Статистические гипотезы

1. Основные определения

Основные задачи МС разделяют на две категории:

- оценивание неизвестных параметров распределения (получение по выборке оценок, наилучших в том или ином смысле);

- проверка статистических гипотез (по выборке принять или отвергнуть некоторое предположение о распределении генеральной совокупности, из которой извлечена выборка).

Статистической гипотезой называют любое предположение о виде (***непараметрическая гипотеза***) или параметрах (***параметрическая гипотеза***) неизвестного распределения.

Статистическую гипотезу называют ***простой***, если она полностью определяет функцию распределения. В противном случае гипотезу называют ***сложной***.

Проверяемую гипотезу называют ***нулевой*** (***основной***) и обозначают H_0. Наряду с нулевой рассматривают ***альтернативную*** (***конкурирующую***) гипотезу $H_1 \ \left(\bar{H}\right)$.

$$H_0 : \ \theta = \theta_0, \quad H_1 : \ \theta \neq \theta_0.$$

$$H_0 : \ \theta = \theta_0, \quad H_1 : \ \theta > \theta_0.$$

Правило, которое позволяет по выборке принять или отвергнуть проверяемую гипотезу, называют ***критерием проверки статистической гипотезы*** (***статистическим критерием***)

Статистическими методами нельзя доказать правильность гипотезы.

Критерий проверки статистической гипотезы позволяет отбросить гипотезу как неправильную, но не позволяет доказать, что она верна, т.е. статистические критерии указывают лишь на отсутствие опровержения со стороны имеющихся экспериментальных данных.

Если по результатам проверки статистическая гипотеза принимается, то говорят, что она ***согласуется с выборочными данными*** или, что она ***не противоречит результатам наблюдений***.

Статистический критерий обычно основывается на некоторой статистике $\overline{\theta}_n$, для которой известно ее точное или приближенное распределение.

Множество всех возможных значений этой статистики разбивают на два непересекающихся множества: ***область принятия нулевой гипотезы*** и ***область отклонения нулевой гипотезы*** (***критическая область***).

Проверяемая гипотеза H_0	H_0 принимается -	H_0 отвергается -
объективно верна	правильное решение	***ошибка 1-го рода***
объективно неверна	***ошибка 2-го рода***	правильное решение

Вероятность ошибки первого рода, т.е. вероятность отвергнуть нулевую гипотезу, когда она верна, называют ***уровнем значимости статистического критерия*** и обозначают α :

$$P\left(H_0 \text{ отвергается} \,\middle|\, H_0 \text{ верна}\right) = \alpha.$$

Вероятность ошибки второго рода, т.е. вероятность ошибочно принять нулевую гипотезу, обозначают β:

$$P\left(H_0 \text{ принимается} \,\middle|\, H_0 \text{ не верна}\right) = \beta.$$

Пользуясь терминологией статистического контроля качества продукции, можно сказать, что $\alpha -$ это риск поставщика (забраковка партии, удовлетворяющей стандарту), а $\beta -$ риск потребителя (принятие партии не удовлетворяющей стандарту)

Мощностью критерия называют вероятность отклонить проверяемую гипотезу H_0, когда она неверна.

$$P\left(H_0 \text{ отвергается} \,\middle|\, H_0 \text{ не верна}\right) = \beta.$$

При построении статистических гипотез требование, чтобы ошибки обоих родов были бы минимальны противоречиво. Невозможно одновременно уменьшить обе ошибки.

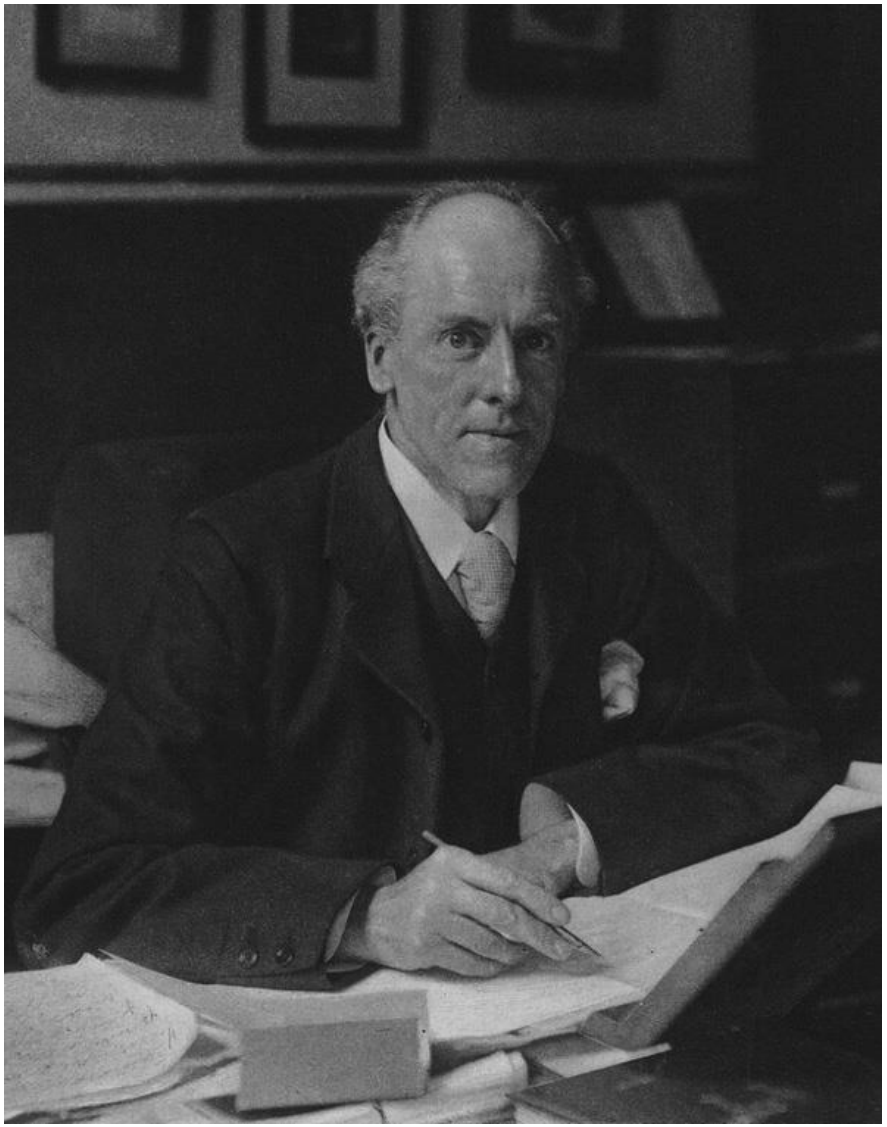
На практике поступают так: задают уровень значимостиα (как правило, равный 0,05, 0,01 или 0,1) а, затем, выбирают статистический критерий так, чтобы ошибка второго рода была наименьшей.

