Тема № 5
«Понятие статистической взаимосвязи»

к. ф.-м. н., доцент
Озёрский Сергей Владимирович

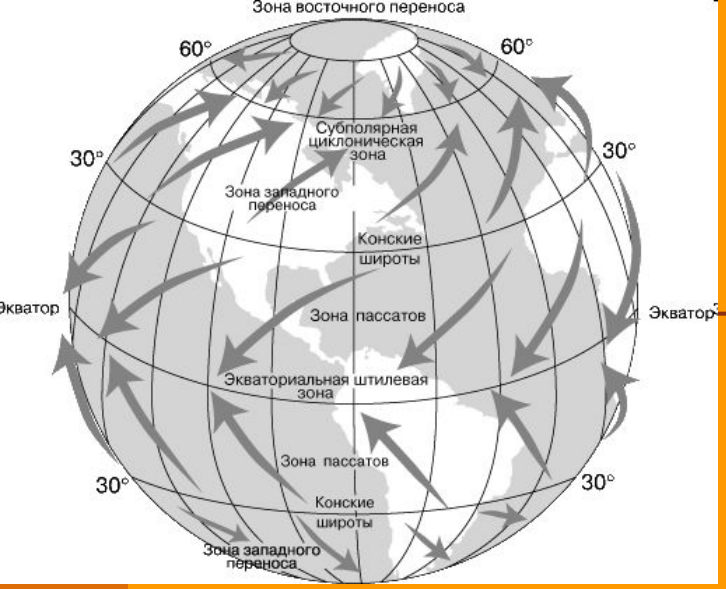
Цель лекции:

- Сформировать у обучаемых систему знаний о сущности методов корреляционного и регрессионного анализа, об их роли в исследовании социально-правовых процессов.



ПЛАН ЛЕКЦИИ

1. Виды зависимостей между величинами
2. Корреляционный анализ
3. Регрессионный анализ
4. Доверительный интервал



1. Виды зависимостей между величинами

Все количественные характеристики объектов в математике обычно называют математическими величинами или просто величинами.

Величины могут быть *постоянными* (constant) и *переменными* (variable).

Величины могут быть *зависимыми* и *независимыми*.

Также величины разделяют на *детерминированные* и *случайные*.

Существует два вида зависимостей:

- *функциональная;*
- *стохастическая* (вероятностная, статистическая; от греч. *stochastikos* – умеющий угадывать, предполагать, строить предположение).

Определение

Зависимость между двумя величинами называется функциональной, если каждому значению одной величины соответствует единственное значение другой величины.

