Критерий Стьюдента

Параметрический критерий, который используют для проверки статистических гипотез по выборкам, распределённым по нормальному закону Гаусса.

Используется:

1) Для определения значимости различия среднего арифметического, полученного для одной выборки, с фиксированным значением.

2) Для определения значимости различия средних арифметических двух выборок.

3) Для определения значимости корреляции двух случайных величин.

Проверка значимости различия среднего арифметического и фиксированного значения μ. Иногда случай μ=0 называют проверкой значимости среднего арифметического

H_0: статистически значимых различий между средним арифметическим и μ нет.

$$t_{эксп} = \frac{\bar{x} - \mu}{S_{\bar{x}}}$$ где $$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n \cdot (n-1)}}$$ ошибка среднего арифметического.

Число степеней свободы $\nu = n - 1$

Находим из таблицы критерия Стьюдента для $\nu = n - 1$ и заданного α, $t_{крит}$

$|t_{эксп}| \le |t_{крит}| \Rightarrow$ H_0 принимаем, сред.ар. не отличается от μ

$|t_{эксп}| > |t_{крит}| \Rightarrow$ H_0 отвергаем, ср.ар. значимо отличается от μ

Пример:

Измерена некоторая случайная величина *X*. Получены следующие результаты: 15, 18, 13, 14

По критерию Стьюдента проверить, значимо ли полученное значение среднего арифметического отличается от нуля. P_D=0,95.

$$\bar{x} = \frac{15+18+13+14}{4} = \frac{60}{4} = 15$$

$$S_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n \cdot (n-1)}} = \sqrt{\frac{(15-15)^2 + (18-15)^2 + (13-15)^2 + (14-15)^2}{4 \cdot (4-1)}} = \sqrt{\frac{0+9+4+1}{12}} =$$

$$\sqrt{\frac{14}{12}} = \sqrt{1,167} = 1,08$$

Таблица критерия Стьюдента

v	*α=0,05*	*α=0,01*
1	12,71	63,66
2	4,30	9,93
3	3,18	5,84
4	2,78	4,60
5	2,57	4,03
6	2,45	3,71
7	2,37	3,50
8	2,31	3,36
9	2,26	3,25
10	2,23	3,17
11	2,20	3,11
12	2,18	3,06
13	2,16	3,01

Находим из таблицы критерия Стьюдента для

$$\nu = 4 - 1 = 3 \qquad \alpha = 1 - 0,95 = 0,05 \qquad t_{крит} = 3,18$$

так как $\left| t_{эксп} \right| > t_{крит} \Rightarrow$ H_0 отвергаем

13,9 3,18

Вывод: $\bar{x}$ значимо (отличается от нуля).

Критери Стьюдента, второе применение. Сравнение средних значений двух выборок.

Имеем две выборочные совокупности:

$X\{x_1, x_2, \dots x_{n1}\}$ и **$Y\{y_1, y_2, \dots y_{n2}\}$**

n_1 –объём первой выборки, **n_2** – объём второй выборки.

H_0: M[X]=M[Y] или M[X]-M[Y]=0, т.е. обе выборки принадлежат одной генеральной совокупности, то есть различия между выборками не значимы.

Задаём уровень значимости α.

$$t_{эксп} = \frac{(\bar{x} - \bar{y})}{S_{\bar{x}-\bar{y}}} \qquad S_{\bar{x}-\bar{y}} = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2}{(n_1 + n_2 - 2)} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

ошибка разности средних арифметических

Число степеней свободы $\nu = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$

Если $n_1 = n_2 = n$ $\nu = 2n - 2$

Ошибку разности можно считать по формуле: $S_{\bar{x}-\bar{y}} = \sqrt{S\frac{2}{x} + S\frac{2}{y}}$

Находим из таблицы критерия Стьюдента для $\nu = n_1 + n_2 - 2$ и заданного α, $t_{крит}$

если $|t_{эксп}| \leq t_{крит} \Rightarrow$ H_0 принимаем

Вывод: обе выборки принадлежат одной генеральной совокупности, различия между выборками не значимы.

если $|t_{эксп}| > t_{крит} \Rightarrow$ H_0 отвергаем

Вывод: обе выборки не принадлежат одной генеральной совокупности, различия между выборками статистически значимы.

Пример:

Исследовалось влияние лекарственного препарата на величину некоторого параметра.

Опыт X	**Контроль Y**
160	**180**
120	**160**
140	**220**
180	**180**
130	**160**
160	**200**
	170

По критерию Стьюдента для уровня значимости **α** =0,05 проверить , эффективен ли препарат.

$n_1 = 6, \qquad n_2 = 7$

Выдвигаем нулевую гипотезу:

$H_0 : M[X] = M[Y]$ Различия между выборками не значимы $\Rightarrow$
препарат не эффективен

$$\bar{x} = \frac{160+120+140+180+130+160}{6} \approx 148$$

$$\bar{y} = \frac{180+160+220+180+160+200+170}{7} \approx 181$$

$$S_{\bar{x}-\bar{y}} = \sqrt{\frac{\sum_{i=1}^{n_1}(n_i-\bar{x})^2 + \sum_{i=1}^{n_2} y_i-\bar{y})^2}{(n_1+n_2-2)} \cdot \left(\frac{1}{n_1}+\frac{1}{n_2}\right)} =$$

$$= \sqrt{\frac{(160-148)^2+(120-148)^2+(140-148)^2+(180-148)^2+(130-148)^2+(160-148)^2+}{(6+7-2)}}$$

$$\sqrt{\frac{+(180-181)^2+(160-181)^2+(220-181)^2+(180-181)^2+(160-181)^2+(200-181)^2}{(6+7-2)}}$$

$$\sqrt{\frac{(170-181)^2}{}\cdot\left(\frac{1}{6}+\frac{1}{7}\right)}=\sqrt{\frac{5371\cdot 13}{11\cdot 42}}\sqrt{\frac{69823}{462}}\approx\sqrt{151}\approx 12,3$$

$$t_{эксп}=\frac{(\bar{x}-\bar{y})}{S_{\bar{x}-\bar{y}}}=\frac{148-181}{12,3}=\frac{-33}{12,3}=-2,68$$