Статистические критерии, с помощью которых проверяют гипотезы о значениях параметров распределения или о соотношениях между ними, в предположении, что тип распределения известен, называют ***критериями значимости*** (***параметрическими критериями***)

Статистические критерии, с помощью которых проверяют гипотезы о виде распределения, называют ***критериями согласия*** (***непараметрическими критериями***)

Наиболее известными являются критериями согласия являются критерий χ^2 Пирсона и критерий Колмогорова.

Карл Пирсон
27.03.1857-27.04.1936



Английский математик и биолог, основатель английской школы биометрики.

Внес существенный вклад в распространение методов статистического анализа в биологии и психологии.

Основные идеи Пирсона были опубликованы в серии из 19 статей под общим названием «Математический вклад в теорию эволюции».

Пирсон считается одним из основоположников современной статистики.

2. Критерий согласия χ^2 Пирсона

Пусть имеется выборка объема n и сгруппированный статистический ряд, в котором k групп (например, в случае непрерывной СВ это будет k интервалов).

Группы выбирают так, чтобы охватить весь диапазон значений предполагаемой СВ. Если диапазон значений СВ неограничен, то крайние интервалы должны быть расширены до $-\infty$ и $+\infty$ соответственно.

В каждый интервал должно входить не менее 5 наблюдений. Группы с малым числом наблюдений объединят с соседними.

Проверяемая гипотеза представляет собой предположение о виде распределения наблюдаемой СВ и является простой (конкретно указывает предполагаемое распределение)

H_0: функция распределения наблюдаемой СВ совпадает с $F(x)$.

H_1: функция распределения наблюдаемой СВ не совпадает с $F(x)$.

Критерий согласия Пирсона основан на сравнении эмпирических и теоретических частот попадания СВ в рассматриваемые группы (интервалы)

n_i – эмпирическая частота наблюдения значений из интервала $\left[x_{i-1}; x_i\right)$

$$n_i' = np_i = nP\left(\xi \in \left[x_{i-1}; x_i\right)\right) = n\left(F\left(x_i\right) - F\left(x_{i-1}\right)\right)$$

n_i' – теоретическое значение соответствующей частоты

По данным выборки вычисляют статистику

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'}$$

Для вычисления статистики $\chi^2_{набл}$ нужно знать сгруппированный статистический ряд и теоретическую функцию распределения $F(x)$.

При этом $F(x)$ может зависеть от одного или нескольких параметров. Пусть r – число неизвестных параметров теоретического распределения. В этом случае вместо значений параметров используют их оценки.

Теорема. Если теоретическая функция распределения зависит от r параметров и оценки этих параметров обладают свойствами асимптотической нормальности и асимптотической эффективности, то, независимо от вида теоретической функции распределения $F(x)$ в пределе (при $n \to \infty$) статистика $\chi^2_{набл}$ имеет распределение χ^2 с числом степеней свободы $k - r - 1$, где k – число интервалов группировки, r – количество параметров теоретической функции распределения, оцениваемых по данной выборке.

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$$

Критерий согласия χ^2 Пирсона заключается в следующем:

если $\chi^2_{набл} < \chi^2_{\alpha, k-r-1}$, где $\chi^2_{\alpha, k-r-1}$ определяют по таблице критических значений распределения χ^2, то гипотеза H_0 принимается (признается непротиворечащей экспериментальным данным; нет оснований отвергнуть гипотезу H_0) на уровне значимости α;

если $\chi^2_{набл} \geq \chi^2_{\alpha, k-r-1}$, то гипотеза H_0 отвергается (не согласуется с данными эксперимента).

Основное достоинство критерия согласия χ^2 Пирсона – его универсальность, т.е. применимость для любого закона распределения, в том числе с неизвестными параметрами

Основное недостаток – необходимость большого объема выборки (не менее 60-100 наблюдений) и произвольность группировки, влияющая на величину $\chi^2_{набл}$.