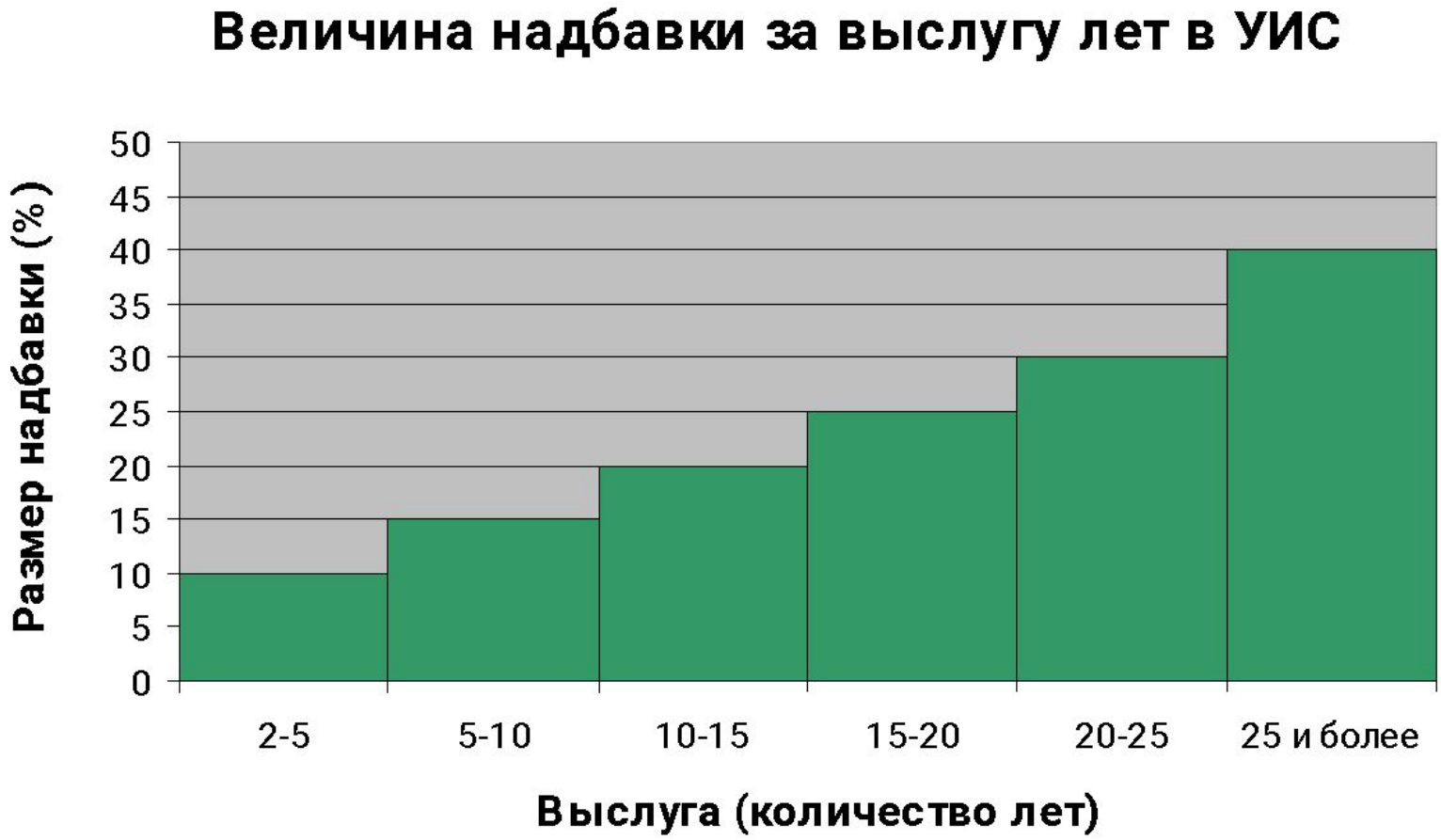
Пример

- Рассмотрим две величины x − выслуга сотрудника УИС (количество лет), y − размер надбавки от оклада по должности (%). Известно, что y зависит от x функционально (т. е. y является функцией от x) и эту зависимость можно представить различными способами.

1. В виде таблицы.

x	2-5	5-10	10-15	15-20	20-25	25 и более
y	10	15	20	25	30	40

2. Графически.



3. Аналитически.

$$y = f(x) = \begin{cases} 0 & при \quad 0 \le x < 2; \\ 10 & при \quad 2 \le x < 5; \\ 15 & при \quad 5 \le x < 10; \\ 20 & при \quad 10 \le x < 15; \\ 25 & при \quad 15 \le x < 20; \\ 30 & при \quad 20 \le x < 25; \\ 40 & при \quad x \ge 25. \end{cases}$$

Определение

Зависимость между двумя величинами называется стохастической, если каждому значению одной величины соответствует множество значений другой величины.

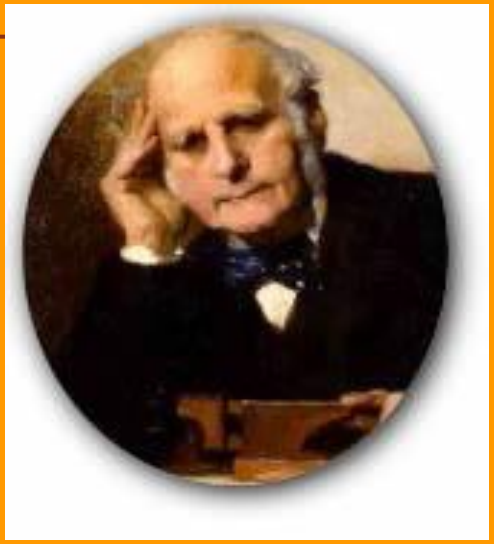
Модель стохастической связи

$$Y=f(X)+\varepsilon,$$

где **Y** − значение результативного признака, **f(X)** − часть результативного признака, сформированного под воздействием факторного признака **X**, **ε** − часть результативного признака, возникшая вследствие влияния других неучтенных факторов.

