**Лабораторная работа №5  
по курсу «Компьютерный анализ статистических данных»  
на тему «Анализ взаимосвязи случайных величин»**

## Цель работы

Изучить виды связи между случайными величинами, понятие корреляции и методы ее расчета. Научиться строить уравнения регрессии средствами Excel и получать по ним прогнозы.

## Теоретические сведения

### Корреляционная и функциональная связь

Часто бывает необходимо проверить наличие зависимости между двумя или более показателями.

*Зависимый* показатель обычно обозначают Y, а влияющий на него (*фактор*) X.

Примеры:

Зависимость заболеваемости раком (Y) от процента курящих (X).

Зависимость рождаемости (Y) от величины субсидий на ребенка (X).

Зависимость числа аварий (Y) от дальности видимости (X).

Зависимость объема продаж (Y) от затрат на рекламу (X).

Зависимость вероятности трудоустроиться (Y) от числа судимостей (X).

Зависимость уровня преступности (Y) от уровня жизни населения (X1), числа больных наркоманией (X2) и алкоголизмом (X3), числа сотрудников полиции (X4).

Зависимость производительности труда следователя (Y) от условий труда (X1), заработной платы (X2), опыта (стажа) (X3), образования (X4), степени автоматизации (X5).

Существуют различные **виды связи**.

*Функциональная связь*: каждому значению X соответствует ровно одно возможное значение Y, которое может быть вычислено по точной формуле.

Y = f(X)

Пример: площадь квадрата от длины его стороны:

S = a2

0

S

a

В экономике, социологии, правовой статистике такие зависимости встречаются крайне редко.

*Статистическая, случайная связь* – каждому значению X соответствует множество возможных случайных значений Y, что можно записать в виде

Y = f(X) + ε

ε – случайная величина.

Чаще всего, статистическая зависимость описывает влияние X на *среднее значение* Y.

mY = f(mX)

Фактические значения могут оказаться далеки от средних.

Пример

При средней безработице 10% (X) число квартирных краж на 100 000 жителей составляет в среднем 3,2 (Y), а при X = 15%, Y в среднем равно 3,8.

Другие классификации связей:

1. *Прямая и обратная*.

Прямая связь (*положительная,* рост): с ростом X растет и Y.

0

Y

X

*Обратная связь* (*отрицательная,* спад): с ростом X значения Y убывают.

0

Y

X

1. *Линейная и нелинейная*. Нелинейные связи могут быть различных видов.

*Линейная* связь описывается уравнением прямой линии:

Y = a + bX + ε

*Нелинейная* связь может быть:

* с ростом или спадом более быстрым, чем линейный:

0

Y

X

* с ростом или спадом более медленным, чем линейный:

0

Y

X

* с переменным направлением:

0

Y

X

Линейная статистическая взаимосвязь называется ***корреляцией***.

Корреляционная связь:

а) Показывает зависимость для средних значений.

б) Не показывает направление связи: X зависит от Y или Y зависит от X, поэтому говорят о взаимосвязи.

в) Проверяет только «синхронность» изменения значений X и Y, но не утверждает, что между ними есть логическая связь.

Явление, когда корреляция между двумя величинами сильна, но причинно-следственной связи между ними нет, называется ***ложной корреляцией***.

Примеры

Статистика показывает, что чем больше в городе пожарных расчетов, тем выше урон от пожаров.

Число самоубийств в США снижалось одновременно с долей пользователей, использующих Internet Explorer.

#### Коэффициент корреляции

*Коэффициент корреляции r* – это величина, показывающая силу и направление линейной связи между величинами.

–1 ≤ *r ≤* 1

Не позволяет судить о наличии нелинейной связи.

Чем ближе *r* к 1 или –1, тем сильнее связь, 0 – связь отсутствует. Положительное значение означает прямую (положительную) связь, отрицательное – обратную (отрицательную) связь.

|r| = 1 строгая функциональная связь.

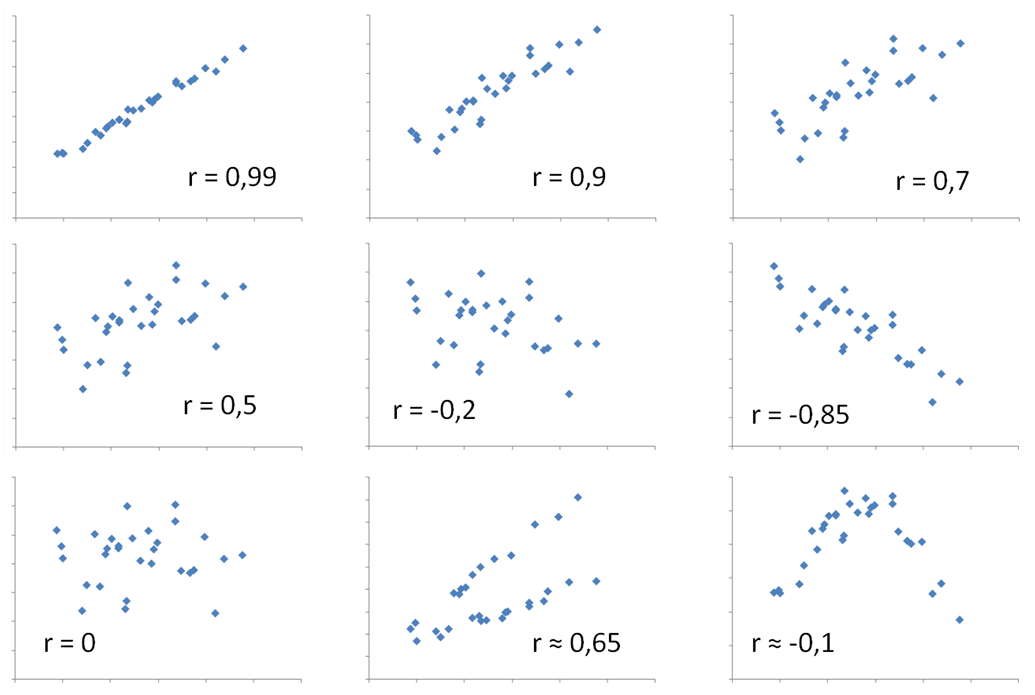
0,9 ≤ |r| < 1 очень сильная (тесная) связь, близкая к функциональной.