Пример. Дано интервальное распределение

Интервалы	n_i
$0-36$	44
$36-72$	24
$72-108$	16
$108-144$	9
$144-180$	2
$180-216$	5
$216-252$	4

После обработки выборки выдвигают гипотезу о показательном распределении с параметром

$$\lambda = \frac{1}{\overline{x}_в} = 0{,}015$$

При $\alpha = 0{,}05$ по критерию Пирсона подтвердите или отвергните выдвинутую гипотезу.

$$\lambda = 0{,}015$$

Находим теоретические (выравнивающие) частоты

$$n_i' = n \cdot P_i = n \cdot P\left(a_i < X_i < a_{i+1}\right) = 104 \cdot \left(e^{-\lambda a_i} - e^{-\lambda a_{i+1}}\right)$$

$$P_1 = e^{-0{,}015 \cdot 0} - e^{-0{,}015 \cdot 36} = 0{,}4173, \ldots$$

Интервалы	n_i	P_i	$n_i' = n \cdot P_i$	$(n_i - n_i')^2$	$(n_i - n_i')^2 / n'$
$0-36$	44	0,4173	43,40	0,36	0,08
$36-72$	24	0,2431	25,28	1,6384	0,065
$72-108$	16	0,1417	14,74	1,5876	0,108
$108-144$	9 (11)	0,0826	8,59 (13,59)	6,7081	0,494
$144-180$	2	0,0481	5,00		
$180-216$	5 (9)	0,0280	2,91 (4,62)	19,1844	4,152
$216-252$	4	0,0164	1,71		
	104	0,9772	101,63		4,827

Сравниваем $\chi^2_{набл} = \sum \frac{(n_i - n_i')^2}{n_i'}$ и

$$\chi^2_{крит}(\alpha, k = l-2) = \chi^2_{крит}(0{,}05, 5-2) = \chi^2_{крит}(0{,}05, 3)$$

$$\chi^2_{набл} = \sum \frac{\left(n_i - n'_i\right)^2}{n'_i} = 4{,}827 \qquad \chi^2_{крит}\left(0{,}05;\, 3\right) = 7{,}815$$

$$\chi^2_{набл} < \chi^2_{крит} \quad \Rightarrow \quad \text{Гипотеза не отвергается}$$

Элементы регрессионного и корреляционного анализа

1. Основные определения

Пусть на основании экспериментальных данных (по выборке объема *n* связанных пар наблюдений $(x_i; y_i))$ изучают связь между двумя величинами.

Две СВ могут быть

1) независимыми;
2) связаны функциональной зависимостью;
3) связаны статистической зависимостью.

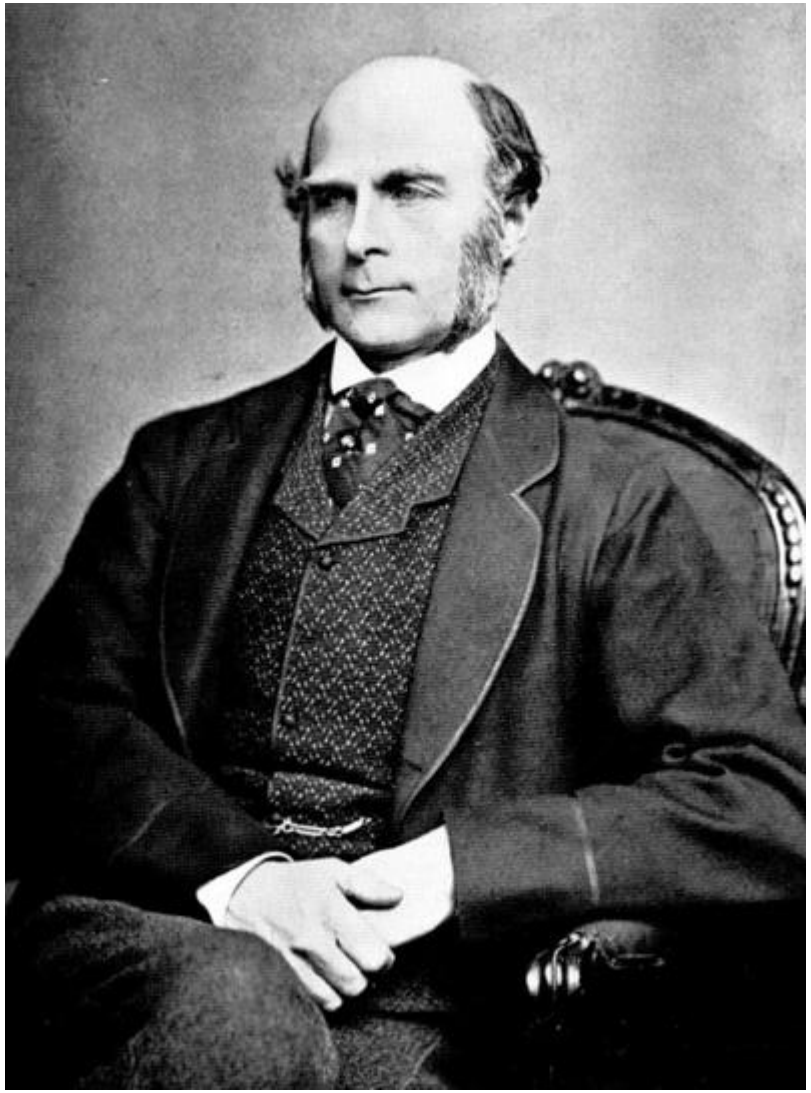
Статистической (стохастической, вероятностной) называют такую зависимость между СВ, при которой каждому значению одной из них соответствует множество возможных значений другой, и изменение значения одной величины влечет изменение ***распределения*** другой, в частности, может изменяться ***среднее значение*** другой.

Регрессионная зависимость между СВ – это функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Основным методом исследования статистических зависимостей является ***корреляционно-регрессионный анализ***.

Понятие корреляции и регрессии появилось во второй половине XIXв. В работах К. Пирсона и Ф. Гамильтона. Слово «корреляция» от лат correlatio – «соотношение, взаимосвязь»; «регрессия» - от лат. regressio – «движение назад». Термин «регрессия» ввел Ф.Гамильтон, который изучал зависимость между ростом отцом и сыновей, обнаружил явление «регрессии к среднему»: у детей родившихся у очень высоких родителей, рост имел тенденцию быть ближе к средней величине.

Фрэнсис Гамильтон
16.02.1822-16.01.1911



Английский исследователь, географ, антрополог, психолог, статистик, основатель дифференциальной психологии и психометрики, основоположник учения евгеники, которое было призвано бороться с явлениями вырождения в человеческом генофонде.