Находим из таблицы критерия Стьюдента для $t_{крит}=2,20$ $\nu=n_1+n_2-2=6+7-2=11$ и заданного **α**=0,05,

Так как $|t_{эксп}|>t_{крит}$ отвергаем $\Rightarrow H_0$

2,68 2,20

Вывод: Различия между выборками стат. значимы $\Rightarrow$ препар эффективен

Опыт	Контроль
160	180
120	160
140	220
180	180
130	160
160	200
	170

Двухвыборочный t-тест с различными дисперсиями		
	Переменная 1	*Переменная 2*
Среднее	148,3333333	181,4285714
Дисперсия	496,6666667	480,952381
Наблюдения	6	7
df	11	
t-статистика	-2,688935101	
P(T<=t) одностороннее	0,010534595	
t критическое одностороннее	1,795884814	
P(T<=t) двухстороннее	0,021069191	
t критическое двухстороннее	2,200985159	

3. Непараметрические (ранговые) критерии

Непараметрические критерии сравнивают сами значения выборок (варианты), они используют ранги.

Ранг -- это порядковый номер в ранжированных по возрастанию вариантах.

Если встречается несколько одинаковых значений, то их ранг равен среднему арифметическому нескольких последовательных порядковых номеров, соответствующих этим значениям в ряду.

Число рангов = n -- количество значений для которых расставляем ранги.

Пример:

X	Ранг
5	7
3	4
2	2,5
5	7
8	9
9	10
5	7
1	1
2	2,5
4	5
N=10	

Ранг «2»= $\frac{2+3}{2} = 2,5$

Ранг «5»= $\frac{6+7+8}{3} = 7$

3.1.Критерий Вилкоксона.

Работает с так называемыми сопряжёнными вариантами, когда варианты из двух выборок измеряются парами (например, значению x_i до воздействия препарата соответствует y_i после воздействия).

Итак, имеем две выборки одинакового объёма $\underline{n_1=n_2=n}$:

$X\{x_1, x_2, \dots x_n\}$ – контроль
$Y\{y_1, y_2, \dots y_n\}$ – опыт

Нас интересует значима ли разница между теми, у кого значение от x к y увеличилось (положительный сдвиг) и теми, у кого значение от x к y уменьшилось (отрицательный сдвиг), для заданного уровня значимости **α**.

H_0: различие между положительным и отрицательным сдвигом не значимо.

Алгоритм вычисления экспериментального значения критерия

1) Вычислить разности: $(x_i - y_i)$ или $(y_i - x_i)$

Если $(x_i - y_i)$ =0, то i-ю строку вычеркнуть и **n=n-k** -- количество вычеркнутых строк.

2) Расставить ранги для разностей, знак разности не учитываем. То есть расставляем ранги для $|x_i - y_i|$

3) Подсчитать суммы рангов, учитывая знаки разностей:

R^+ -- сумма рангов для $(x_i - y_i)$ >0

R^- -- сумма рангов для $(x_i - y_i)$<0

4) $T_{эксп} = \min(R^+, R^-)$, то есть выбираем меньшее из двух чисел

Определить по таблице критерия Вилкоксона для α и числа степеней свободы=n $T_{крит.}$

Если $T_{эксп} \leq T_{крит}$ то H_0 отвергаем.

если $T_{эксп} > T_{крит}$ то H_0 принимаем.

Пример: Значимы ли различия между выборками для уровня значимости α=0,05?

№	Контроль **X**	Опыт **Y**	Разности	Ранг разности
1	32	21	11	7
2	31	19	12	8
3	29	27	2	2,5
4	28	29	-1	1
5	30	30	0	
6	27	29	-2	2,5
7	29	22	7	6
8	33	27	6	5
9	26	21	5	4

H_0: Различия между выборками не значимы.

n=9-1=8 R^{-}=1+2,5=3,5 R^{+}=7+8+2,5+6+5+4=32,5

Следовательно $T_{эксп}$=3,5.

По таблице для n=8 и α=0,05 находим: $T_{крит}$=4.

$T_{эксп} < T_{крит} \Rightarrow$ H_0 отвергаем.

3,5 4

Вывод: Различия между выборками значимы.

Табличные значения критерия Вилкоксона

N	***α=0,05***	***α=0,01***
6	0	–
7	2	–
8	4	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	20
17	35	23
18	40	28
19	46	32
20	52	38
21	59	43
22	66	49
23	73	55
24	81	61
25	89	68

Критерий Манна-Уитни

Этот непараметрический критерий можно использовать для двух выборок как одинаковых, так и разных объёмов. Объём меньшей выборки обозначают **n_1.**

То есть, если $n_1 \neq n_2$ то $n_1 < n_2$

Обе выборки объединяют в один ряд и ранги расставляют для всех $n_1 + n_2$ чисел.

H_0: различие между выборками не значимо.

Алгоритм проверки статистической гипотезы:

1) Расставить ранги для всех $n_1 + n_2$ значений.

