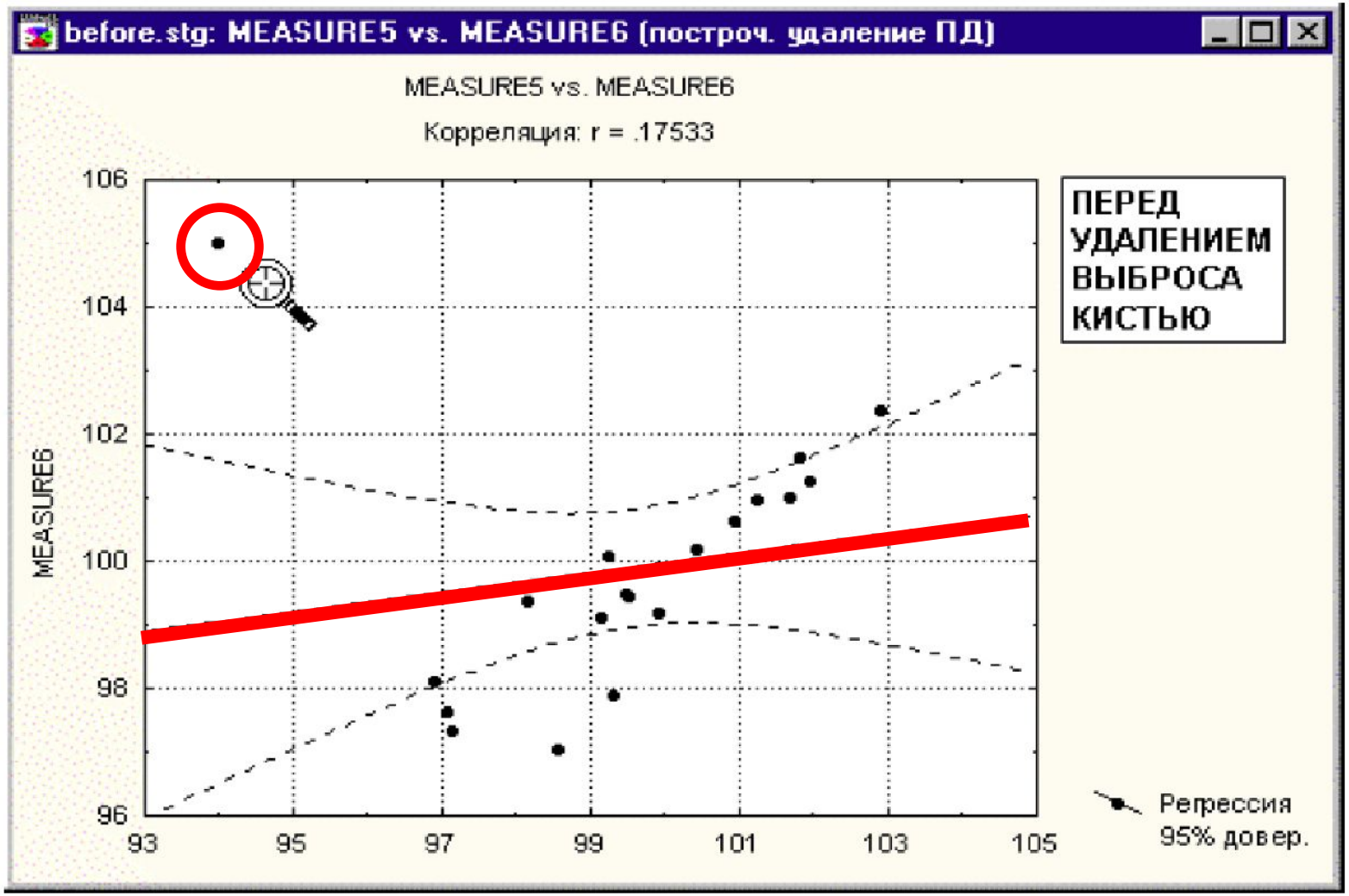


Техника расчета *r* Пирсона

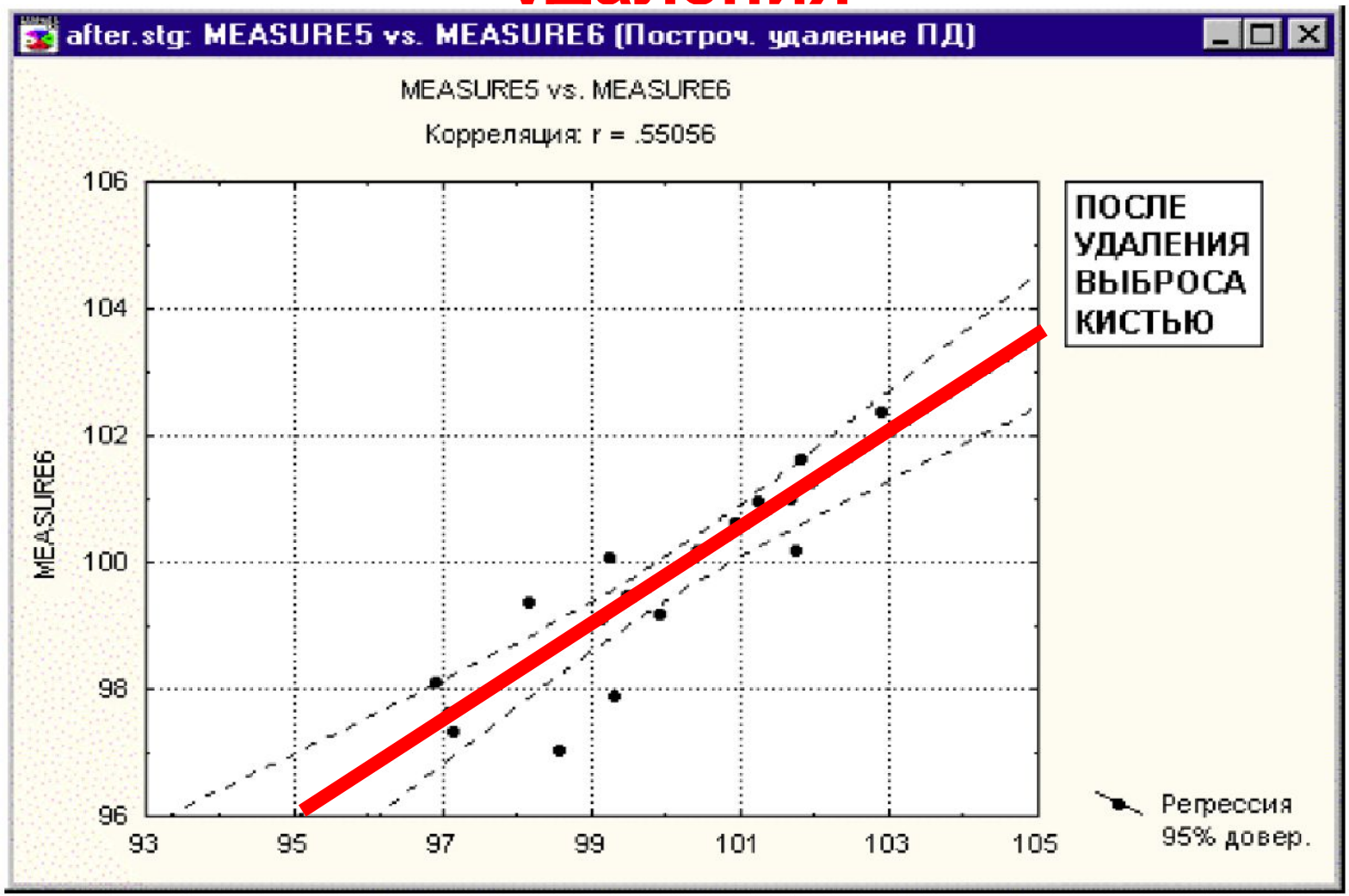


- **«Пример_тм_токсичность_преобразования.xls» (Cu_хлорид * Cd_хлорид вместе и по зонам)**
- **Пары переменных или матрицы;**
- **Просмотр результатов в разном расширении;**
- **Иллюстрации;**
- **Категоризированные зависимости**