Основными задачами корреляционного анализа являются выявление связи между наблюдаемыми СВ и оценка тесноты этой связи.

Основными задачами регрессионного анализа является установление формы зависимости между наблюдаемыми величинами и определение по экспериментальным данным уравнения зависимости, которое называют ***выборочным (эмпирическим) уравнением регрессии***, а также прогнозирование с помощью уравнения регрессии среднего значения зависимой переменной при заданном значении независимой переменной.

2. Выборочный коэффициент корреляции

Пусть на основании экспериментальных данных изучают связь между двумя величинами.

Количественной мерой ***линейной связи*** между двумя наблюдаемыми величинами служит выборочный коэффициент корреляции.

$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

Свойства выборочного коэффициента корреляции

1) $-1 \le r_{xy} \le 1.$

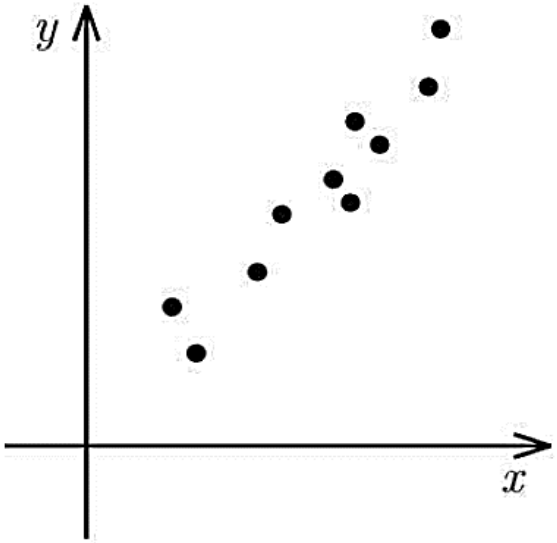
2) Если наблюдаемые величины независимыми, то $r_{xy} \approx 0.$

3) Если $\left|r_{xy}\right| = 1$ (или близок к 1), то наблюдаемые величины связаны линейной зависимостью.

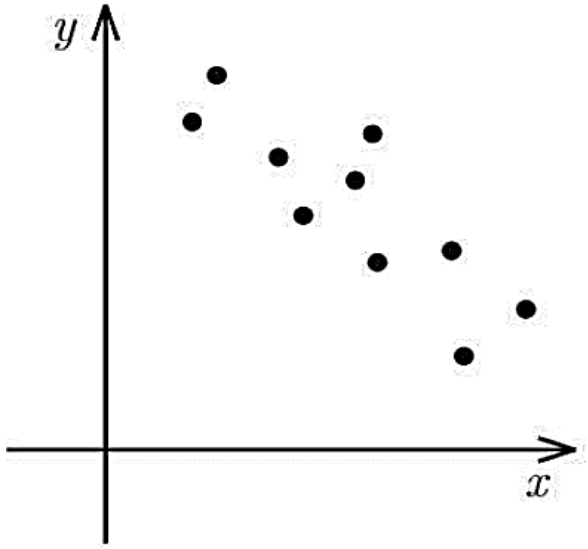
$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

4) Если $r_{xy} > 0$, то с ростом значений одной величины, значения другой также возрастают;

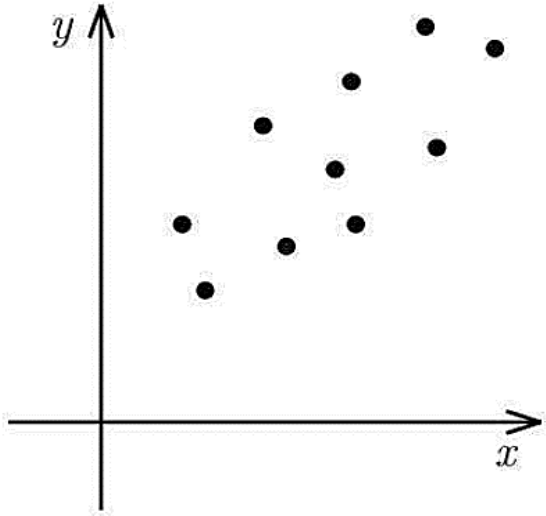
если $r_{xy} < 0$, то с ростом значений одной величины, значения другой убывают;