0,7 ≤ |r| < 0,9 сильная связь.

0,5 ≤ |r| < 0,7 присутствует слабая связь.

|r| < 0,5 связь очень слабая или отсутствует.

График, на которым показаны пары значений X-Y называют ***корреляционным полем***.



Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***№*** | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* |
| ***X*** | 27 | 32 | 42 | 8 | 41 | 23 | 12 | 22 | 24 | 31 | 27 | 43 | 28 | 50 | 54 | 47 |
| ***Y*** | 47 | 72 | 65 | 78 | 38 | 44 | 63 | 39 | 60 | 54 | 59 | 46 | 56 | 47 | 31 | 29 |

Корреляционное поле выглядит следующим образом:

Можно заметить, что с ростом X значения Y уменьшаются, но зависимость не очень сильная.

Формул для расчета коэффициента корреляции существует несколько.

Мы рассмотрим только *парную* корреляцию, т.е. между двумя признаками.

#### Метод параллельных рядов

*Основная суть*: при прямой связи бо́льшим (относительно среднего) значениям X соответствуют бо́льшие значения Y. При обратной связи бо́льшим значениям X соответствуют меньшие значения Y.

Алгоритм:

1. Отсортировать значения по возрастанию X.
2. Вычислить средние значения X и Y.
3. Отметить + значения, большие среднего и – значения, меньшие среднего. Если хотя бы одно из значений X или Y равно среднему, такая пара исключается из расчетов.
4. Подсчитать количество совпадений и несовпадений знаков  и ****.
5. Вычислить *коэффициент корреляции знаков Фехнера* по формуле:



Недостаток: не учитывает значения X и Y, поэтому позволяет определить наличие и направление связи, но не позволяет определить, насколько она сильная.

Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **№** | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* | ***m*** |
| **X** | 8 | 12 | 22 | 23 | 24 | 27 | 27 | 28 | 31 | ~~32~~ | 41 | 42 | 43 | 47 | 50 | 54 | **32** |
| **Зн.X** | – | – | – | – | – | – | – | – | – | ~~0~~ | + | + | + | + | + | + |  |
| **Y** | 78 | 63 | 39 | 44 | 60 | 59 | 47 | 56 | 54 | ~~72~~ | 38 | 65 | 46 | 29 | 47 | 31 | **52** |
| **Зн.Y** | + | + | – | – | + | + | – | + | + | ~~+~~ | – | + | – | – | – | – |  |
| **Совп.** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

 = 4

 = 11

KФ = (4 – 11) / (4 + 11) = –7 / 15 = –0,467

**Вывод**: KФ < –0,5 связь очень слабая или отсутствует.

#### Ранговая корреляция Спирмена

Учитывает не сами значения X и Y, а их ранги (порядковые номера).

Алгоритм:

1. Вычислить ранги каждого значения X и Y.
2. Вычислить разность между рангами *d* = ранг X – ранг Y
3. Вычислить квадраты разностей *d*2 и их общую сумму.
4. Вычислить коэффициент ранговой корреляции Спирмена по формуле:



Для каждого повторяющегося ранга (связки) в формулу вводят следующие поправку:



 – число повторяющихся значений в каждой связке.



Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***№*** | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* |
| ***X*** | 27 | 32 | 42 | 8 | 41 | 23 | 12 | 22 | 24 | 31 | 27 | 43 | 28 | 50 | 54 | 47 |
| ***Ранг*** | 6,5 | 10 | 12 | 1 | 11 | 4 | 2 | 3 | 5 | 9 | 6,5 | 13 | 8 | 15 | 16 | 14 |
| ***Y*** | 47 | 72 | 65 | 78 | 38 | 44 | 63 | 39 | 60 | 54 | 59 | 46 | 56 | 47 | 31 | 29 |
| ***Ранг*** | 7,5 | 15 | 14 | 16 | 3 | 5 | 13 | 4 | 12 | 9 | 11 | 6 | 10 | 7,5 | 2 | 1 |
| ***d*** | –1 | –5 | –2 | –15 | 8 | –1 | –11 | –1 | –7 | 0 | –4,5 | 7 | –2 | 7,5 | 14 | 13 |
| ***d2*** | 1 | 25 | 4 | 225 | 64 | 1 | 121 | 1 | 49 | 0 | 20,25 | 49 | 4 | 56,25 | 196 | 169 |

*n* = 15 (без учета одинаковых рангов)

 = 985,5

Связки:

A1 = A2 = 2

A = 1/12 \* (2 \* (4 – 1) + 2 \* (4 – 1)) = 12/12 = 1

*RC* = 1 – (6 \* 985,5 + 1) / (15 \* (225 – 1)) = –0,76

**Вывод**: присутствует сильная ранговая корреляция.

#### Коэффициент корреляции Пирсона

Обычно под «коэффициентом корреляции» понимают именно его. Вычисляется по формуле:



 – среднее значение произведений X на Y.