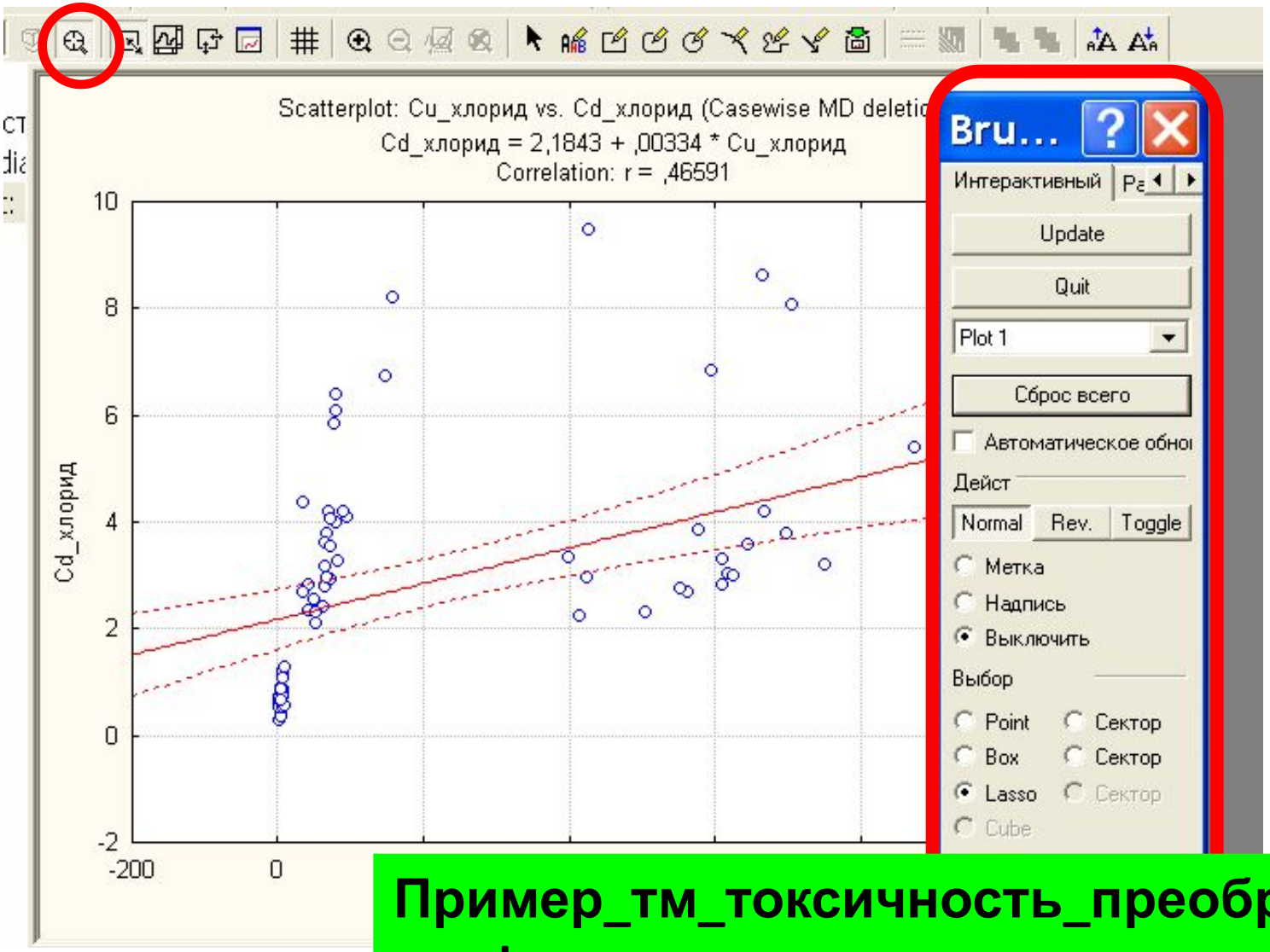
Нарушение «нормальности»: управление выбросами: перед удалением



Нарушение «нормальности»: управление выбросами: после удаления



Управление выбросами: инструмент «кисть»



Пример_тм_токсичность_преобразование.xls;