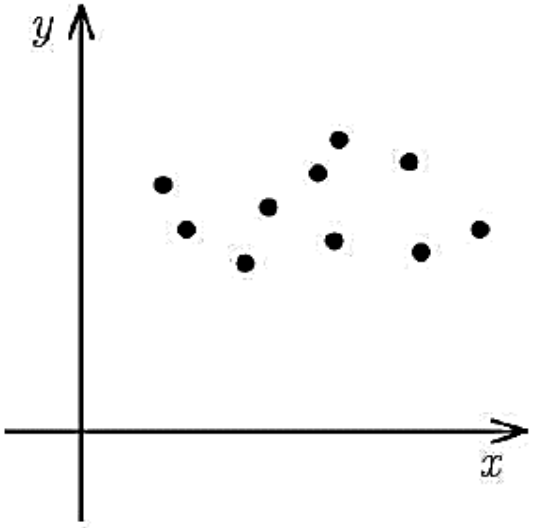
$r_{x,y} \approx 0,95$



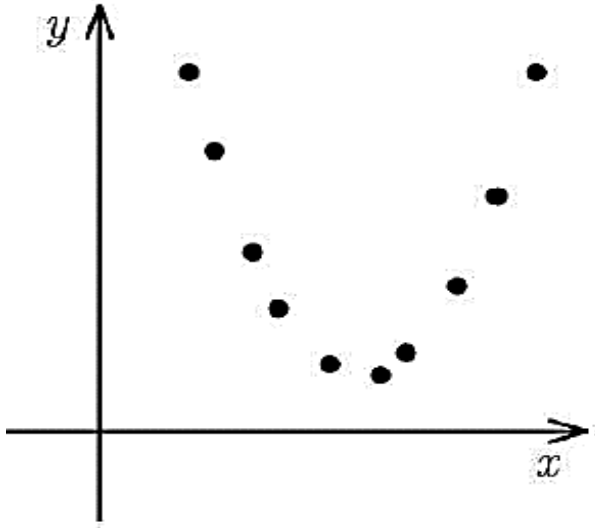
$r_{x,y} \approx -0,85$



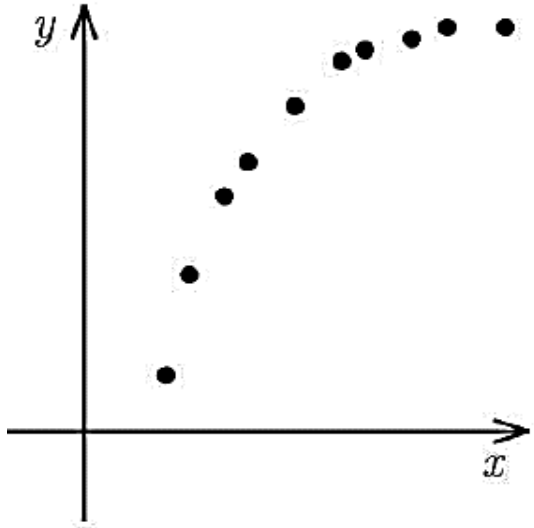
$r_{x,y} \approx 0,7$



$r_{x,y} \approx 0$



$r_{x,y} \approx 0$



$r_{x,y} \approx 0,9$

3. Проверка значимости коэффициента корреляции

Проверка значимости коэффициента корреляции – это проверка гипотезы о том, что коэффициент корреляции значимо отличается от нуля.

Т.к. выборка произведена случайно, то нельзя утверждать, что если $r_{xy} \neq 0$, то $r_{ген.} \neq 0$.

Если выборка из нормального распределения, то проверка производится по критерию Стьюдента

$$t_{набл} = \left|r_{xy}\right| \sqrt{\frac{n-2}{1-r_{xy}^2}} > t_{крит}\left(\alpha; n-2\right)$$

то при заданном уровне значимости α коэффициент корреляции считают значимо отличающимся от нуля, а, следовательно, связь между величинами признается статистически значимой.

Коэффициент корреляции является мерой именно линейной зависимости.

Уи́льям Си́ли Го́ссет

13.06.1876-16.10.1937

Британский химик и статистик, работавший на пивоваренном заводе «Гиннесс» (Arthur Guinness Son & Co). Один из основоположников теории статистических оценок и проверки гипотез. Разработал математическое обоснование «закона ошибок» для малых статистических выборок.

Госсет более известен под своим псевдонимом Стьюдент (Student), поскольку по условиям контракта с корпорацией «Гиннесс» не имел права открыто публиковать результаты своих Первым, кто понял значение работ Госсета по оценке параметров малой выборки, был английский статистик Р. Э. Фишер, считавший, что Госсет совершил «логическую революцию» в математической статистике.

Пусть имеется выборка объема n наблюдений над двумя величинами X и $У$, и принята гипотеза о линейной зависимости между $У$ и X.

Для определения коэффициентов линейного эмпирического уравнения регрессии

$$\bar{y}_x = a + bx$$

используют метод наименьших квадратов.

$$\begin{cases} a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i; \\ an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i. \end{cases}$$

Пример. При контроле качества пищевых продуктов для определения концентрации определенных веществ находят эмпирическое линейное уравнение зависимости оптической плотности градуировочного раствора от концентрации. Имеются данные для определения концентрации фосфора в мясных изделиях.

Концентрация р-ра, мг/кг	0,02	0,04	0,06	0,08	0,10
Оптическая плотность р-ра	0,035	0,070	0,150	0,140	0,175