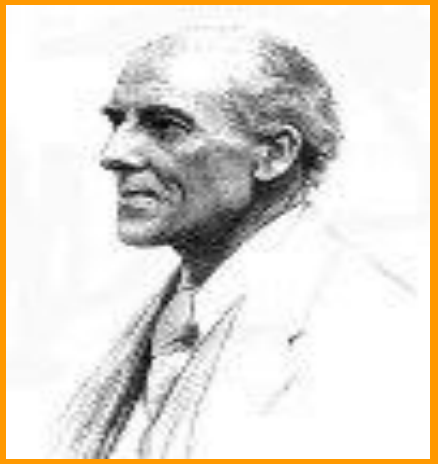
2. Корреляционный анализ

Понятия *корреляция* и *регрессия* появились в середине XIX в. благодаря работам английских статистиков **Ф. Гальтона** и **К. Пирсона.**



Фрэнсис Гальтон (1822-1911)

Первый термин произошёл от латинского *correlation* (соотношение, взаимосвязь), второй также от латинского *regressio* (движение назад).



Карл Пирсон (1857-1936)

Определение

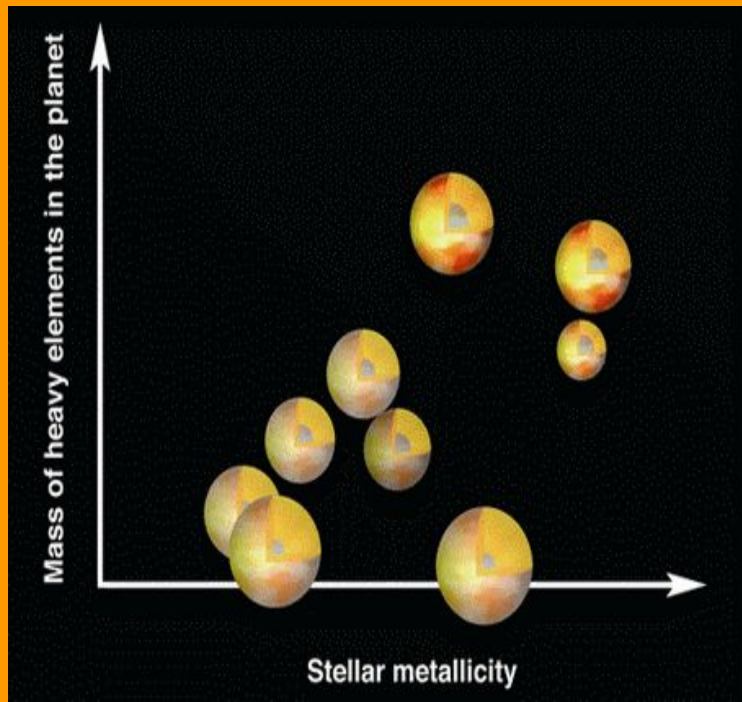
Корреляционная зависимость (или просто корреляция) – это статистическая зависимость между случайными величинами, при которой каждому значению одной величины соответствует определённое значение условного математического ожидания (среднего значения) другой.

Виды корреляции

- ***Парная корреляция*** – связь между двумя признаками.
- ***Частная корреляция*** – зависимость между результативным и одним факторным признаками при фиксированном значении других факторных признаков.
- ***Множественная корреляция*** – зависимость результативного признака и двух или более факторных признаков.

Основные задачи корреляционного анализа

- определение существования и тесноты корреляционной связи;
- установление достоверности суждения о наличии этой связи.



Коэффициент корреляции

$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \cdot \sqrt{n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2}}.$$

Основные свойства коэффициента корреляции (при достаточно большом объёме выборки n).