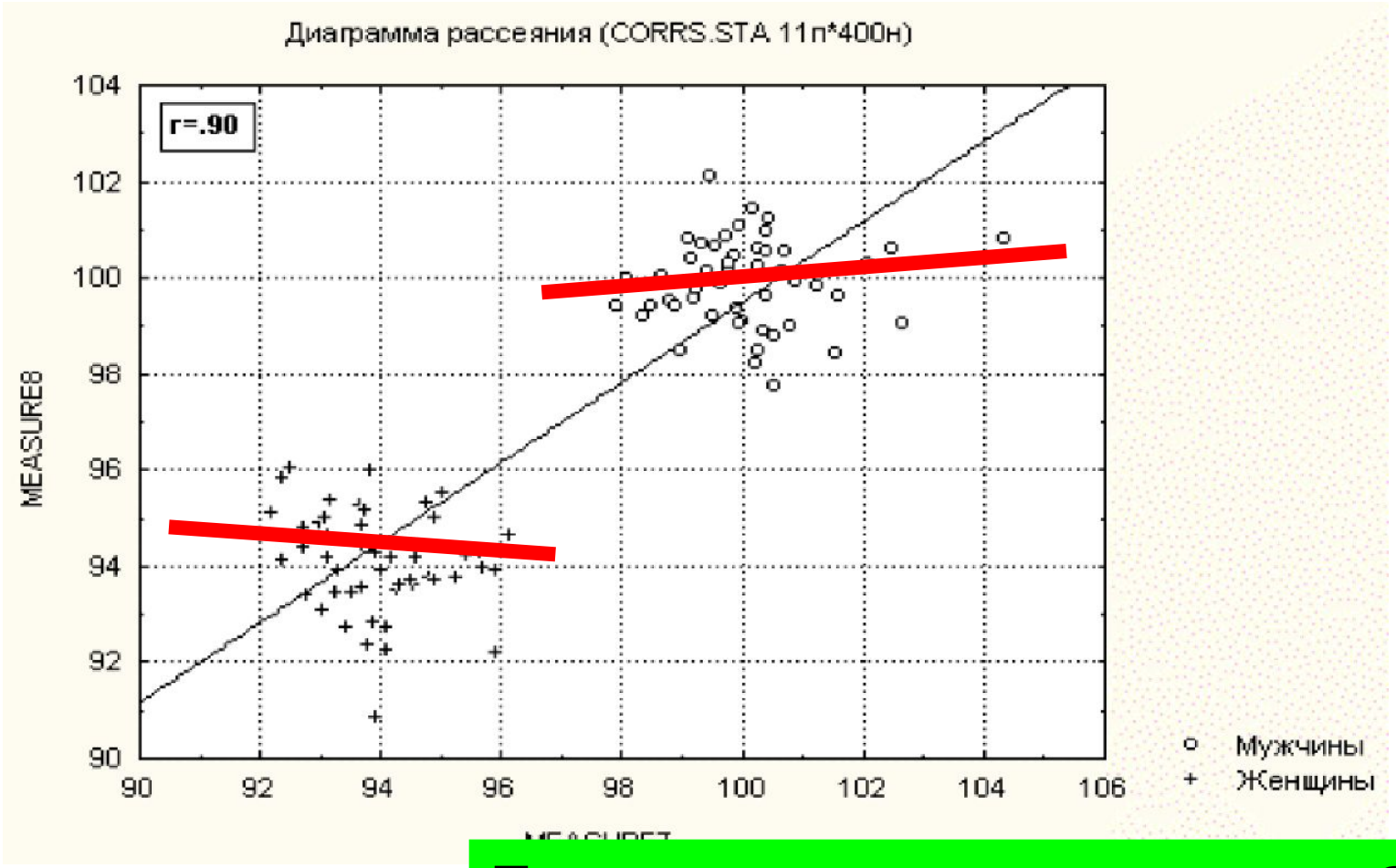
Cu_хлорид * Cd_хлорид вместе

Управление выбросами: общие правила отсутствуют

Количественный подход к выбросам. Некоторые исследователи применяют численные методы удаления выбросов. Например, исключаются значения, которые выходят за границы ±2 стандартных отклонений (и даже ±1.5 стандартных отклонений) вокруг выборочного среднего. В ряде случаев такая "чистка" данных абсолютно необходима. Например, при изучении реакции в когнитивной психологии, даже если почти все значения экспериментальных данных лежат в диапазоне 300-700 *миллисекунд*, то несколько "странных времен реакции" 10-15 *секунд* совершенно меняют общую картину.

К сожалению, определение выбросов субъективно, и решение должно приниматься индивидуально в каждом эксперименте (с учетом особенностей эксперимента и/или "сложившейся практики" в данной области). Следует заметить, что в некоторых случаях относительная частота выбросов к численности групп может быть исследована и разумно проинтерпретирована с точки зрения самой организации эксперимента.

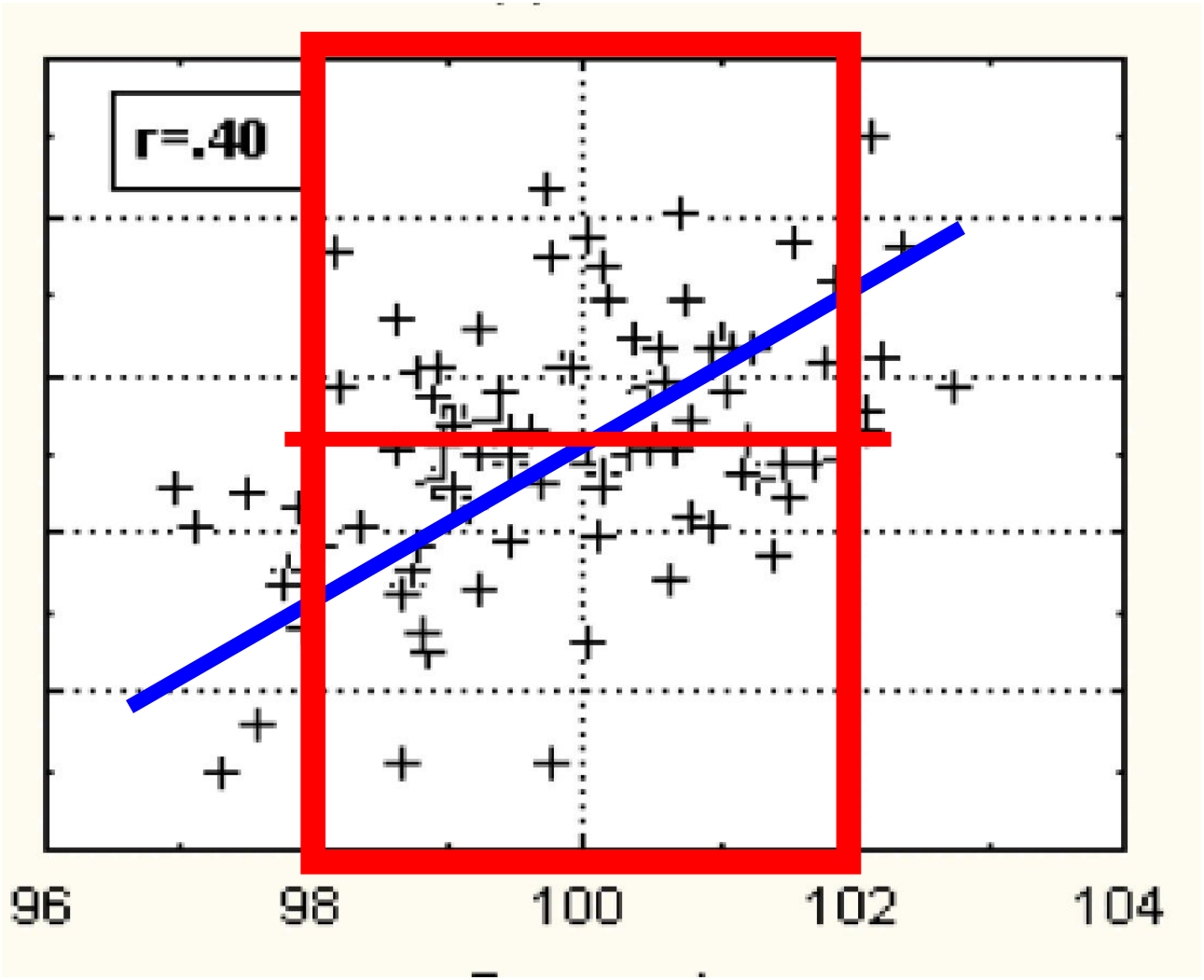
Осторожно:
корреляция в неоднородных группах !



Пример_тм_токсичность_преобразование.xls;

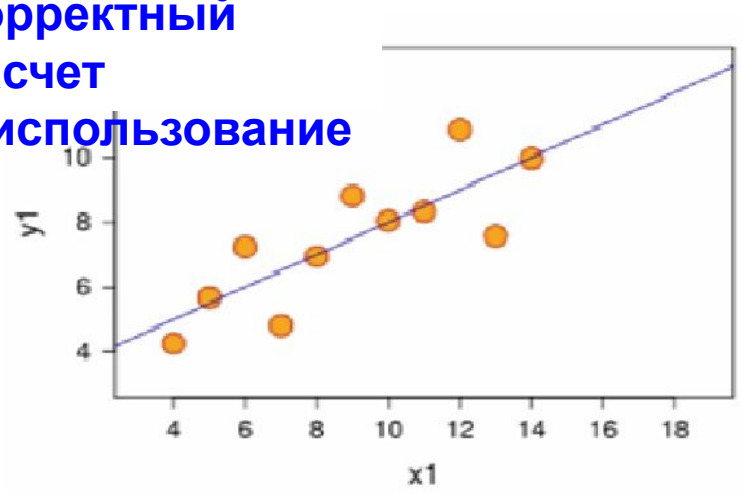
Cu_хлорид * Cd_хлорид вместе и по

Условие продуктивного использования коэффициентов корреляции: достаточная дисперсия данных