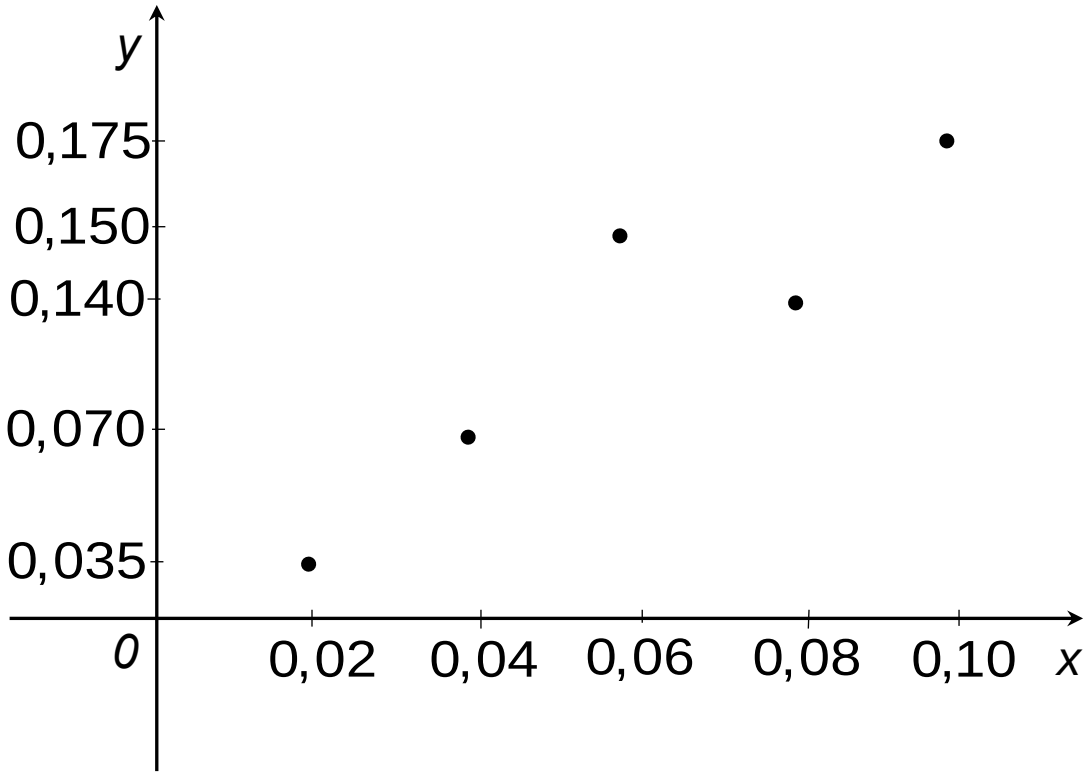
По имеющимся данным

1) построить корреляционное поле;

2) найти выборочный коэффициент корреляции и проверить его значимость при $\alpha = 0{,}05$;

3) определить коэффициенты линейного эмпирического уравнения регрессии, построить прямую на корреляционном поле.

Концентрация р-ра, мг/кг	0,02	0,04	0,06	0,08	0,10
Оптическая плотность р-ра	0,035	0,070	0,150	0,140	0,175



По виду корреляционного поля можно предположить, что коэффициент корреляции положителен и значимо отличается от нуля

$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$\sigma_x = \sqrt{D_в}, \qquad D_в(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2,$$

$$\sigma_y = \sqrt{D_в}, \qquad D_в(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2$$

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	0,02	0,035	0,0007	0,0004	0,001225
2	0,04	0,070	0,0028	0,0016	0,004900
3	0,06	0,150	0,0090	0,0036	0,022500
4	0,08	0,140	0,0112	0,0064	0,019600
5	0,10	0,175	0,0175	0,0100	0,030625
$\sum$	0,30	0,570	0,0412	0,0220	0,078850

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{0,30}{5} = 0,06$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{y} = \frac{0{,}570}{5} = 0{,}114$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \qquad \overline{xy} = \frac{0{,}0412}{5} = 0{,}00824$$

$$D_x = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\overline{x}\right)^2 \qquad D_x = \frac{0{,}022}{5} - 0{,}06^2 = 0{,}0008$$

$$D_y = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\overline{y}\right)^2 \qquad D_y = \frac{0{,}07885}{5} - 0{,}114^2 = 0{,}002774$$

$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y},$$

$$r_{xy} = \frac{0,00824 - 0,06 \cdot 0,114}{\sqrt{0,0008 \cdot 0,002774}} \approx 0,94$$

Оценим значимость выборочного коэффициента корреляции $r_{xy} = 0,94$ для генеральной совокупности $(X, У)$ при заданном уровне значимости $\alpha = 0,05$.

Выдвигаем нулевую и альтернативную гипотезы:

$H_0: r_{ген} = 0$ (в генеральной совокупности нет линейной зависимости).

$H_1: r_{ген} \neq 0$ (в генеральной совокупности есть линейная зависимость между СВ X и $У$).

Для проверки воспользуемся критерием Стьюдента. Если

$$t_{набл} = \left| r_{xy} \right| \sqrt{\frac{n-2}{1-r_{xy}^2}} > t_{крит}\left(\alpha; n-2\right)$$

то при заданном уровне значимости α коэффициент корреляции считают значимо отличающимся от нуля, а, следовательно, связь между величинами признается статистически значимой.

$$t_{набл} = |r_{xy}| \sqrt{\frac{n-2}{1-r_{xy}^2}}$$

$$t_{набл} = 0,94 \cdot \sqrt{\frac{5-2}{1-0,94^2}} \approx 4,76$$

$$t_{крит}(\alpha; n-2)$$

По таблице «Критические точки распределения Стьюдента».

$$t_{крит}(0,05; 5-2) = 3,18$$

$$t_{набл} = |r_{xy}| \sqrt{\frac{n-2}{1-r_{xy}^2}} > t_{крит}(\alpha; n-2)$$

$$t_{набл} = 0,94 \cdot \sqrt{\frac{5-2}{1-0,94^2}} \approx 4,76$$

$$t_{крит}(0,05; 5-2) = 3,18$$

$$t_{набл} > t_{крит}$$

Значит, при уровне значимости $\alpha = 0,05$ коэффициент корреляции значимо отличается от нуля, а, следовательно, величины *X* и *У* связаны линейной зависимостью $\bar{y}_x = a + bx$.

Найдем коэффициенты уравнения $\bar{y}_x = a + bx$ МНК.

$$\begin{cases} a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i; \\ an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i. \end{cases}$$

$$\begin{cases} 0,3a + 0,022b = 0,0412 \\ 5a + 0,3b = 0,570 \end{cases} \qquad \Delta = \begin{vmatrix} 0,3 & 0,022 \\ 5 & 0,3 \end{vmatrix} = -0,02$$

$$\Delta_a = \begin{vmatrix} 0,0412 & 0,022 \\ 0,57 & 0,3 \end{vmatrix} = -0,00018 \qquad \Delta_b = \begin{vmatrix} 0,3 & 0,0412 \\ 5 & 0,57 \end{vmatrix} = -0,035$$

$$\bar{y}_x = a + bx$$

$$\Delta = \begin{vmatrix} 0,3 & 0,022 \\ 5 & 0,3 \end{vmatrix} = -0,02 \qquad \Delta_a = \begin{vmatrix} 0,0412 & 0,022 \\ 0,57 & 0,3 \end{vmatrix} = -0,00018$$