2) Вычислить: $U_1 = R_1 - \frac{n_1 \cdot (n_1 + 1)}{2}; \quad U_2 = R_2 - \frac{n_2 \cdot (n_2 + 1)}{2};$

где R_1 -- *сумма рангов для первой выборки,*

R_2 -- *сумма рангов для второй выборки.*

3) $U_{эксп} = \min(U_1, U_2)$

4)

Если $n_2 \leq 8,$ то в таблице для n_2
по n_1 и $U_{эксп}$ находим число -- P.
если P≥α то H_0 принимаем
если P<α то H_0 отвергаем
Где α -- заданный уровень значимости.

Если, $n_2 > 8$ то существует другая таблица. В ней для n_1 и n_2 находим $U_{крит}$

Если $U_{эксп} \leq U_{крит}$ то H_0 отвергаем

если $U_{эксп} > U_{крит}$ то H_0 принимаем

6). Записать вывод.

Пример 1: даны две выборки. По критерию Манна-Уитни проверить, значимы ли различия между выборками для уровня значимости α=0,05?

1-я выборка	*Ранг*	**2-я выборка**	*Ранг*
1	*1*	**3**	*2*
5	*3*	**8**	*5*
7	*4*	**10**	*7*
9	*6*	**12**	*8*
		13	*9*
n_1=4	*R_1=14*	**n_2=5**	*R_2=31*

H_0: Различия между выборками не значимы.

$n_1 + n_2 = 4+5=9$

$R_1 = 1+3+4+6=14$ $\quad U_1 = 14 - \frac{4 \cdot (4+1)}{2} = 4$

$R_2 = 2+5+7+8+9=31$ $\quad U_2 = 31 - \frac{5 \cdot (5+1)}{2} = 16$

$$U_{эксп} = \min(4;16) = 4$$

В таблице для $n_2=5$, находим для $n_1=4$ и $U_{эксп} = 4$

$$P = 0,095 > 0,05 = \alpha \Rightarrow$$ H_0 принимаем.

Вывод: Различия между выборками не значимы.

Таблицы вероятностей, связанных со значениями критерия Манна-Уитни .

	$N_1(N_2=5)$				
U	1	2	3	4	5
0	0.167	0.047	0.018	0.008	0.004
1	0.333	0.095	0.036	0.016	0.008
2	0.500	0.190	0.071	0.032	0.016
3	0.667	0.286	0.125	0.056	0.028
4		0.429	0.196	0.095	0.048
5		0.571	0.286	0.143	0.075
6			0.393	0.206	0.111
7			0.500	0.278	0.155
8			0.607	0.365	0.210
9				0.452	0.274
10				0.548	0.345
11					0.421
12					0.500

Пример 2: Изучалось действие различных лекарственных препаратов на двух группах животных. Получены следующие результаты:

1-я группа	1	2	2	3	4	5			
2-я группа	3	3	4	4	5	8	9	10	12

По критерию Манна-Уитни для уровня значимости α=0,05 выяснить, значима ли разница между действием этих препаратов.

H_0: Различия между выборками не значимы, то есть разница между действием препаратов не значима.

1-ая группа	Ранг	**2-ая группа**	Ранг
1	1	**3**	5
2	2,5	**3**	5
2	2,5	**4**	8
3	5	**4**	8
4	8	**5**	10,5
5	10,5	**8**	12
		9	13
		10	14
		12	15
$n_1 = 6$	$R_1 = 29,5$	$n_2 = 9$	$R_2 = 90,5$

$$n_1 + n_2 = 6 + 9 = 15$$

$$ранг(2) = \frac{2+3}{2} 2,5$$

$$ранг(3) = \frac{4+5+6}{3} = 5$$

$$ранг(4) = \frac{7+8+9}{3} = 8$$

$$ранг(5) = \frac{10+11}{2} = 10,5$$

$$R_1 = 1 + 2,5 + 2,5 + 5 + 8 + 10,5 = 29,5 \quad U_1 = 29,5 - \frac{6 \cdot (6+1)}{2} = 8,5$$

$$R_2 = 5 + 5 + 8 + 8 + 10,5 + 12 + 13 + 14 + 15 = 90,5$$

$$U_2 = 90,5 - \frac{9 \cdot (9+1)}{2} = 45,5$$

$$U_{эксп} = \min(8,5; 45,5) = 8,5$$

По таблице критических значений критерия Манна-Уитни для уровня значимости α=0,05.

$$U_{эксп} < U_{крит} \Rightarrow$$ H_0 отвергаем.

Вывод: разница между действием препаратов значима.

Таблица критических значений критерия Манна-Уитни для уровня значимости $\alpha=0,05$.

	N_2									
N_1	**9**	**10**	**11**	**13**	**13**	**14**	**15**	**16**	**17**	**18**
1										
2	0	0	0	1	1	1	1	1	2	2
3	2	3	3	4	4	5	5	6	6	7
4	4	5	6	7	8	9	10	11	11	12
5	7	8	9	11	12	13	14	15	17	18
6	10	11	13	14	16	17	19	21	22	24
7	12	14	16	18	20	22	24	26	28	30
8	15	17	19	22	24	26	29	31	34	36
9	17	20	23	26	28	31	34	37	39	42
10	20	23	26	29	33	36	39	42	45	48
11	23	26	30	33	37	40	44	47	51	55
12	26	29	33	37	41	45	49	53	57	61
13	28	33	37	41	45	50	54	59	63	67
14	31	36	40	45	50	55	59	64	67	74
15	34	39	44	49	54	59	64	70	75	80
16	37	42	47	53	59	64	70	75	81	86
17	39	45	51	57	63	67	75	81	87	93
18	42	48	55	61	67	74	80	86	93	99
19	45	52	58	65	72	78	85	92	99	106
20	48	55	62	69	76	83	90	98	105	112

Контрольные вопросы.