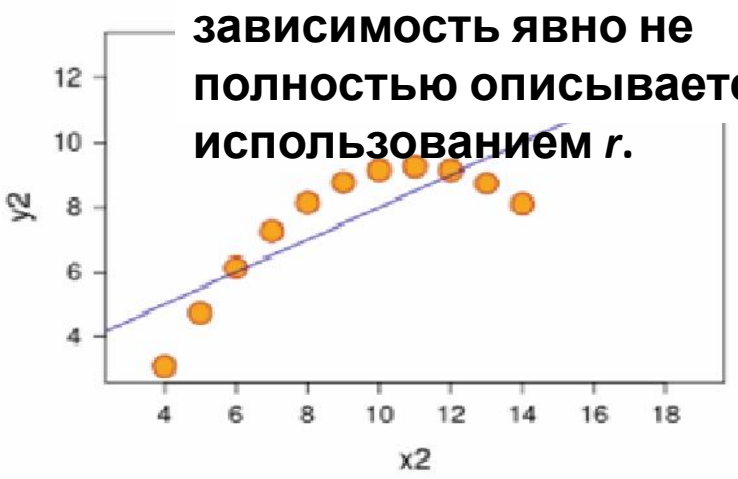


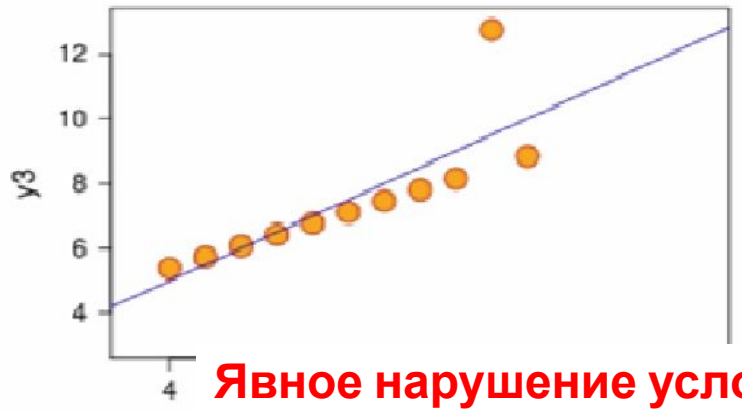
Правильный/неправильный расчет и интерпретация *r* (во всех случаях *r*=0,816 и *P* одинаковая)

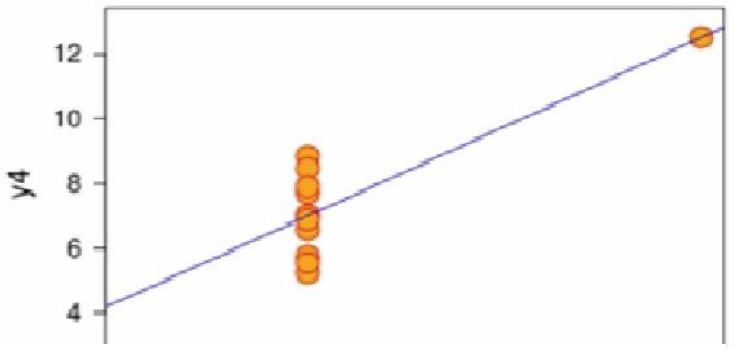
Корректный расчет и использование



Так делать можно, но зависимость явно не полностью описывается с использованием *r*.





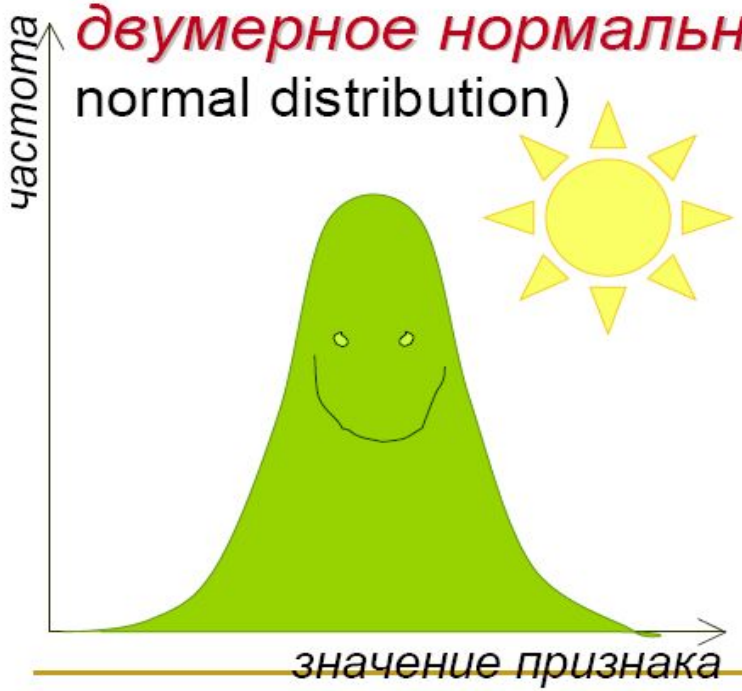


Явное нарушение условий использования *r*: «выбросы» и отклонение от нормального распределения

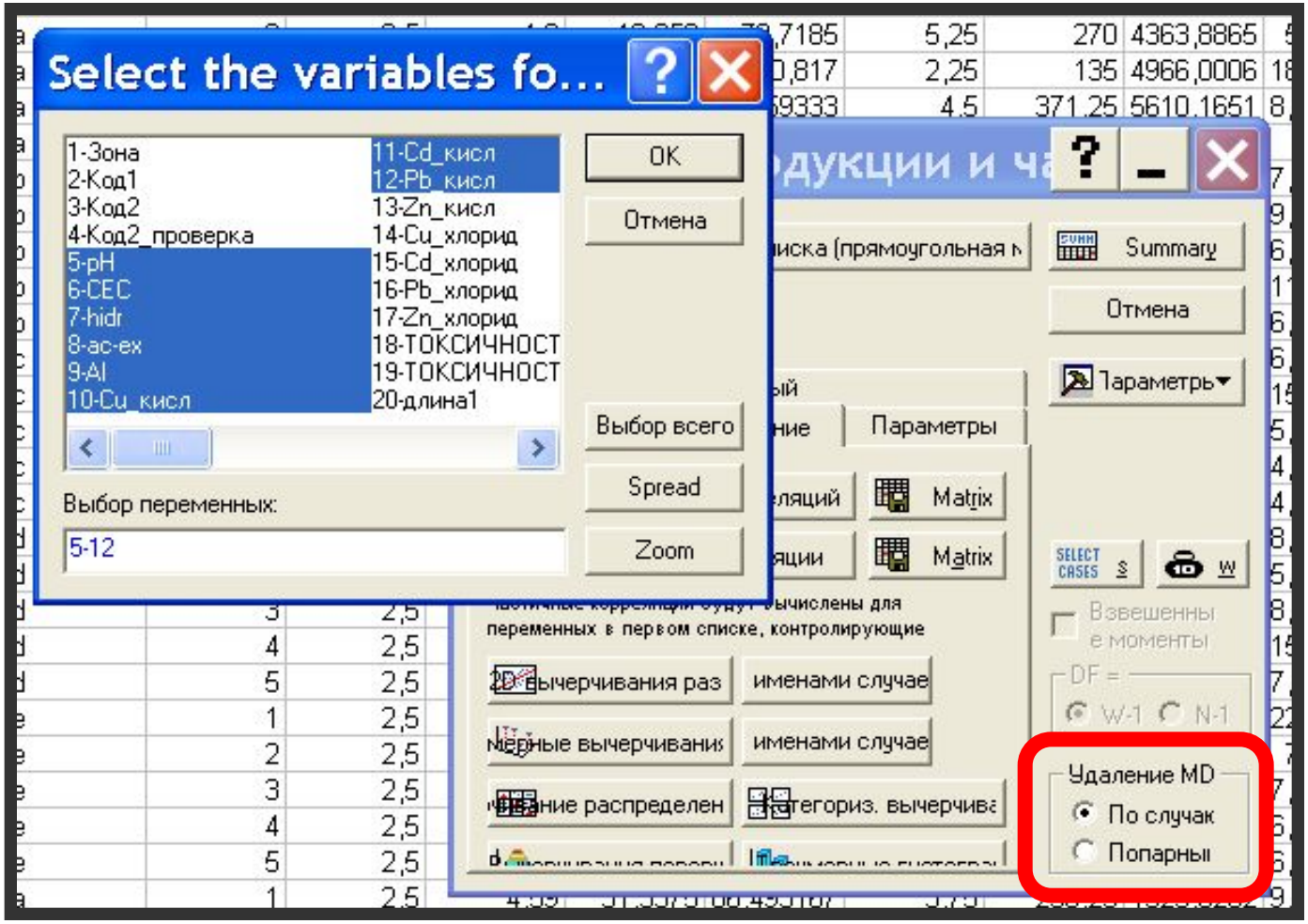
Требование к выборке для тестирования гипотезы о коэффициенте корреляции Пирсона:

Для каждого X значения Y должны быть распределены нормально, и для каждого Y все X должны иметь нормальное распределение -

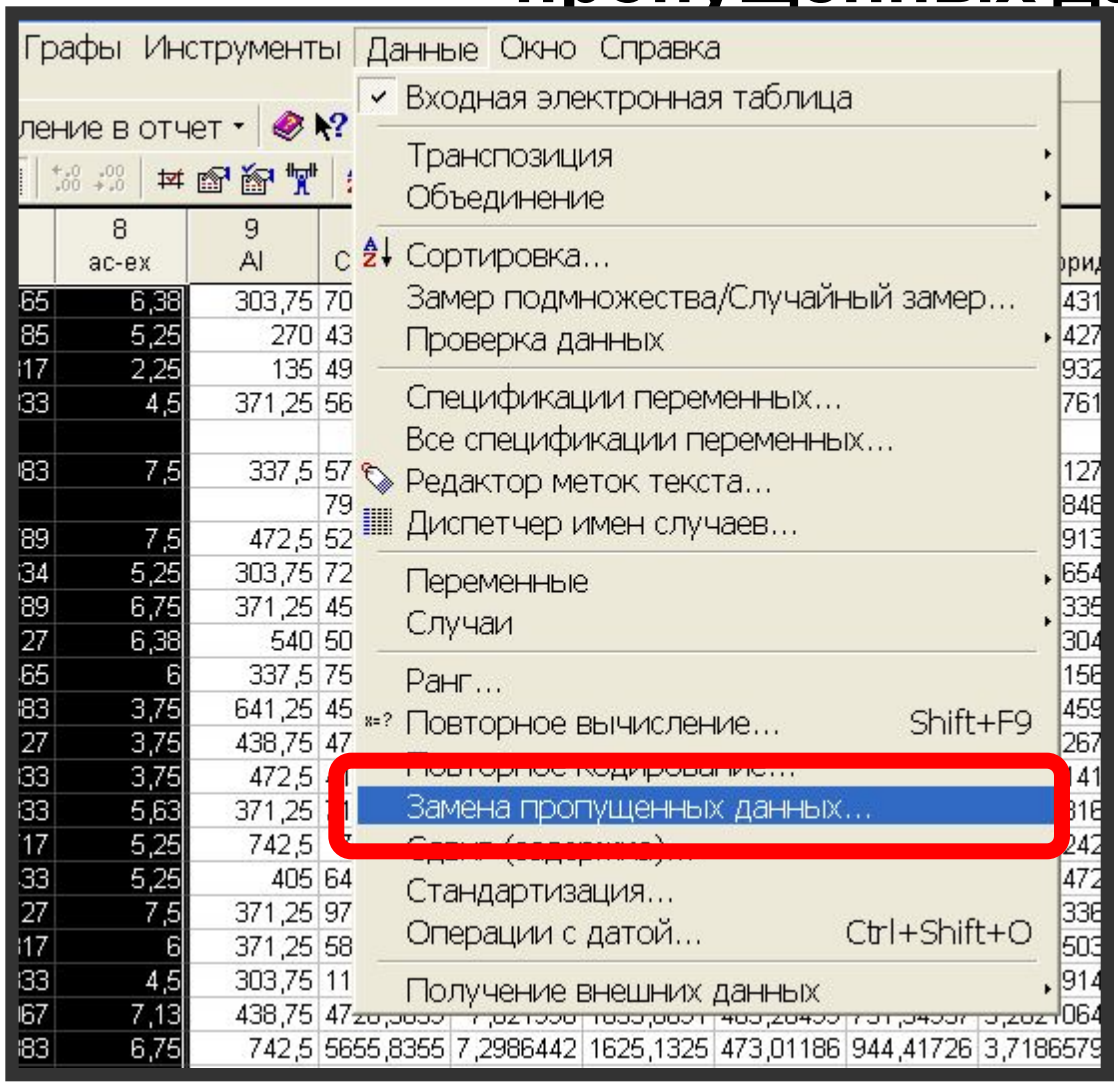
двумерное нормальное распределение (bivariate normal distribution)



Оперирование «пропущенными значениями» при расчете корреляционных матриц



Замена пропущенных значений средними: возможность минимизировать ущерб от пропущенных данных