$$\Delta_b = \begin{vmatrix} 0,3 & 0,0412 \\ 5 & 0,57 \end{vmatrix} = -0,035$$

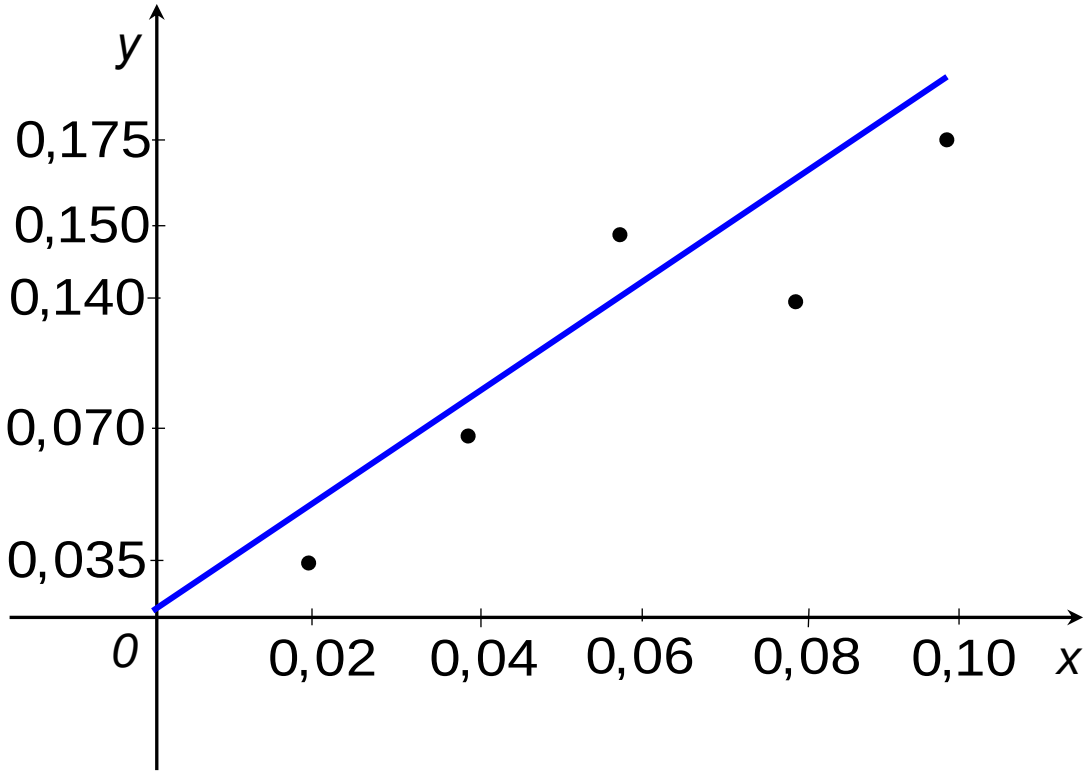
$$a = \frac{\Delta_a}{\Delta} = \frac{-0,00018}{-0,02} = 0,009$$

$$b = \frac{\Delta_b}{\Delta} = \frac{-0,035}{-0,02} = 1,75$$

$$\bar{y}_x = 0,009 + 1,75x$$

$$\bar{y}_x = 0{,}009 + 1{,}75x$$

x	0	0,1
$\bar{y}_x$	0,009	0,184



$$\begin{cases} a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i; \\ an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i. \end{cases}$$

$$\bar{y}_x - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \bar{x}_y - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Пример. Дана таблица распределения двух СВ *X* и *У*. Известно, что между *X* и *У* существует линейная корреляционная зависимость. Требуется:

1)составить уравнения прямых регрессии *У* на *X* и *X* на *У*;

2) построить на графике прямые регрессии и корреляционное поле;

3) оценить тесноту корреляционной зависимости и значимость выборочного коэффициента корреляции.

x \ y	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	m_x
120 – 140					3	4	7
140 – 160				2	5	2	9
160 – 180			3	6	3		12
180 – 200		5	9	8			22
200 – 220	1	4	2				7
220 – 240	3	2					5
240 – 260	3						3
m_y	7	11	14	16	11	6	65

x \ y	15	25	35	45	55	65	m_x	xm_x	x^2m_x
130					3	4	7	910	118300
150				2	5	2	9	1350	202500
170			3	6	3		12	2040	346800
190		5	9	8			22	4180	794200
210	1	4	2				7	1470	308700
230	3	2					5	1150	264500
250	3						3	750	187500
m_y	7	11	14	16	11	6	65	11850	2222500
ym_y	105	275	490	720	605	390	2585		
y^2m_y	1575	6875	17150	32400	33275	25350	116625		

$$\bar{y}_x - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \bar{x}_y - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{11850}{65} = 182{,}31$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \bar{y} = \frac{2585}{65} = 39{,}77$$

$$D_x = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 \qquad D_x = \frac{2222500}{65} - 182{,}31^2 = 955{,}37$$

$$D_y = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2 \qquad D_y = \frac{116625}{65} - 39{,}77^2 = 212{,}58$$

$$\sigma_x = \sqrt{955{,}37} = 30{,}91 \qquad \sigma_y = \sqrt{212{,}58} = 14{,}58$$

$$\overline{xy} = 130 \cdot (55 \cdot 3 + 65 \cdot 4) + 150 \cdot (45 \cdot 2 + 55 \cdot 5 + 65 \cdot 2) +$$
$$+170 \cdot (35 \cdot 3 + 45 \cdot 6 + 55 \cdot 3) + 190 \cdot (25 \cdot 5 + 35 \cdot 9 + 45 \cdot 8) +$$
$$+210 \cdot (15 \cdot 1 + 25 \cdot 4 + 35 \cdot 2) + 210 \cdot (15 \cdot 3 + 25 \cdot 2) +$$
$$+250 \cdot 3 = 445250$$

$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

$$r_{xy} = \frac{445250 - 182{,}31 \cdot 39{,}77}{30{,}91 \cdot 14{,}58} \approx -0{,}88$$

Близость $r_{xy} = -0{,}88$ к единице говорит о достаточно тесной линейной зависимости между *X* и *У.*

Т.к. с возрастанием значений одной СВ значения другой убывают, то $r_{xy} < 0$.