1. $-1 \le r \le 1$. При $0 < |r| \le 0{,}3$ связь практически отсутствует; при $0{,}3 < |r| \le 0{,}5$ слабая; при $0{,}5 < |r| \le 0{,}7$ умеренная; при $0{,}7 < |r| \le 0{,}9$ существенная; при $0{,}9 < |r| \le 1$ сильная.

При $|r| = 1$ корреляционная связь представляет собой линейную функциональную зависимость

вида $\bar{y}_x = ax + b$.

2. Знак коэффициента корреляции указывает направление связи: при $0 < r \leq 1$ связь прямая, а при $-1 \leq r < 0$ связь обратная.

3. Если величины X и Y статистически независимы, то $r_{xy} = 0$. Обратное утверждение неверно. Равенство $r_{xy} = 0$ говорит только об отсутствии линейной корреляционной зависимости, но не вообще об отсутствии корреляционной, а тем более статистической, зависимости.

Использование MS Excel

Для вычисления коэффициента корреляции используется стандартная функция

=КОРРЕЛ(Массив 1; Массив 2).

Для вычисления критического значения распределения Стьюдента используется функция

=СТЬЮДРАСПОБР(p; n-2).

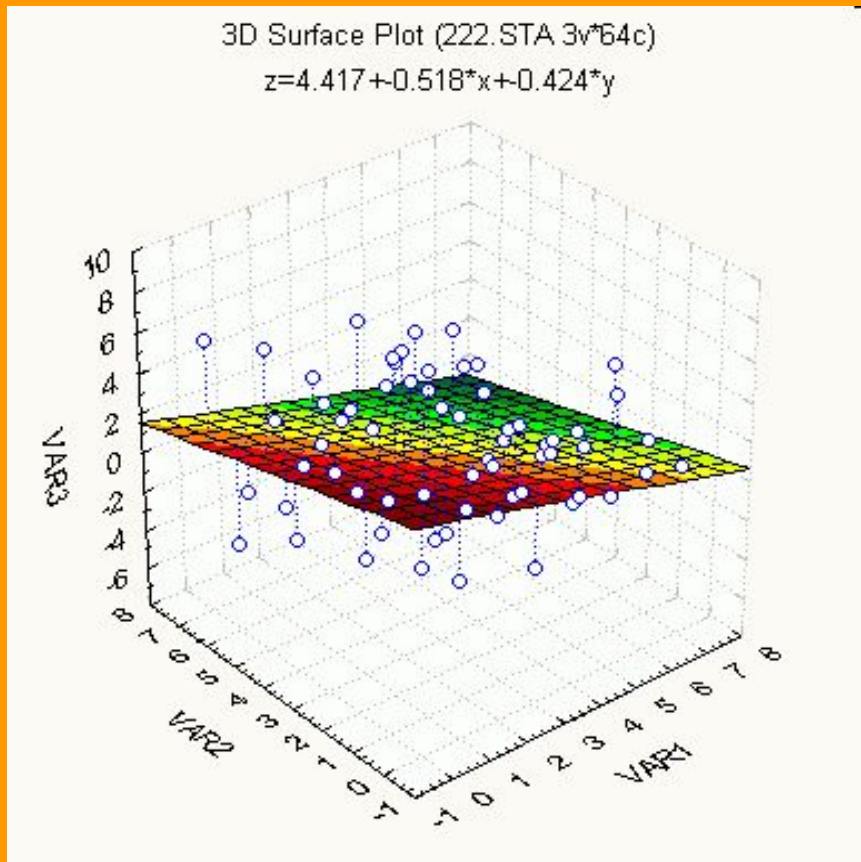
Результаты расчёта

J	K	L	M

	Корреляция		
	r_{TX}	r_{TY}	r_{TZ}
	-0,12	0,84	-0,82
$T_{набл.}$	-0,51	6,62	-6,03
$T_{крит.}$	2,10	2,10	2,10

3. Регрессионный анализ

Определение. *Регрессионный анализ – это совокупность методов, с помощью которых устанавливают форму стохастической зависимости между величинами.*