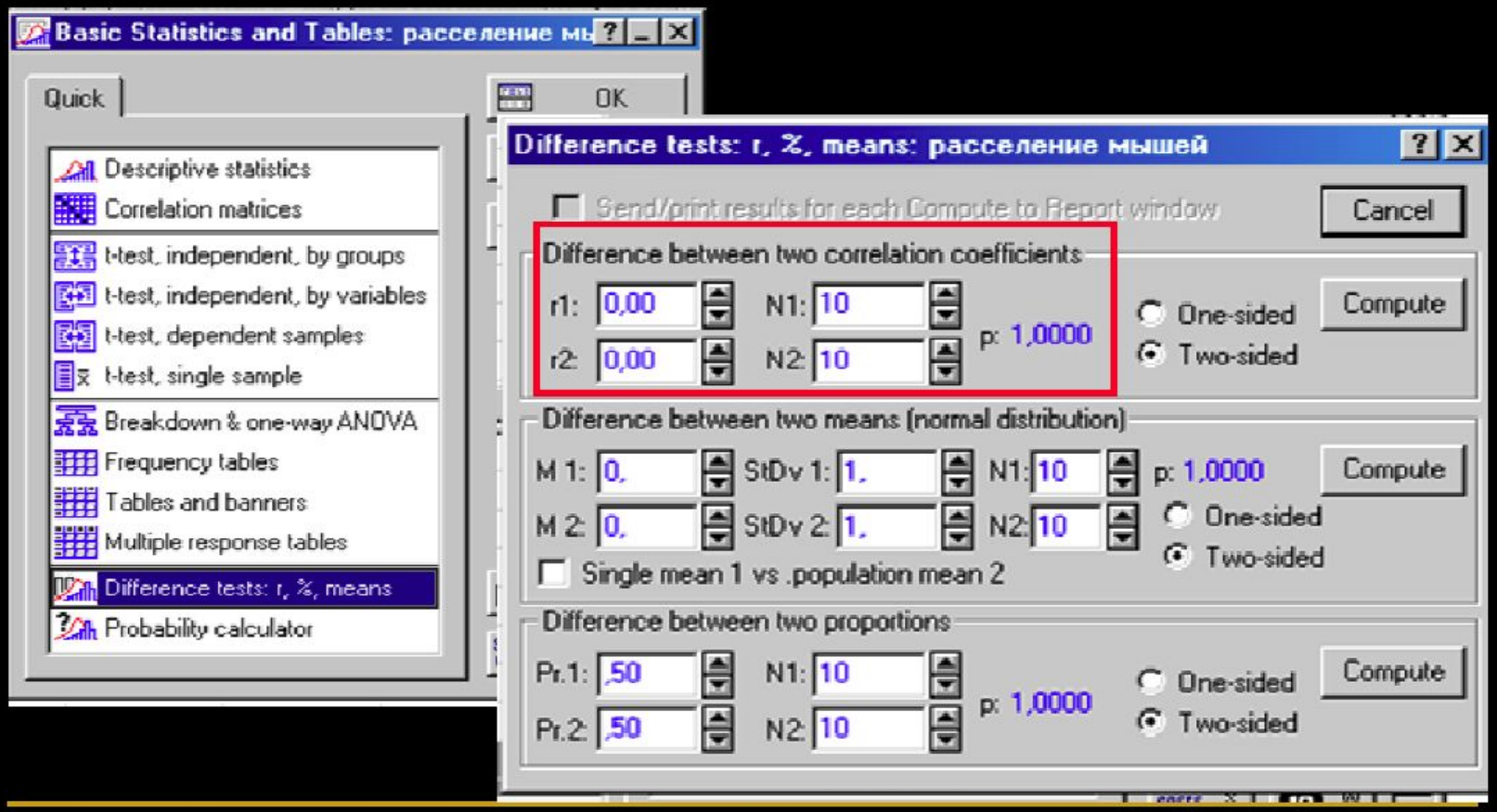


Пример_тм_токсичность_преобразования.xls;

Операции с переменными:
«CEC»
«Hidr»
«ac-ex»
«Al»

С учетом «зон»!

Можно сравнить два коэффициента корреляции от двух выборок

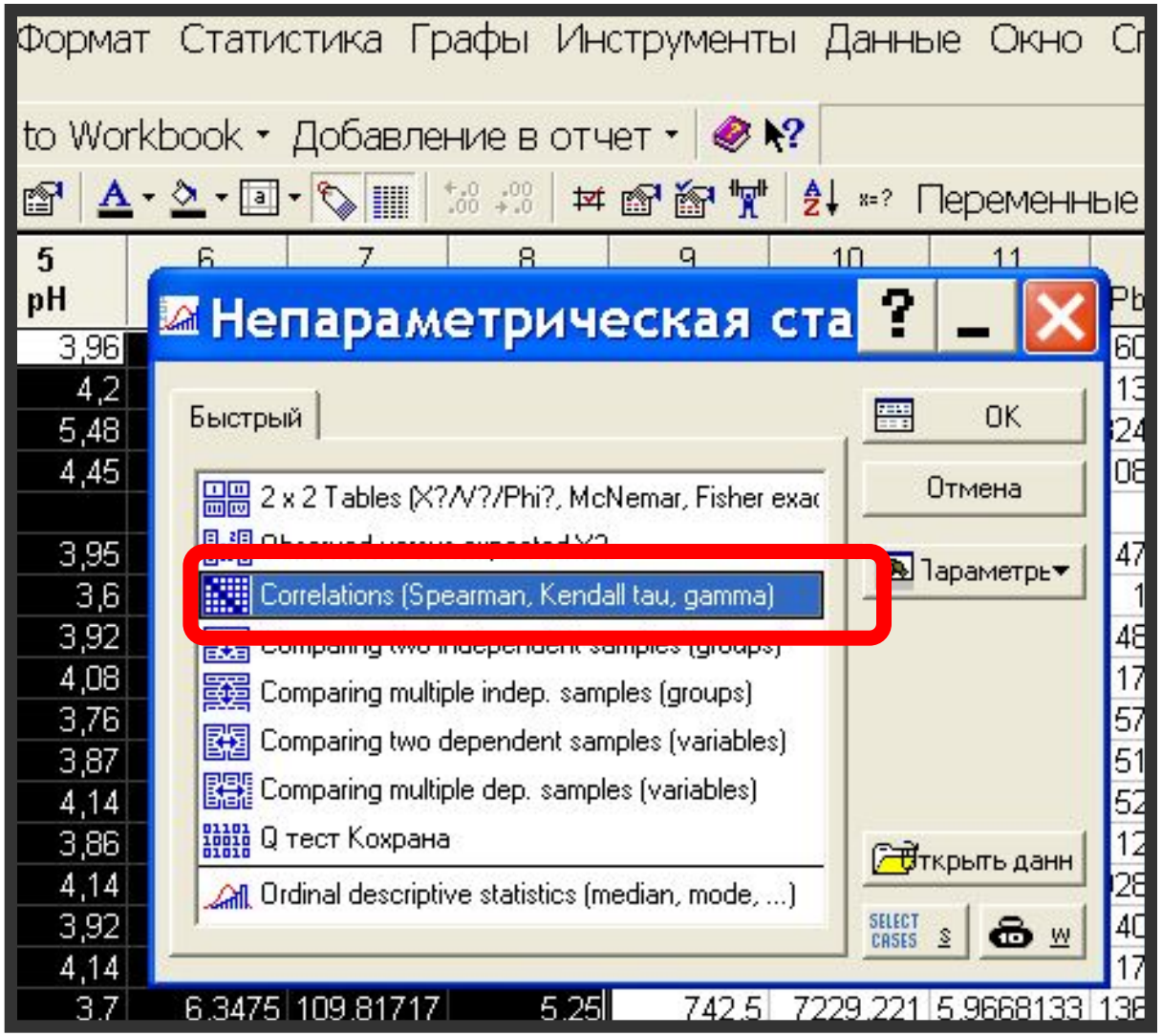


Для двумерного нормального распределения

Непараметрическая корреляция

Задача	Количественные нормально распределенные переменные	Количественные ненормально распределенные переменные или ранги
Оценить взаимосвязь между двумя переменными	Коэффициент парной корреляции Пирсона	**Коэффициенты ранговых корреляций (Спирмена, Кендалла)**

Коэффициент корреляции Спирмена



аналог коэффициента Пирсона;

подходит для расчета корреляционных матриц;

Размер выборки: >10.

Линейная (парная) регрессия

Задача: предсказать значение одной переменной на основании другой на основе аппроксимации – линии.

Переменные: зависимая (Y) и независимая (X).