Коэффициент корреляции Пирсона позволяет оценить силу и направление линейной связи и учитывает все значения величин, но довольно чувствителен к случайным выбросам.

Для него существует статистический тест, позволяющий проверить наличие корреляционной связи для конкретного объема выборки.

**Проверка наличия корреляционной связи по T-критерию Стьюдента**

X и Y должны иметь нормальное распределение.

|  |  |
| --- | --- |
| *Гипотезы* | H0: rXY = 0  H1: rXY ≠ 0 |
| *Критерий* |  |
| *Критическое значение* |  |
| *Критическая область* | |*T*| ≥ *tкр*. => rXY ≠ 0 |

Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***№*** | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* |
| ***X*** | 27 | 32 | 42 | 8 | 41 | 23 | 12 | 22 | 24 | 31 | 27 | 43 | 28 | 50 | 54 | 47 |
| ***Y*** | 47 | 72 | 65 | 78 | 38 | 44 | 63 | 39 | 60 | 54 | 59 | 46 | 56 | 47 | 31 | 29 |
| ***X\*Y*** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

mX = 31,94

mY = 51,75

mXY = 1546,9

SX = 13,22

SY = 14,10

**rXY** =(1546,9 – 31,94 \* 51,75) / (13,22 \* 14,10) = **–0,5675**

T = –0,5675 \* √((16 – 2) / (1 – 0,56752)) = – 2,58

tкр. = 2,12 (α = 0,05)

**Вывод**: |T| > tкр. Присутствует корреляционная связь между X и Y.

### Регрессионный анализ

**Регрессия** – это математическая модель, описывающая зависимость одного статистического показателя (Y) от другого (X).



Y называют регрессором, X – фактором, влияющим на регрессор,  – ошибки модели (случайные отклонения).

Наличие в модели случайных отклонений указывает на то, что регрессия – это статистическая связь, а не функциональная.

0

Y

X

ε

Регрессия может быть как линейной, так и нелинейной.

В работе рассматриваются

* линейная ;
* парабола (полином степени 2) ;
* степенная модель ;
* экспоненциальная (показательная) модель ;
* логарифмическая модель .

 называются параметрами модели, это числовые значения, которые необходимо найти (оценить), так, чтобы модель проходила как можно ближе к исходным данным, т.е. чтобы значения ошибок были минимальными.

В расчетах по модели ε не используется, модель без случайных остатков (расчетные значения) обозначается звездочкой, например:



Основная цель построения таких моделей – *прогнозирование*, т.е. определение, чему в среднем будет равен *Y\** при заданном значении *X\**.

Чтобы выбрать лучшую регрессию, нужно *сравнить их по точности*. Для оценки точности регрессии наиболее часто используют *коэффициент детерминации R*2.

Коэффициент детерминации изменяется в пределах от 0 до 1. **Чем больше *R*2, тем точнее модель.** Возможно следующее толкование коэффициента детерминации: *R*2 показывает долю исходных данных, которые описываются моделью, т.е. если *R*2 = 0,95, то модель описывает 95% данных.

Если *R*2 = 1, то регрессия проходит точно через пары значений X и Y, все ε = 0. Однако на практике такие случаи не встречаются.

Считается, что регрессия обладает хорошей точностью, если *R*2 > 0,7. Если *R*2 < 0,5, то модель считается неудовлетворительной, ненадежной.

Выбирать следует модель с самым большим R2. Но если для двух моделей разница составляет менее 0,01, то лучше выбрать более простую модель (линейная модель самая простая, парабола – самая сложная, для всех остальных сложность примерно одинаковая).

Для линейной модели коэффициент детерминации равен квадрату коэффициента корреляции Пирсона:

.

Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***№*** | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *14* | *15* | *16* |
| ***X*** | 27 | 32 | 42 | 8 | 41 | 23 | 12 | 22 | 24 | 31 | 27 | 43 | 28 | 50 | 54 | 47 |
| ***Y*** | 47 | 72 | 65 | 78 | 38 | 44 | 63 | 39 | 60 | 54 | 59 | 46 | 56 | 47 | 31 | 29 |

Какие значения Y следует ожидать при X = 30? X = 40? X = 50?

Были получены следующие уравнения регрессии:

1. **Y = 72,3–0,64X** R² = 0,366
2. Y = 151,4X–0,33 R² = 0,352
3. Y = 76,13e–0,01X **R² = 0,375**

Самый большой R2 у модели (3), но он только на 0,009 больше, чем у линейной регрессии. Поэтому выбираем линейную регрессию.

Тогда:

*X\** = 30 *Y\** = 72,3 – 0,64 \* 30 = 53,1

*X\** = 40 *Y\** = 72,3 – 0,64 \* 40 = 46,7

*X\** = 50 *Y\** = 72,3 – 0,64 \* 50 = 40,3

В целом, все значения R2 маленькие (меньше 0,5), поэтому полученный прогноз будет очень ненадежным, грубым, неточным.