Уравнение прямой регрессии *У* на *X*

$$\bar{y}_x - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$\bar{y}_x - 39,77 = -0,88 \cdot \frac{14,58}{30,91}(x - 182,31)$$

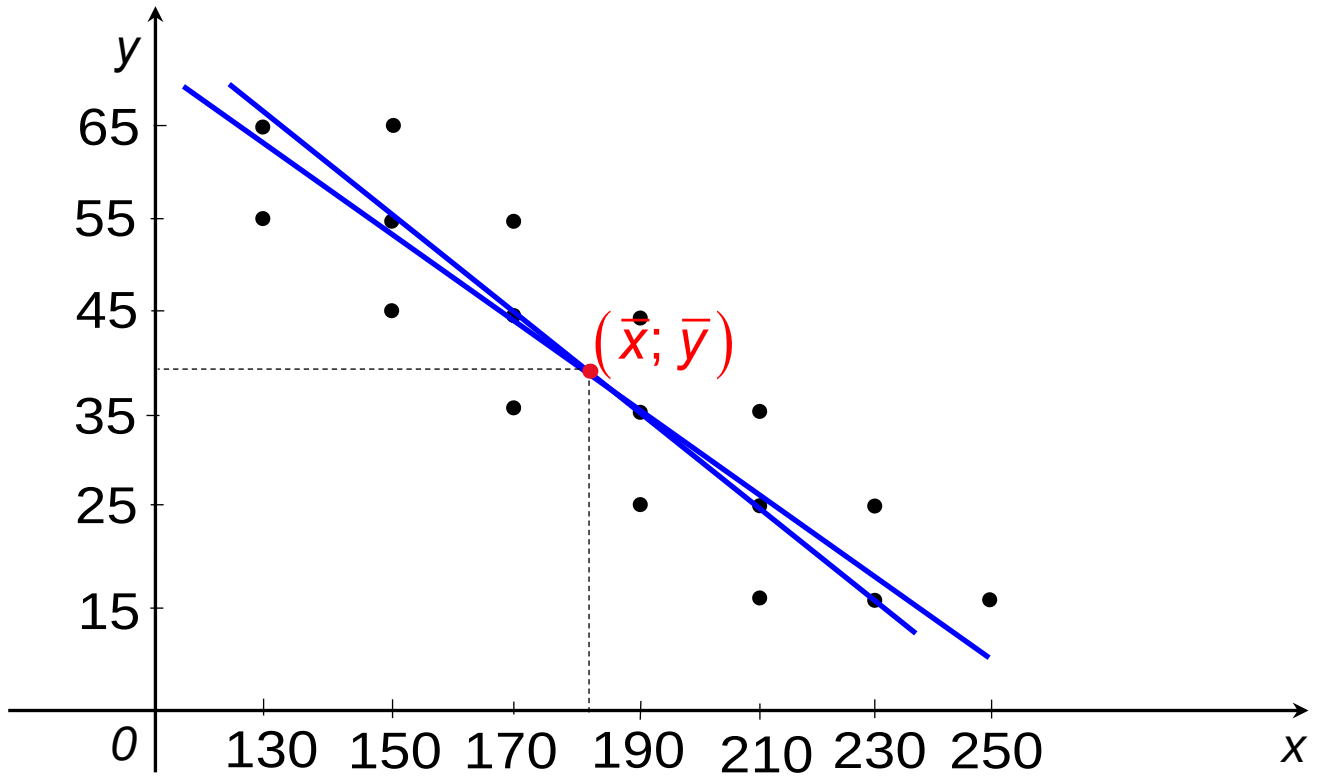
$$\bar{y}_x = -0,42x + 116,34$$

Уравнение прямой регрессии *X* на У

$$\bar{x}_y - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$\bar{x}_y - 182,31 = -0,88 \cdot \frac{30,91}{14,58}(y - 39,77)$$

$$\bar{x}_y = -1,88y + 257,08$$



x	130	250
$\bar{y}_x$	61,74	11,34

$$\bar{y}_x = -0,42x + 116,34$$

y	15	65
$\bar{x}_y$	228,88	134,88

$$\bar{x}_y = -1,88y + 257,08$$

Оценим значимость выборочного коэффициента корреляции $r_{xy} = -0,88$ для генеральной совокупности $(X, У)$ при заданном уровне значимости $\alpha = 0,05$.

Выдвигаем нулевую и альтернативную гипотезы:

$H_0: r_{ген} = 0$ (в генеральной совокупности нет линейной зависимости).

$H_1: r_{ген} \neq 0$ (в генеральной совокупности есть линейная зависимость между СВ X и $У$).

$$t_{набл} = |r_{xy}| \sqrt{\frac{n-2}{1-r_{xy}^2}}$$

$$t_{набл} = 0,88 \cdot \sqrt{\frac{65-2}{1-0,88^2}} = 14,71$$

$$t_{крит}(\alpha; n-2)$$

По таблице «Критические точки распределения Стьюдента».

$$t_{крит}(0,05; 65-2) = 2$$

$$t_{набл} > t_{крит}$$

Значит, при уровне значимости $\alpha = 0,05$ коэффициент корреляции значимо отличается от нуля, а, следовательно, величины *X* и *У* связаны линейной зависимостью

$$\bar{y}_x = -0,42x + 116,34$$

$$\bar{x}_y = -1,88y + 257,08$$