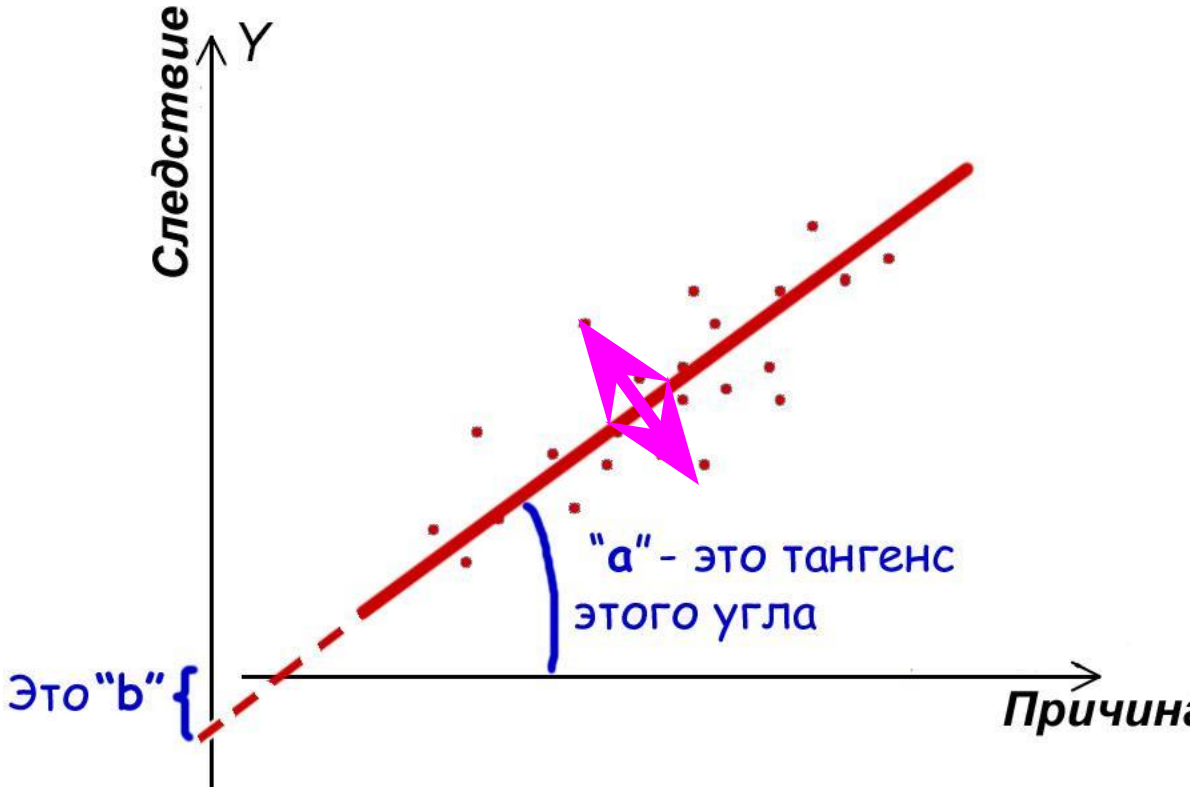
Предположения:

- линейная зависимость между переменными;
- независимость измерений отдельных X и Y от других измерений X и Y;
- двумерное нормальное распределение и нормальное распределение «остатков», т.е. разностей между наблюдаемыми и предсказываемыми величинами Y.
-

Интерпретация (при правильной постановке вопроса и правильном расчете): причинная и объясняющая.

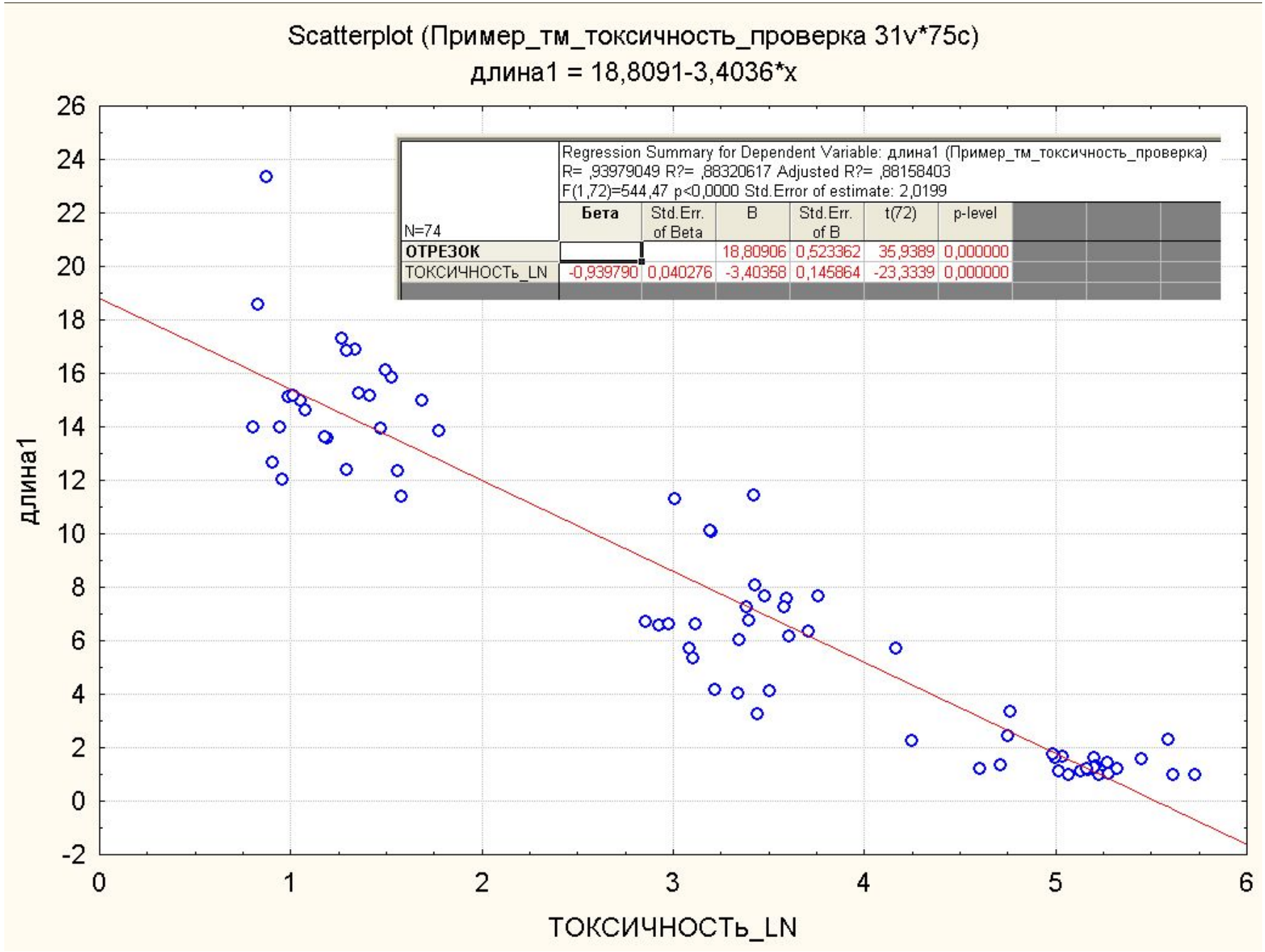
Формальное выражение: $Y = aX + b$.