### Рекомендуемая литература

1. **Лялин, В.С. Правовая статистика**: учебник для студентов вузов, обучающихся по специальности «Юриспруденция», для курсантов и слушателей образовательных учреждений МВД. – М.: ЮНИТИ-ДАНА, 2010. – 255 с. (***Глава 9***)
2. **Савюк, Л.К. Правовая статистика**: Учебник.— М.: Юристъ, 2004. — 588 с. (***Глава 14***)

### Задание

Имеется выборки двух показателей X и Y. Проверить наличие, силу и направление связи между ними, используя:

1. Графический анализ (корреляционное поле).
2. Метод параллельных рядов (корреляцию знаков Фехнера).
3. Ранговый коэффициент корреляции Спирмена.
4. Коэффициент корреляции Пирсона, проверить его значимость с помощью T-статистики.
5. Регрессионный анализ, построить модель зависимости Y от X. Спрогнозировать значение Y для заданного X\*.

### Подготовка исходных данных

Работа выполняется в том же файле, что и другие, но на новом листе. Добавьте новый лист и переименуйте в «Л.р.5». Скопируйте данные для своего варианта на лист «Л.р.5».

Обратите внимание, ниже рядов X и Y находится описание показателей X и Y и значения X\*.

### Пример

Все расчеты показаны на примере зависимости числа преступлений, связанных с незаконным оборотом наркотических веществ в некотором регионе Y от числа лиц, страдающих наркоманией и состоящих на учете в наркологических диспансерах X (тыс. чел.):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X** | **Y** | **X** | **Y** | **X** | **Y** | **X** | **Y** |
| 10,89 | 3146 | 9,85 | 3329 | 7,67 | 1591 | 8,08 | 2542 |
| 9,09 | 3520 | 9,41 | 3264 | 10,33 | 4418 | 7,89 | 2926 |
| 8,01 | 2279 | 14,56 | 9140 | 8,82 | 2903 | 3,56 | 570 |
| 11,56 | 3371 | 14,6 | 7557 | 8,72 | 1735 | 7,53 | 1107 |
| 3,28 | 262 | 8,87 | 2131 | 10,58 | 3618 | 9,76 | 2404 |
| 11,73 | 4323 | 11,22 | 3646 | 7,22 | 1105 | 11,01 | 3877 |
| 10,28 | 3193 | 3,43 | 543 | 13,67 | 8507 | 4,17 | 960 |
| 12,82 | 5416 | 9,92 | 3032 | 11,94 | 3455 | 12,51 | 3074 |

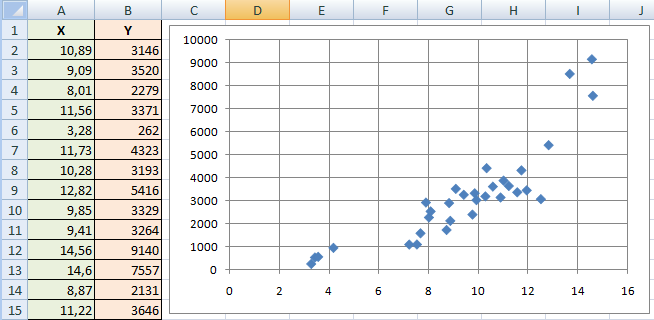
X\* = 11,5

Примечание При качественном анализе этих показателей необходимо учитывать, что это только официальная статистика. По результатам анонимных опросов, число лиц, употребляющих наркотики, как правило, в несколько раз выше числа официально зарегистрированных.

## Указания к выполнению работы

1. Построение корреляционного поля – наиболее простой и наглядный способ проверить наличие зависимости между исходными данными.

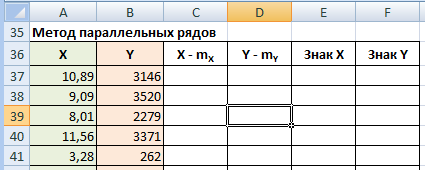
Выделите данные для X и Y и добавьте диаграмму типа «Точечная». Настройте внешний вид диаграммы: удалите лишние надписи, добавьте вертикальные линии сетки.



**Вывод**: Очевидно, что с ростом числа лиц, употребляющих наркотические вещества, увеличивается и число связанных с ними преступлений, т.е. наблюдается прямая (положительная) зависимость между показателями X и Y.

Однако по графику нельзя однозначно сказать, линейная зависимость или нет. Можно заметить, что для регионов с очень большими значениями X (три точки в верхней правой части графика), значения Y растут немного быстрее. В то же время в центральной части графика точки расположены ближе к прямой линии.

1. Проанализируем зависимость между X и Y с помощью метода параллельных рядов. Скопируйте исходные данные для X и Y и вставьте их ниже на листе. Добавим заголовки 4 столбцов для расчетов.



Данные необходимо отсортировать по X. Для этого выделите оба ряда (X и Y) вместе с заголовками. На ленте на вкладке Главная –Сортировка и фильтр – Настраиваемая сортировка... Выполните сортировку по столбцу X по возрастанию.