1. Критерий Стьюдента.

2. Критерий Вилкоксона.

3. Критерий Манна-Уитни.

Основы корреляционного анализа.

Наиболее простой вид связи между переменными величинами -- это функциональная зависимость:y=f(x). Каждому значению x соответствует одно значение y.

В медицине и биологии чаще встречается более сложный вид зависимости, когда каждому x соответствует множество значений y -- это корреляционная зависимость.

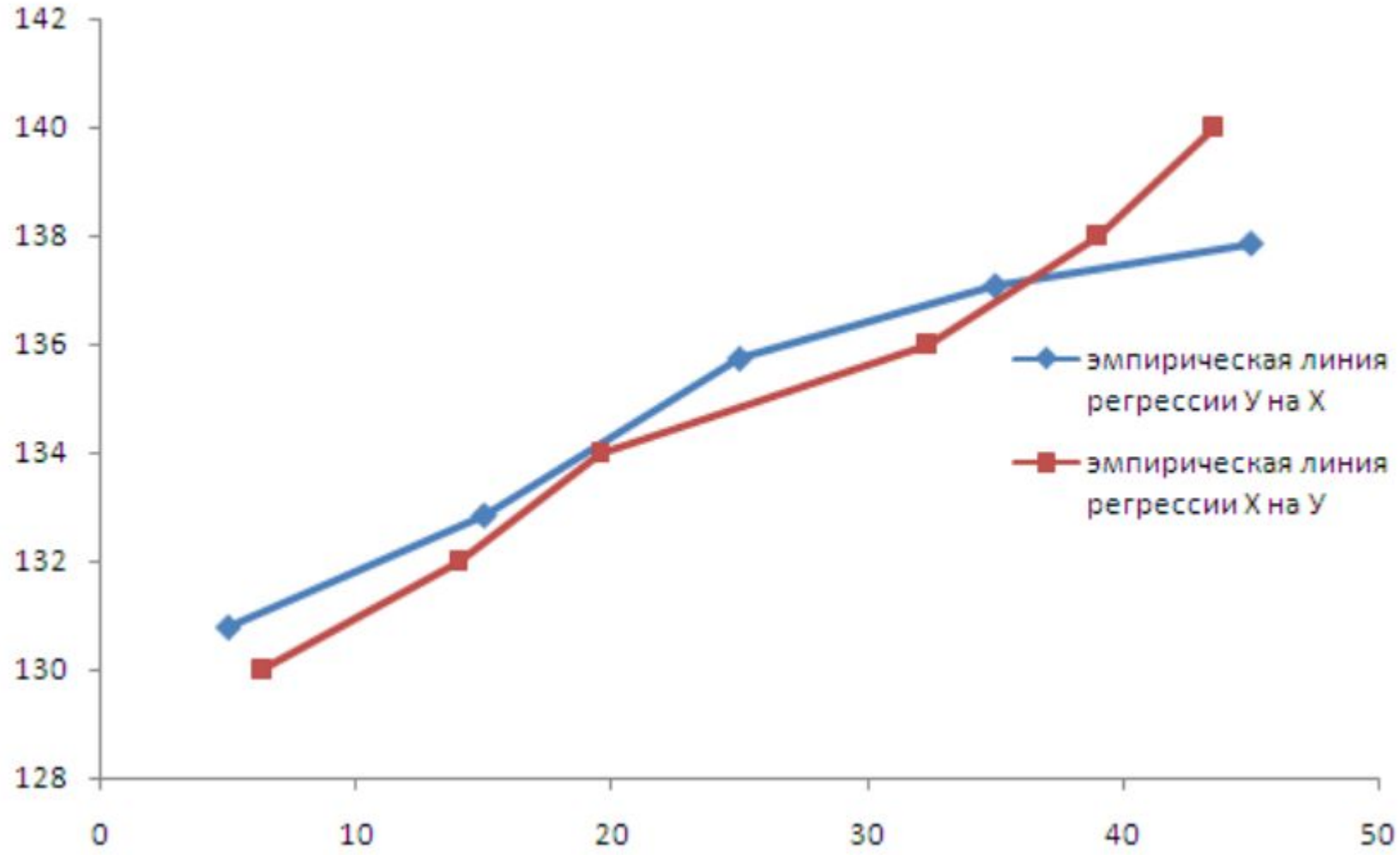
Каждому значению $\mathbf{x}_i$ соответствует множество значений **y**, среднее арифметическое этих значений $\bar{\mathbf{y}}_i$ называется условным средним.

$\overline{x},\overline{y}$ -- общие средние. Это средние арифметические, вычисленные по всем значениям x и y.

Среди множества точек с изменением x можно выделить точки, соответствующие условным средним y: $\overline{y}_1, \overline{y}_2, \overline{y}_3, \ldots \overline{y}_n$. Если соединить эти точки кривой линией, то получим линию регрессии, а соответствующая ей функция

$y=\overline{y}(x)$ -- функция регрессии.

Точно также, при изменении значений y, каждому y_i соответствует множество значений x, их средние арифметические $\overline{x}_i$ -- условные средние, соединив их кривой , получаем вторую линию регрессии, ей соответствует функция регрессии: $x=\overline{x}(y)$.


142
140
138
136
134
132
130
128
0
10
20
30
40
50
эмпирическая линия регрессии У на Х
эмпирическая линия регрессии Х на У

Следовательно, в отличии от функциональной зависимости, корреляционная зависимость характеризуется двумя линиями регрессии.