Y = aX + b

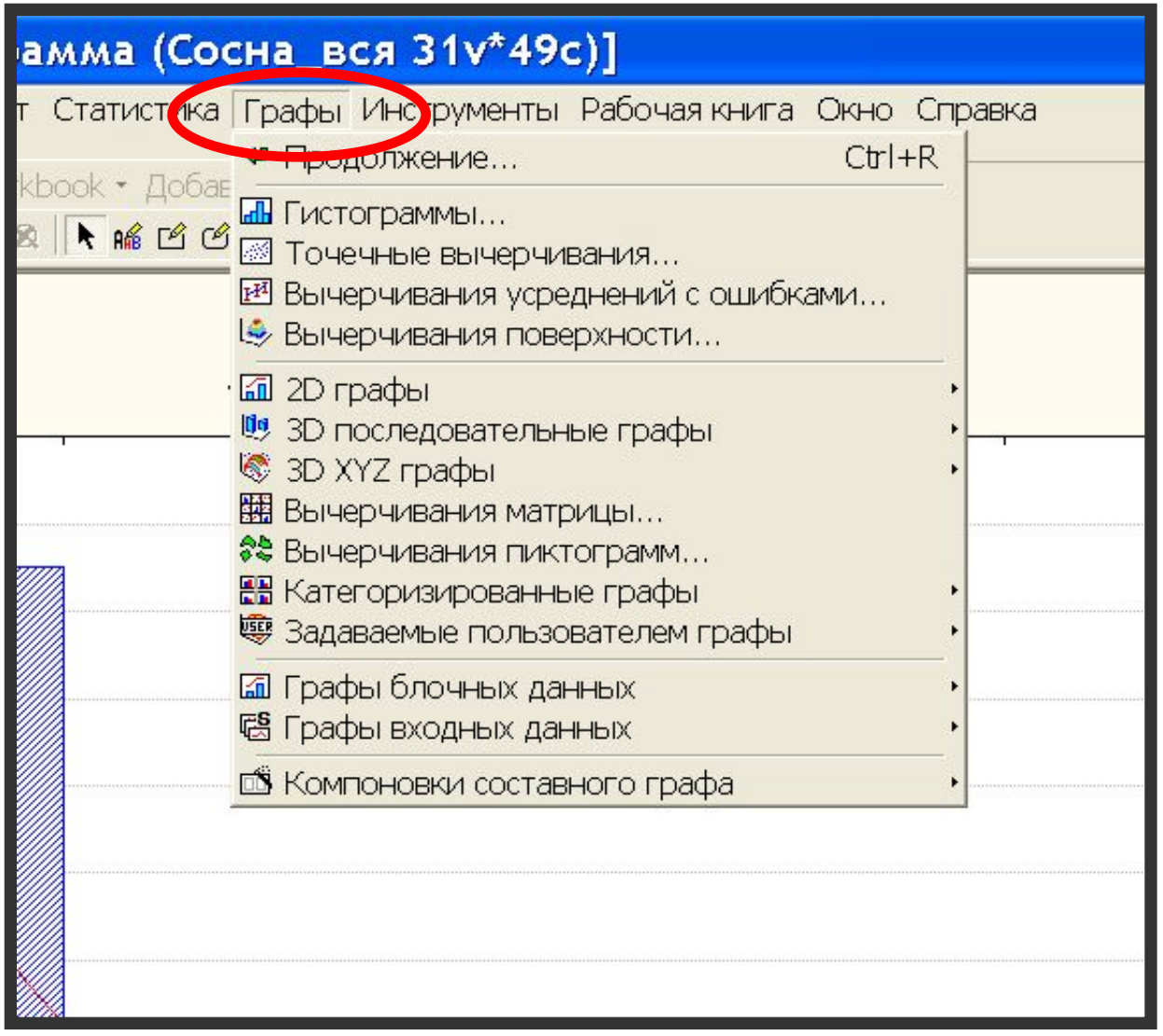


Изменчивость данных возле линии регрессии характеризует параметр R^2 – простой квадрат коэффициента корреляции Пирсона (в случае линейной регрессии).

НО!!!
Показатель R^2 приемлем и для нелинейных и для множественных зависимостей. Интерпретируется он как…..?

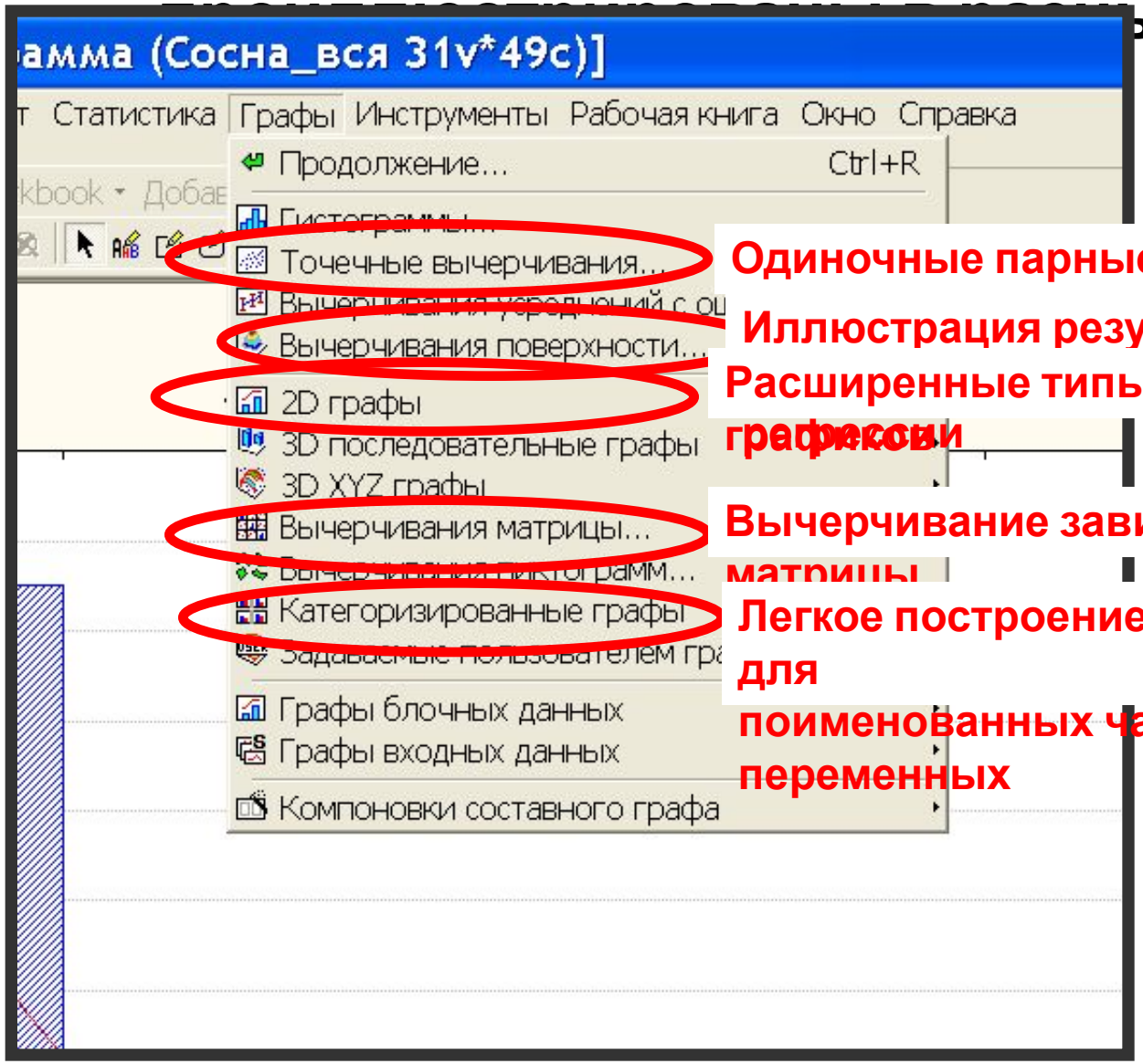

Scatterplot (Пример_тм_токсичность_проверка 31v*75c)
длина1 = 18,8091-3,4036*x
Regression Summary for Dependent Variable: длина1 (Пример_тм_токсичность_проверка)
R= ,93979049 R?= ,88320617 Adjusted R?= ,88158403
F(1,72)=544,47 p<0,0000 Std.Error of estimate: 2,0199
N=74
Бета
Std.Err. of Beta
B
Std.Err. of B
t(72)
p-level
ОТРЕЗОК
18,80906
0,523362
35,9389
0,000000
ТОКСИЧНОСТЬ_LN
-0,939790
0,040276
-3,40358
0,145864
-23,3339
0,000000
длина1
ТОКСИЧНОСТЬ_LN
26
24
22
20
18
16
14
12
10
8
6
4
2
0
-2
0
1
2
3
4
5
6

Пункт меню «Графы»: ШИРОЧАЙШИЕ возможность построения диаграмм как без расчета статистик, так и с расчетом таковых



Корреляционные и регрессионные зависимости могут быть

ых пунктах:



Одиночные парные

Иллюстрация результатов

Расширенные типы

Вычерчивание зависимостей в виде

Легкое построение зависимостей для

поименованных частей переменных

STATISTICA™

Том II:
ГРАФИКА

Настройка вида графиков: Инструменты → Параметры → Графы