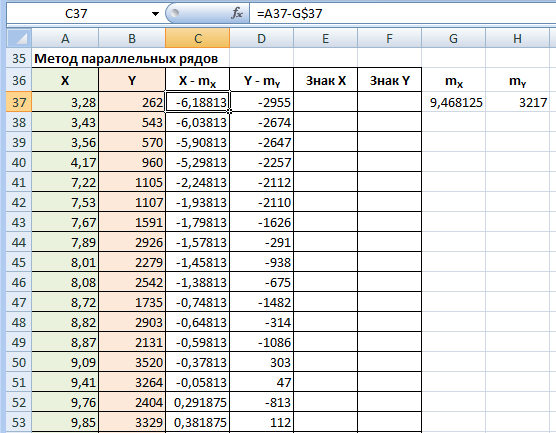
Результат сортировки для рассматриваемого примера:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X** | **Y** | **X** | **Y** | **X** | **Y** | **X** | **Y** |
| 3,28 | 262 | 8,01 | 2279 | 9,85 | 3329 | 11,56 | 3371 |
| 3,43 | 543 | 8,08 | 2542 | 9,92 | 3032 | 11,73 | 4323 |
| 3,56 | 570 | 8,72 | 1735 | 10,28 | 3193 | 11,94 | 3455 |
| 4,17 | 960 | 8,82 | 2903 | 10,33 | 4418 | 12,51 | 3074 |
| 7,22 | 1105 | 8,87 | 2131 | 10,58 | 3618 | 12,82 | 5416 |
| 7,53 | 1107 | 9,09 | 3520 | 10,89 | 3146 | 13,67 | 8507 |
| 7,67 | 1591 | 9,41 | 3264 | 11,01 | 3877 | 14,56 | 9140 |
| 7,89 | 2926 | 9,76 | 2404 | 11,22 | 3646 | 14,6 | 7557 |

Таким образом, даже без детальных вычислений, по отсортированным данным можно заметить, что при увеличении значений X, Y тоже увеличивается.

Подтвердим это с помощью знаковой корреляции Фехнера. Вычислим средние значения X и Y, а затем из каждого исходного значения вычтем соответствующее среднее.

Если расчеты выполнены верно, то значения X–mX будут возрастать от отрицательных к положительным. Для Y это не соблюдается.



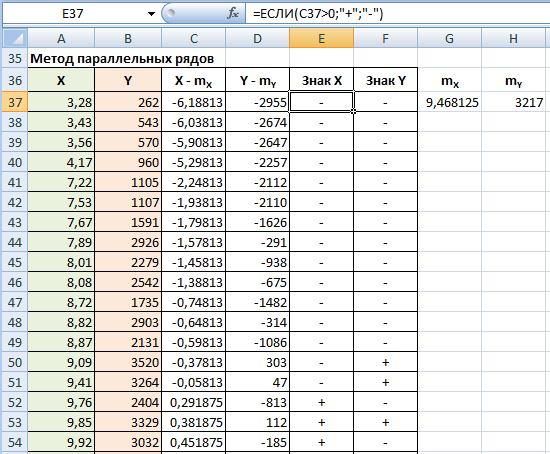
Теперь необходимо определить знаки разностей.

Примечание На практике знаки отдельно вычислять необязательно, можно использовать сами разности, но с ними повышается наглядность расчетов.

Для этого воспользуемся функцией ЕСЛИ(*условие*; *значение1*; *значение2*) , которая в зависимости от условия возвращает одно из двух значений: если *условие* выполнено – то *значение1*, если не выполнено – то *значение2*.

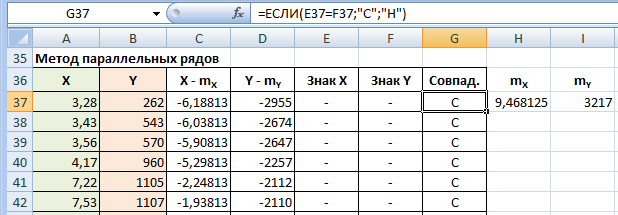
Нам необходимо подставить в ячейку +, если разность >0, и –, если нет. Знаки + и – в Excel необходимо вписать в кавычках, т.к. это текст, а не числа.

Примечание Если в вашем варианте есть разности, строго равные 0, удалите соответствующие строки из расчета.

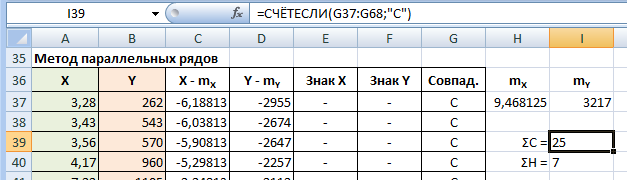


Видим, что в большинстве случаев знаки разностей совпадают, но не всегда. Вычислим число совпадений и число несовпадений знаков.

Добавим еще один столбец «Совпадение». Совпадающие знаки будем обозначать «C», несовпадающие – «Н». Расчет выполним также через функцию ЕСЛИ.

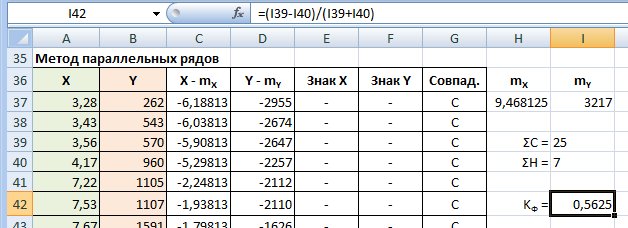


Количество совпадений и несовпадений (т.е. букв С и Н) можно подсчитать с помощью функции СЧЁТЕСЛИ(*диапазон*; *условие*). Диапазон – столбец «Совпадение», условие – буква С или Н.



Осталось вычислить коэффициент корреляции знаков Фехнера по формуле:





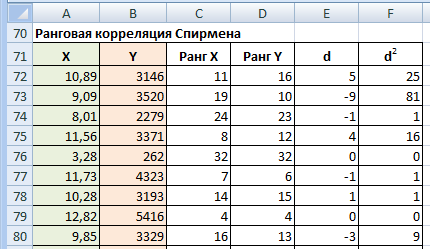
**Вывод**: коэффициент корреляции знаков Фехнера положительный, превосходит 0,5, следовательно, наблюдается прямая положительная зависимость между X и Y.