Уравнение регрессии

В настоящее время изучение различных корреляций является важным разделом многих биологических дисциплин, поэтому возникает потребность в количественном измерении корреляции.

Для этого служит ряд методов, наиболее распространённым из которых является вычисление коэффициента корреляции -- это количественная характеристика связи (зависимости) между исследуемыми величинами.

1. Дисперсия суммы случайных величин. Корреляционный момент.

X и **Y** -- случайные величины.

(1) $Z=X+Y$ -- их сумма.

(2) $M[Z]=M[X]+M[Y]$

Найдём $D[Z]=D[X+Y]$, для этого вычтем из уравнения (1) уравнение (2):

(3) $Z-M[Z]=X+Y-M[X]-M[Y]=(X-M[X])+(Y-M[Y])$

Для сокращения записи обозначают:

$Z-M[Z]=\Delta Z$

$X-M[X]=\Delta X$

$Y-M[Y]=\Delta Y$

} Эти величины называют моментами

Момент -- это отклонение каждого значения случайной величины от её математического ожидания.

Возведём уравнение (3) в квадрат:

$(Z-M[Z])^2=((X-M[X])+(Y-M[Y]))^2$ $\Delta Z^2=(\Delta X+\Delta Y)^2$, тогда

$\Delta Z^2=\Delta X^2+\Delta Y^2+2\cdot\Delta X\cdot\Delta Y$ -- это сумма квадратов отклонений.

Математическое ожидание от суммы квадратов отклонений это дисперсия:

$D[Z]=D[X+Y]=M[\Delta Z^2]=M[\Delta X^2]+M[\Delta Y^2]+2\cdot M[\Delta X\cdot\Delta Y]=D[X]+D[Y]+2\cdot M[\Delta X\cdot\Delta Y]$

Принято обозначение: $M[\Delta X \cdot \Delta Y]=K[X,Y]$ -- корреляционный момент. Основное свойства корреляционного момента: если величины X и Y независимы, то их корреляционный момент $K[X,Y]=0$. Обратное утверждение неверно.

Из последнего утверждения следует:

2. Теорема сложения дисперсий.

Если величины Хи Y независимы, то:

$D[X+Y]= D[X]+D[Y]$

Этой теоремой пользуются в теории погрешностей, при обработке результатов косвенных измерений. Так как входящие в расчётные формулы величины в большинстве случаев независимы, то подсчитывая среднюю квадратическую погрешность, суммируют квадраты всех их погрешностей.

3. Коэффициент корреляции (параметрический).

Корреляционный момент K[X,Y] – размерная величина, то есть зависит от выбора единицы измерения. Это затрудняет сравнение корреляционных моментов различных случайных величин, поэтому удобнее использовать безразмерную величину -- коэффициент корреляции:

$$\rho[X,Y]=\frac{K[X,Y]}{\sigma[X]\cdot\sigma[Y]}$$ -- это коэффициент корреляции для генеральной совокупности.

$$K[X,Y]=M\left[(X-M[X])\cdot(\mathrm{Y}-\mathrm{M}[Y])\right]=\frac{\sum_{i=1}^{n}(\mathrm{x}_i-M[X])\cdot(\mathrm{y}_i-M[Y])}{n}$$

$$\sigma[X]=\sqrt{\frac{\sum_{i=1}^{n}(\mathrm{x}_i-M[X])^2}{n}}$$

$$\sigma[Y]=\sqrt{\frac{\sum_{i=1}^{n}(\mathrm{y}-M[Y])^2}{n}}$$

Но мы имеем дело с выборкой, n конечно, выборочные оценки M[X] и M[Y] -- это $\bar{x}$ и $\bar{y}$ -- общие средние (средние арифметические всех значений X и Y, которые мы имеем из выборки).Поэтому для вычисления коэффициента корреляции для выборки, используют формулу:

$$R[X,Y]=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Свойства коэффициента корреляции:

1). $-1\le R[X,Y]\le+1$

если $R[X,Y]>0$ то корреляция называется положительной,
если $R[X,Y]<0$ то корреляция называется отрицательной.

2). если $R[X,Y]\approx1$, зависимость между X и Y близка к линейной.

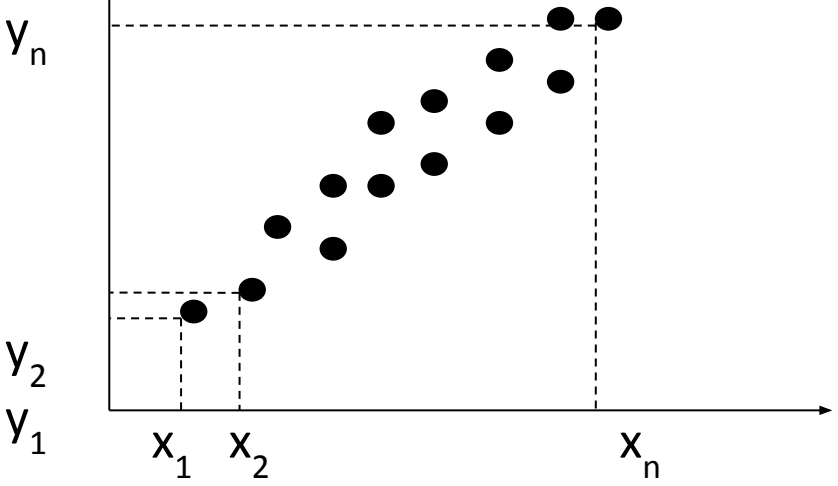
3). $R[X,Y]=\pm1$, то X и Yсвязаны линейной зависимостью:

y=ax+b

x=cy+d

Так как мы имеем дело с выборочной совокупностью, то имеем не множество значений X и Y, а несколько пар выборочных значений: (x_i, y_i), i=1 : n.