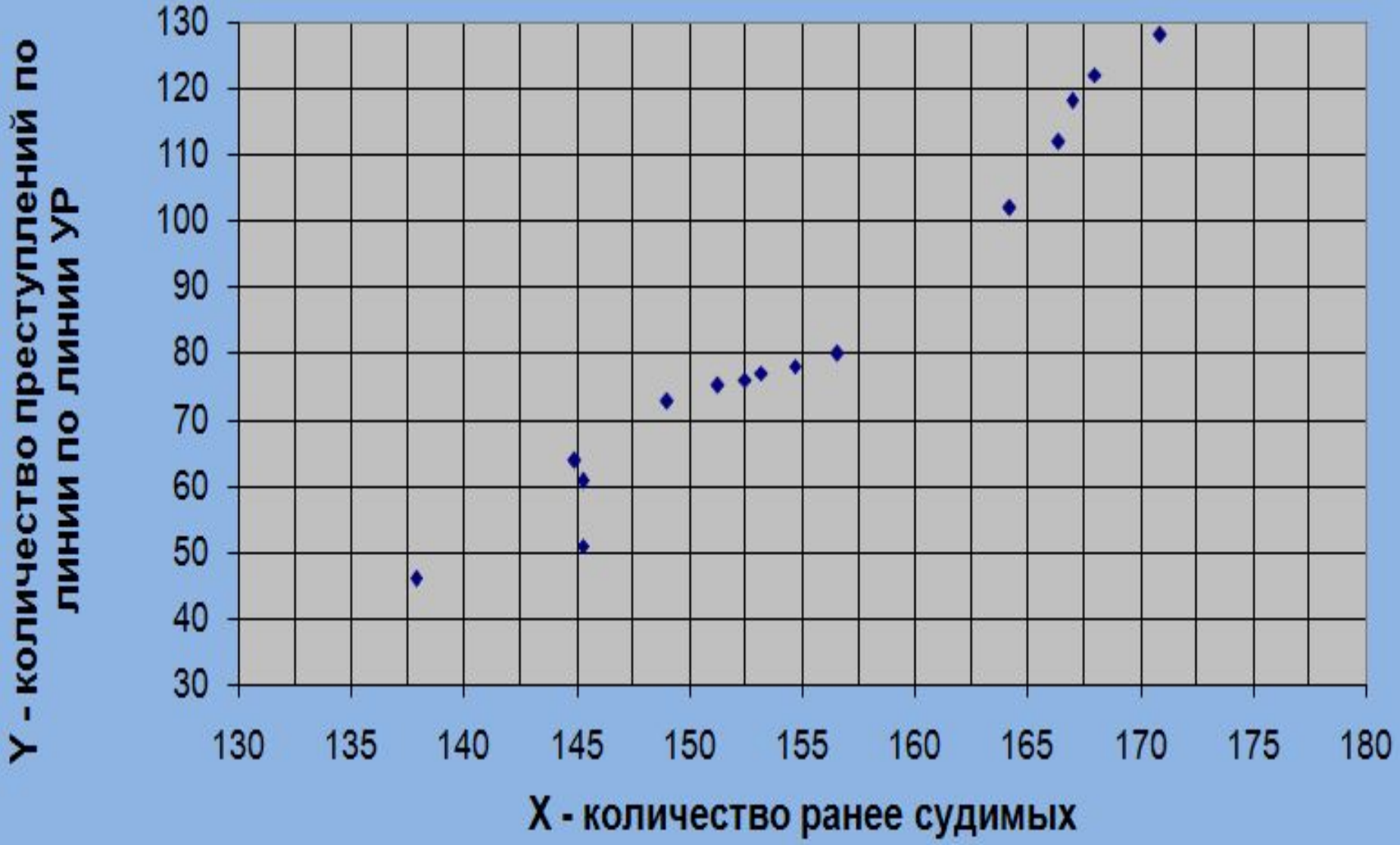
Пример

1. На рабочем листе в диапазон ячеек **B3:B17** введём значения величины **X**, а в диапазон ячеек **C3:C17** − величины **Y**.
2. Вычислим выборочный коэффициент корреляции R_{XY} с помощью стандартной функции **=КОРРЕЛ(B3:B17;C3:C17)**. В результате получаем R_{XY}**=0,98**. Так как коэффициент корреляции близок к **1**, то между признаками наблюдается тесная связь, близкая к линейной.