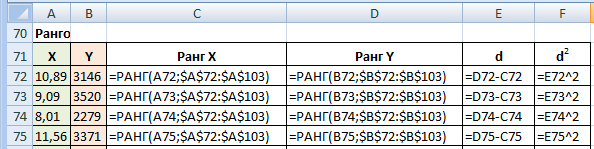
Отчет

Добавьте в отчет формулу и результаты расчета коэффициента корреляции знаков Фехнера и сделайте вывод по его значению.

1. Теперь проверим наличие зависимости с помощью другого коэффициента корреляции – ранговой корреляции Спирмена.

Сделайте еще одну копию исходных данных (без сортировки), добавьте 4 столбца для расчетов: ранги для X и Y, их разности *d* и квадрат разности *d*2. Заполните столбцы соответствующими формулами.



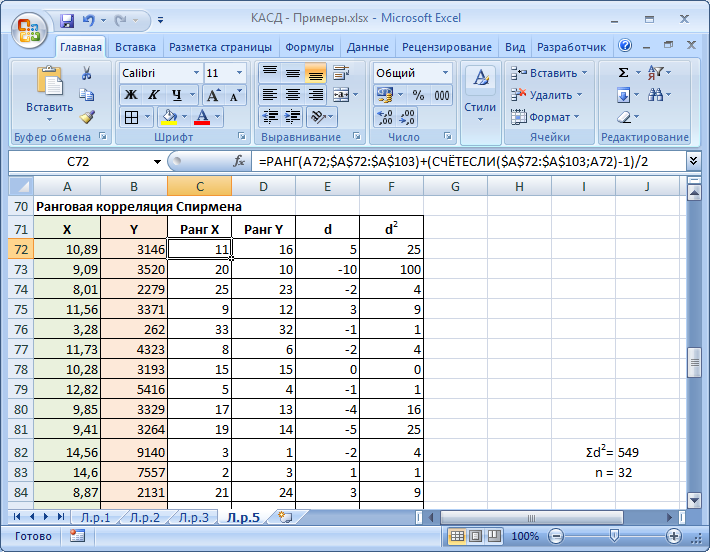


Внимание!

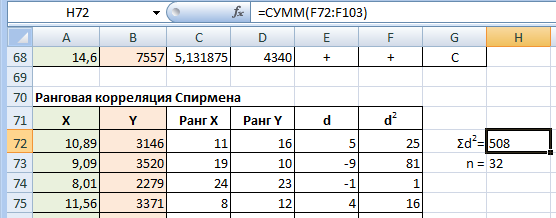
В данном примере нет повторяющихся значений (связок) ни у X, ни у Y. Если они есть в вашем варианте, необходимо рассчитать связанные ранги и добавить поправку A на их количество в числитель формулы *R*С.

Связанный ранг = Ранг + (Количество повторов – 1) / 2

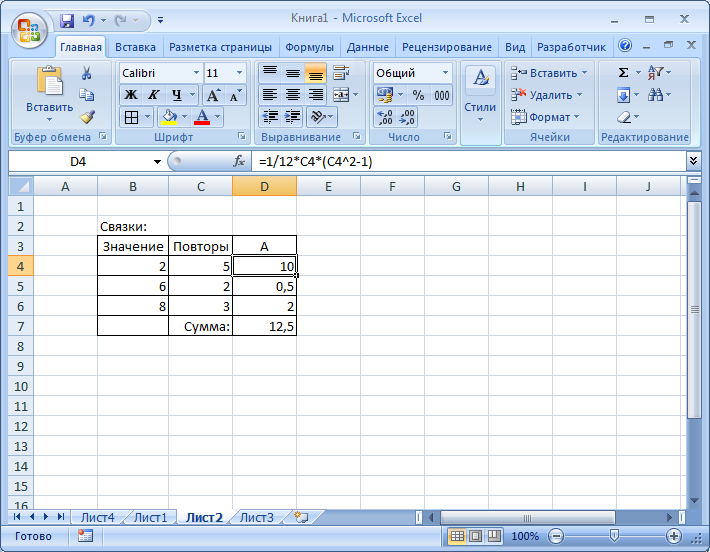
Количество повторов можно вычислить с помощью функции СЧЁТЕСЛИ:



Вычислим сумму *d*2 и число пар наблюдений (считать только по X или только по Y).



Если в вашем варианте есть повторяющиеся значения, выпишите их в таком виде (к рассматриваемому примеру не относится):



Здесь, например, значение 2 встречается 5 раз, а значение 6 – 2 раза.

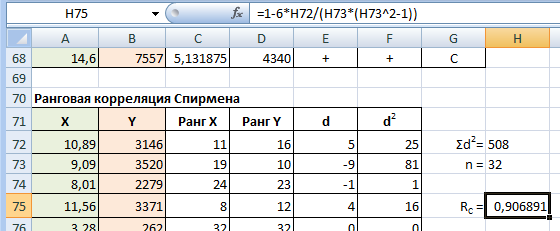
Получившуюся сумму нужно прибавить к числителю в формуле ранговой корреляции.

Вычислим коэффициент ранговой корреляции Спирмена по формуле:



При наличии повторов значений (связок):





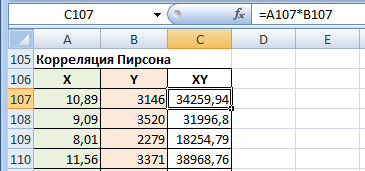
**Вывод**: Коэффициент ранговой корреляции Спирмена равен 0,907, что показывает наличие сильной положительной зависимости между X и Y.