Сильная положительная корреляция:



R≈+1

Например: X -- нагрузка ↑

Y -- частота пульса ↑

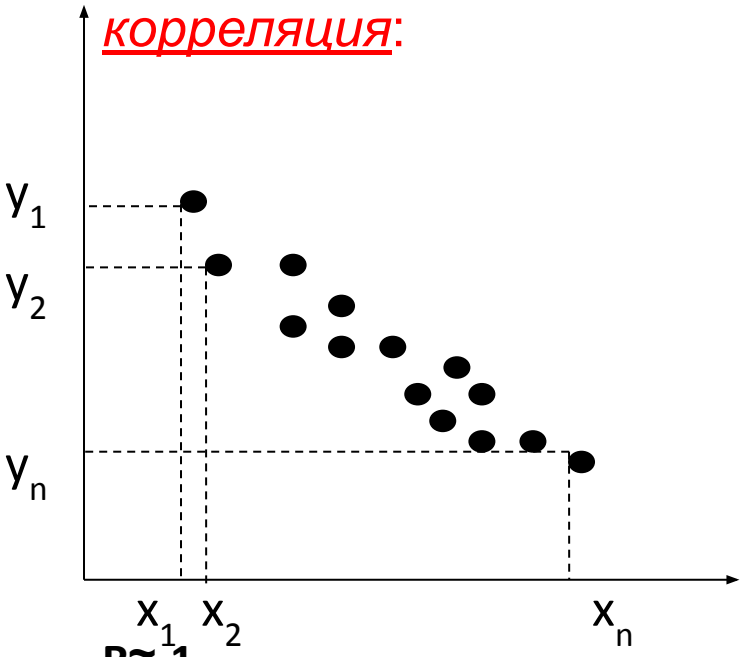
Слабая положительная корреляция:

R≠+1, R>0

Например: X -- число пятен на солнце ↑

Y – количество инфарктов ↑

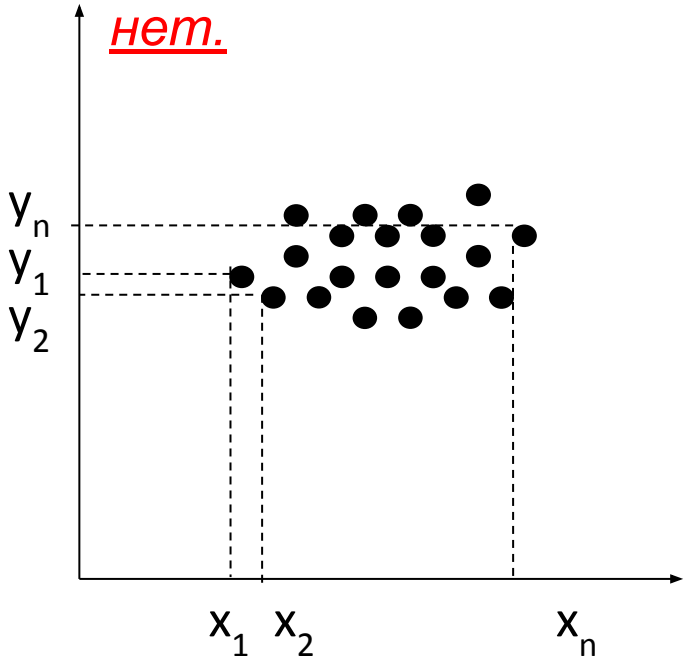
Сильная отрицательная корреляция:



R≈-1

X↑, Y↓

Корреляции (зависимости) нет.



R=0

Так как коэффициент корреляции R[X,Y] вычисляется по выборке, то есть является статистической оценкой ρ[X,Y]-- коэффициента корреляции генеральной совокупности, то R[X,Y] вычислен с ошибкой. Встаёт вопрос: значимо ли значение выборочного коэффициента корреляции (отличается от нуля)?

4. Оценка значимости параметрического коэффициента корреляции.

Проверка коэффициента корреляции на значимость осуществляется по критерию Стьюдента.

H_0: ρ=0,следовательно R[X,Y] не достоверен (то есть коэффициент корреляции генеральной совокупности ρ=0, следовательно зависимости между X и Y нет).

$$t_{эксп} = \frac{R-\rho}{\sigma_R} = \frac{R}{\sigma_R}$$

где σ_R -- ошибка выборочного коэффициента корреляции R[X,Y]

$$t_{эксп} = \frac{R\cdot\sqrt{n-2}}{\sqrt{1-R^2}}$$

для n<30 $\sigma_R = \sqrt{\frac{1-R^2}{n-2}}$ где **n-2** число степеней свободы.

В таблице критерия Стьюдента по заданному уровню значимости **α** и числу степеней свободы: **n-2** находим $t_{крит}$

если $|t_{эксп}| \le t_{крит} \Rightarrow$ H_0 принимаем.

Вывод: R недостоверен, зависимости между X и Y нет.

если $|t_{эксп}| > t_{крит} \Rightarrow$ H_0 отвергаем.

Вывод: R достоверен, зависимость (корреляция) между X и Y есть.

Если коэффициент корреляции статистически значим, то можно переходить к построению линий регрессии и записать уравнение регрессии.