Алгоритм решения

3. Для графического определения вида формы связи построим корреляционное поле, используя *стандартную точечную диаграмму*. Расположение точек на корреляционном поле подтверждает сделанную выше гипотезу о линейной зависимости между **X** и **Y**. Тогда функция регрессии имеет вид $\mathbf{y_x = a + bx}$.

Поле корреляции



Алгоритм решения

4. Найдём значения параметров регрессии. Для этого используем инструмент **Сервис→Анализ данных→Регрессия**. В появившемся диалоговом окне **«Регрессия»** указываем диапазоны входных данных для **X** и **Y**, а также в выходном интервале указываем ссылку на левую верхнюю ячейку выходного диапазона для вывода итогов. Затем кнопка **OK.**

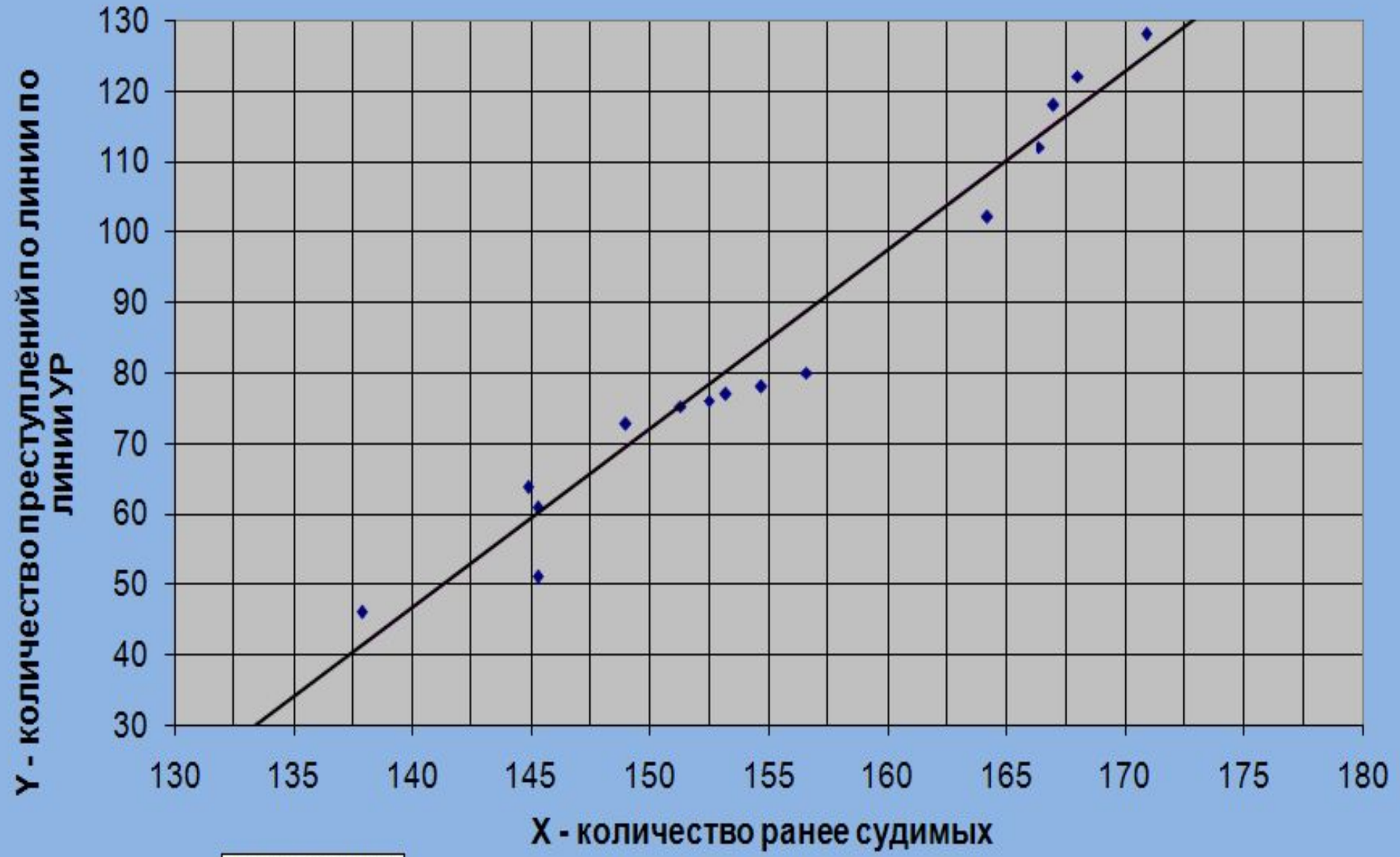
Алгоритм решения

ВЫВОД ИТОГОВ	
Регрессионная статистика	
Множественный R	0,982574732
R-квадрат	0,965453104
Нормированный R-квадрат	0,962574196
Стандартная ошибка	5,21107256
Наблюдения	14
Дисперсионный анализ	
	df
Регрессия	1
Остаток	12
Итого	13
	Коэффициенты
Y-пересечение	-309,0530771
152,5	2,535082117

Алгоритм решения

5. Среди появившихся итогов находим коэффициенты регрессии **b=2,54** и **a=-309**. Тогда уравнение регрессии y_x**=-309+2,54x**.
6. На корреляционном поле построим прямую **y=-309+2,54x.** Видно, что выборочные значения располагаются достаточно близко от этой прямой. Следовательно, полученная модель в некоторых случаях может быть использована для прогнозирования

Прямая регрессии



Проверка значимости модели регрессии с помощью критерия Фишера (Алгоритм)

1. Вычисляют факторную дисперсию.

$$S^2_{факт} = \frac{\sum (\hat{y}_x - \bar{y})^2}{m};$$

Проверка значимости модели регрессии с помощью критерия Фишера

2. Вычисляют остаточную дисперсию.

$$S^2_{ост} = \frac{\sum (\hat{y}_x - y)^2}{n - m - 1}.$$