Отчет

Добавьте в отчет формулу и результаты расчета коэффициента ранговой корреляции Спирмена и сделайте вывод по его значению. Укажите, были ли в ваших данных повторяющиеся значения и использовали ли вы поправку в расчетах.

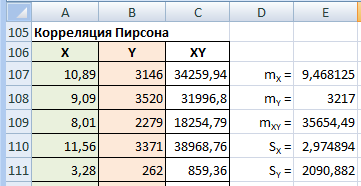
1. Рассчитаем коэффициент корреляции Пирсона, который чаще всего применяется на практике.

Сделайте новую копию исходных данных, добавьте столбец с расчетом произведений XY.



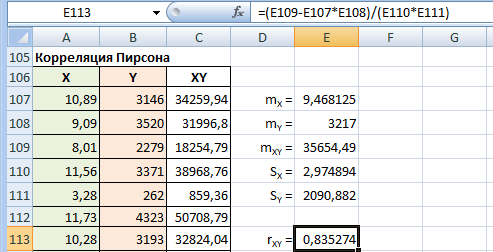
Для расчета корреляции вычислим средние значения и стендартные отклонения mXY, mX, mY, SX, SY.

Примечание mX, mY можно взять из расчета корреляции знаков.

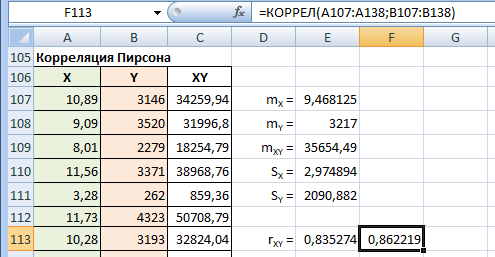


Рассчитаем значение коэффициента корреляции по формуле:





Проверим правильность расчетов, используя встроенную формулу Excel для расчета корреляции КОРРЕЛ(*X*;*Y*).



Есть небольшая разница в значениях, поскольку мы использовали приближенную формулу расчета корреляции, но в целом результаты совпадают. Коэффициент корреляции положительный и превосходит 0,7 – наблюдается сильная линейная связь.

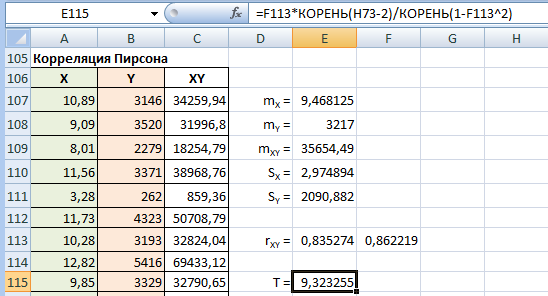
Примечание Чтобы получить такое же значение, как у функции КОРРЕЛ, нужно при расчете SX, SY в формуле делить не на (*n*–1), а на *n*, или использовать не СТАНДОТКЛОН, а СТАНДОТКЛОН**П**.

Проверим значимость коэффициента корреляции с помощью *T*-статистики Стьюдента.

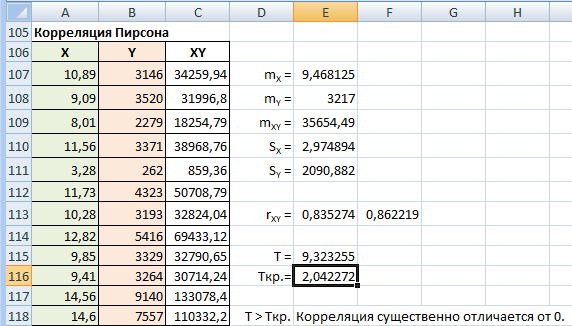
Вычислим *T*-статистику по формуле:



Для расчета будем использовать более точное значение корреляции, полученное с помощью функции КОРРЕЛ.



Вычислим критическое значение критерия для уровня значимости α = 0,05 по формуле СТЬЮДРАСПОБР(α;*n–2*) и сделаем вывод.



**Вывод**: Коэффициент корреляции Пирсона  существенно отличается от нуля, между рассматриваемыми показателями имеется тесная положительная связь.

Отчет

Добавьте в отчет формулу и результаты расчета коэффициента корреляции Пирсона и сделайте вывод по его значению. Что показывает тест Стьюдента?

1. Все коэффициенты корреляции показали наличие положительной линейной связи между показателями. Но они не учитывают возможность наличия нелинейной связи.

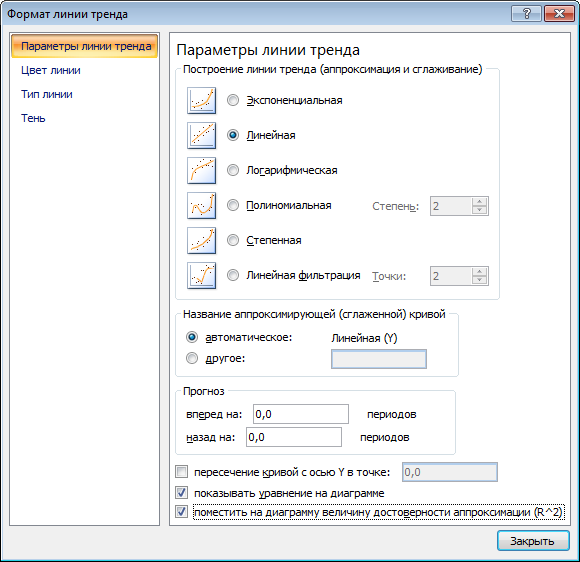
Выполним регрессионный анализ, чтобы проверить различные модели зависимости X от Y и выбрать наилучшую из них.