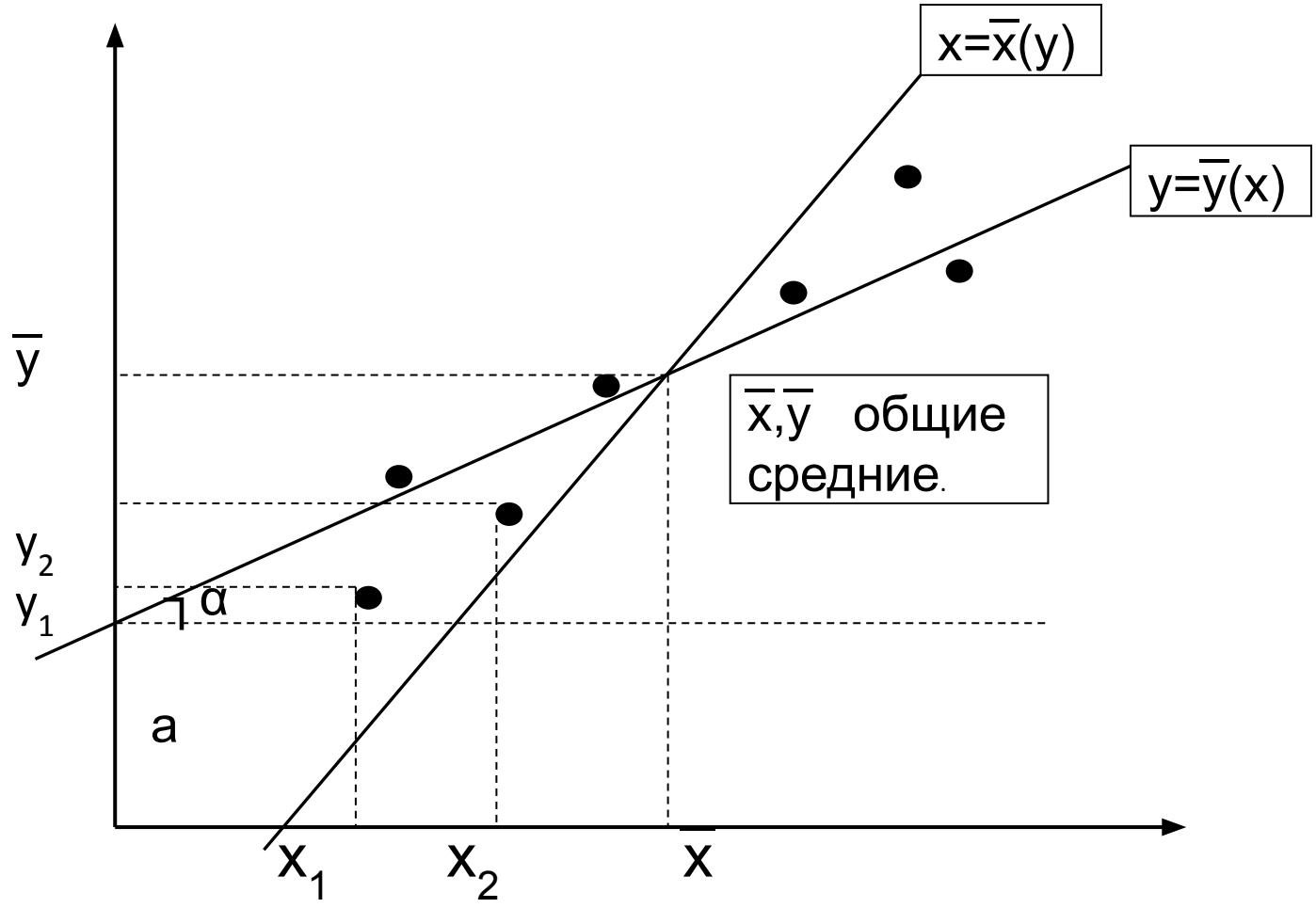
5. Построение линий регрессии.

Коэффициент корреляции R[X,Y] указывает лишь на наличие связи двух величин, но не даёт возможности судить, как количественно изменяется одна величина относительно другой. Ответ на этот вопрос даёт регрессионный анализ.

Корреляционная зависимость характеризуется двумя линиями регрессии:

Если R[X,Y] =1, то это линейная зависимость: $y=bx+a$.

Если R[X,Y] ≠1,то условные средние $\overline{y}_i$ не лягут на одну прямую, но при R[X,Y] ≈1 (R=0,95, R=0,85) можно провести усредняющую прямую: $y=b\cdot x+a$



b=tg α -- угловой коэффициент.

Так как имеем две линии регрессии:

$$y = b_{y/x} \cdot x + a_y$$

$$x = b_{x/y} \cdot y + a_x$$

$$R[X,Y] = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

то для $y=\bar{y}(x)$: $$b_{y/x} = \frac{R(X,Y) \cdot \sigma(Y)}{\sigma(X)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

то для $x=\bar{x}(y)$: $$b_{x/y} = \frac{R(X,Y) \cdot \sigma(X)}{\sigma(Y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$\sigma(Y)=S_n(Y)$ выборочные оценки среднего
$\sigma(X)=S_n(X)$ квадратического отклонения.

Так как обе линии регрессии проходят через точку с координатами $(\bar{x},\bar{y})$, где $\bar{x},\bar{y}$ -- общие средние, вычисленные по выборке, то уравнение регрессии имеет вид:

$$\begin{cases} y-\bar{y}=b_{y/x}\cdot(x-\bar{x}) \\ x-\bar{y}=b_{x/y}\cdot(y-\bar{y}) \end{cases}$$

или:

$$\begin{cases} y=b_{y/x}\cdot x-(b_{y/x}\cdot\bar{x}-\bar{y}) \\ \qquad\qquad\qquad\quad a_y \\ x=b_{x/y}\cdot y-(b_{x/y}\cdot\bar{y}-\bar{x}) \\ \qquad\qquad\qquad\quad a_x \end{cases}$$

6. Ранговый коэффициент корреляции.