Проверка значимости модели регрессии с помощью критерия Фишера

3. Вычисляют наблюдаемое значение критерия Фишера.

$$F_{факт} = \frac{S^2_{факт}}{S^2_{ост}}.$$

Проверка значимости модели регрессии с помощью критерия Фишера

4. Задают уровень значимости **α:**

$$0{,}01 < \alpha < 0{,}1.$$

Проверка значимости модели регрессии с помощью критерия Фишера

5. С помощью стандартной функции MS Excel находят теоретическое значение критерия Фишера $F_{теор}$.

$$=F.ОБР(1-\alpha;m;n-m-1)$$

Проверка значимости модели регрессии с помощью критерия Фишера

6. Делают вывод.

Если $F_{факт} > F_{теор}$, то модель регрессии признаётся статистически значимой в целом, и может быть использована для прогнозирования.

4. Доверительный интервал

Доверительным интервалом называется интервал, который с заданной надёжностью (или доверительной вероятностью) β покрывает оцениваемый параметр.

В общем виде доверительный интервал имеет вид:

$$x_в - \Delta < x_{ген} < x_в - \Delta.$$

Доверительный интервал для генеральной средней (математического ожидания)

Если среднее квадратическое отклонение **σ** исследуемой случайной величины неизвестно, то для оценки математического ожидания ***a*** случайной величины служит доверительный интервал:

$$\bar{x}_в - \Delta < \bar{x}_{ген} < \bar{x}_в - \Delta.$$

Алгоритм нахождения доверительного интервала для среднего значения

1. Для вычисления выборочного среднего значения используется стандартная функция

=СРЗНАЧ(Массив)

2. Для вычисления выборочного среднего квадратического отклонения S_x используют функцию

=СТАНДОТКЛОН.В(Массив)

Использование MS Excel

3. Задают доверительную вероятность β

$$0{,}9 < \beta < 0{,}99.$$

4. Для вычисления допустимой предельной ошибки Δ используется функция

$$=ДОВЕРИТ.СТЬЮДЕНТ(1-\beta;S_x;n)$$


Задавайте вопросы