Excel позволяет автоматически рассчитывать следующие модели регрессии (линии тренда):

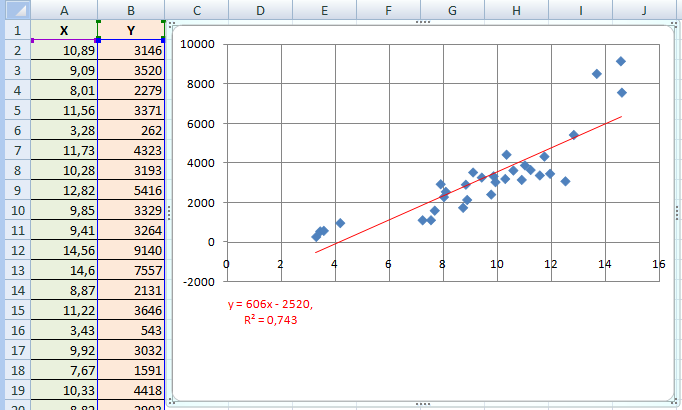
1. линейную;
2. экспоненциальную;
3. логарифмическую;
4. полиномиальную, в том числе параболу (для степени 2);
5. степенную;
6. кусочно-линейную (линейная фильтрация)

Вернитесь к построенному в первом пункте графику. Выделите построенный ряд (левый клик по точкам на графике), щелкните по ним правой кнопкой мыши и выберите «Добавить линию тренда».

В окне выберите «Линейную» и отметьте галочки «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации».

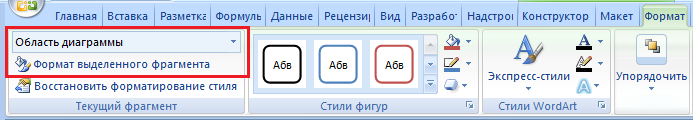


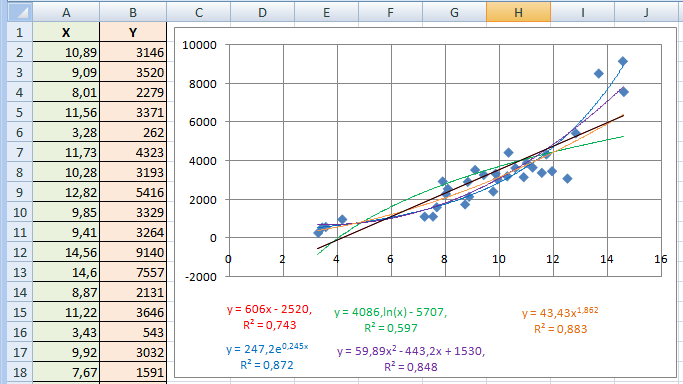
По умолчанию все линии тренда рисуются черными, и их очень трудно различать. Сразу перекрасьте линию и соответствующее ей уравнение в другой цвет. Разместите уравнение справа или ниже графика.



По аналогии добавьте другие линии тренда, кроме линейной фильтрации. Если на одном графике они сливаются, можно построить на нескольких отдельных графиках.

Если не получается выделить линию тренда на графике, чтобы ее перекрасить, то список всех линий на графике можно найти на ленте на вкладке «Формат» (диаграмма должна быть выделена). Там же есть кнопка «Формат выделенного фрагмента».





**Вывод**: У экспоненциальной, параболической и степенной регрессии коэффициент детерминации (линейной аппроксимации) значительно выше, чем у линейной, т.е. они лучше описывают зависимость Y от X. Это означает, что при увеличении числа официально зарегистрированных больных наркоманией (X) число преступлений, связанных с оборотом наркотиков, (Y) растет еще быстрее. Причин тому может быть несколько:

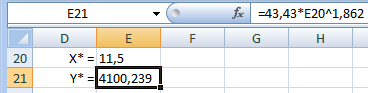
1. Наркодилерам невыгодно «вкладываться» в регионы с небольшим количеством покупателей, и выгодно вкладываться в регионы с их большим количеством, поэтому большая часть оборота наркотиков приходится на регионы с большим числом больных наркоманией.
2. В исходных данных используется официальный учет как больных наркоманией, так и числа преступлений. Возможно, между реальными показателями зависимость линейная. Но а) для ряда регионов с большим оборотом наркотиков меньшая доля наркоманов стоит на учете, или б) в регионах с высоким оборотом наркотиков выше раскрываемость соответствующих преступлений.
3. Не учитывается, что в разных регионах разное количество населения. Следовательно, меньшее число лиц, страдающих от наркомании, будет в регионах с меньшим населением, и наоборот. В крупных регионах больше зарегистрированных больных наркоманией и больше оборот наркотрафика.

Наибольшей точностью по R2 обладает степенная модель

Y = 43,43x1,862

Построим по этой модели прогноз среднего числа преступлений для числа больных наркоманией X\* = 11,5:

Y\* = 43,43∙11,51,862= 4100



Таким образом, при 11,5 тыс. официально состоящих на учете больных наркоманией, число преступлений, связанных с оборотом наркотиков, ожидается в районе 4100.

Отчет

Добавьте в отчет выбранное уравнение регрессии и полученный прогноз для X\* из вашего варианта. Поясните, какими причинами может быть обусловлена такая зависимость.