Определить, есть ли связь между признаками XиY, можно и с помощью рангового коэффициента корреляции. Он менее точен и не может использоваться для построения линий регрессии, но так как ранговый коэффициент корреляции легко вычислить, есть смысл воспользоваться им для первоначальной оценки связи между признаками.

Ранговый коэффициент корреляции вычисляется по формуле:

$$R_{Sксп} = 1 - \frac{6 \cdot \sum_{i=1}^{n} (r_{x_i} - r_{y_i})^2}{n \cdot (n^2 - 1)}$$

n -- число пар x,y (объём выборки).
r_{xi} -- ранг по признаку X
r_{yi} -- ранг по признаку Y.
То есть ранги отдельно расставляются для X и Y.

Так как $R_{S\ эксп}$ вычислен по выборке, то он также нуждается в проверке на значимость.

Для этого $R_{S\ эксп}$ сравнивают с $R_{S\ критич}$, найденным в таблице ранговой корреляции ,для заданного уровня значимости α и числа степеней свободы n-2.

H_0: $R_{S\ эксп}$ не достоверен.

Если $|R_{S\ эксп}| \leq R_{S\ критич}$ то R_S ,полученный по выборке не достоверен, корреляции (зависимости) между X и Y нет.

Если $|R_{S\ эксп}| > R_{S\ критич}$ то R_S ,полученный по выборке достоверен, корреляция (зависимость) между X и Y есть.

Пример: По критерию ранговой корреляции выяснить, есть ли корреляция между признаками XиY для уровня значимости α=0,05.

X	Y	r_{xi}	r_{yi}	$(r_{xi}-r_{yi})$	$(r_{xi}-r_{yi})^2$
23	**5**	4	5	-1	**1**
21	**1**	3	1	2	**4**
24	**4**	5	4	1	**1**
20	**3**	1,5	3	-1,5	**2,25**
28	**8**	9,5	8	1,5	**2,25**
20	**2**	1,5	2	-0,5	**0,25**
26	**7**	7	7	0	**0**
25	**6**	6	6	0	**0**
27	**10**	8	10	-2	**4**
28	**9**	9,5	9	0,5	**0,25**
					Σ=15

n=10 α=0,05

$$R_{Sэксп} = 1 - \frac{6 \cdot \sum_{i=1}^{n} (r_{x_i} - r_{y_i})^2}{n \cdot (n^2 - 1)}$$

$$R_{Sэксп} = 1 - \frac{6 \cdot 15}{10 \cdot (10^2 - 1)} = 1 - \frac{90}{990} = 1 - 0,09 = 0,91$$

H_0: $R_{S\ эксп}$ не значим, корреляции нет.

По таблице ранговой корреляции для α=0,05 и числа степеней свободы 10-2=8 находим: $R_{S\ критич}$=0,64.

Так как

$$R_{Sэксп} > R_{Sкрит} \Rightarrow \mathbf{H_0}\ \text{отвергаем.}$$

Вывод: Ранговый коэффициент корреляции статистически значим, корреляция между признаками X и Y есть.

Коэффициент корреляции рангов

N	***α=0,05***	***α=0,01***
5	0,94	-
6	0,85	-
7	0,78	0,94
8	0,72	0,88
9	0,68	0,83
10	0,64	0,79
11	0,61	0,76
12	0,58	0,73
13	0,56	0,70
14	0,54	0,68

Вычислим параметрический коэффициент корреляции:

X	Y	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
23	**5**	-1,2	-0,5	0,6	1,44	0,25
21	**1**	-3,2	-4,5	14,4	10,24	20,25
24	**4**	-0,2	-1,5	0,3	0,04	2,25
20	**3**	-4,2	-2,5	10,5	17,64	6,25
28	**8**	3,8	2,5	9,5	14,44	6,25
20	**2**	-4,2	-3,5	14,7	17,64	12,25
26	**7**	1,8	1,5	2,7	3,24	2,25
25	**6**	0,8	0,5	0,4	0,64	0,25
27	**10**	2,8	4,5	12,6	7,84	20,25
28	**9**	3,8	3,5	13,3	14,44	12,25
$\bar{x}$=24,2	**$\bar{y}$=5,5**			**Σ=79**	**Σ=87,6**	**Σ=82,5**
					$\sqrt{\sum}$=9,36	**$\sqrt{\sum}$=9,08**

$$R[X,Y]=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{79}{9,36\cdot 9,08}\approx 0,93$$

Проверка на достоверность:

H_0: ρ=0,следовательно R[X,Y] не достоверен (то есть коэффициент корреляции генеральной совокупности ρ=0, следовательно зависимости между XиY нет).

$$t_{эксп} = \frac{R \cdot \sqrt{n-2}}{\sqrt{1-R^2}} = \frac{0,93 \cdot \sqrt{10-2}}{\sqrt{1-0,93^2}} \approx 7,1$$

В таблице критерия Стьюдента по заданному уровню значимости α=0,05 и числу степеней свободы: n-2=10-2=8

находим $t_{крит}$ =2,31

Так как $|t_{эксп}| > t_{крит} \Rightarrow$ H_0 отвергаем.

Вывод: R статистически значим, зависимость (корреляция) между X и Y есть.

Таблица критерия Стьюдента.

v	***α=0,05***	***α=0,01***
1	12,71	63,66
2	4,30	9,93
3	3,18	5,84
4	2,78	4,60
5	2,57	4,03
6	2,45	3,71
7	2,37	3,50
8	2,31	3,36
9	2,26	3,25
10	2,23	3,17
11	2,20	3,11
12	2,18	3,06
13	2,16	3,01

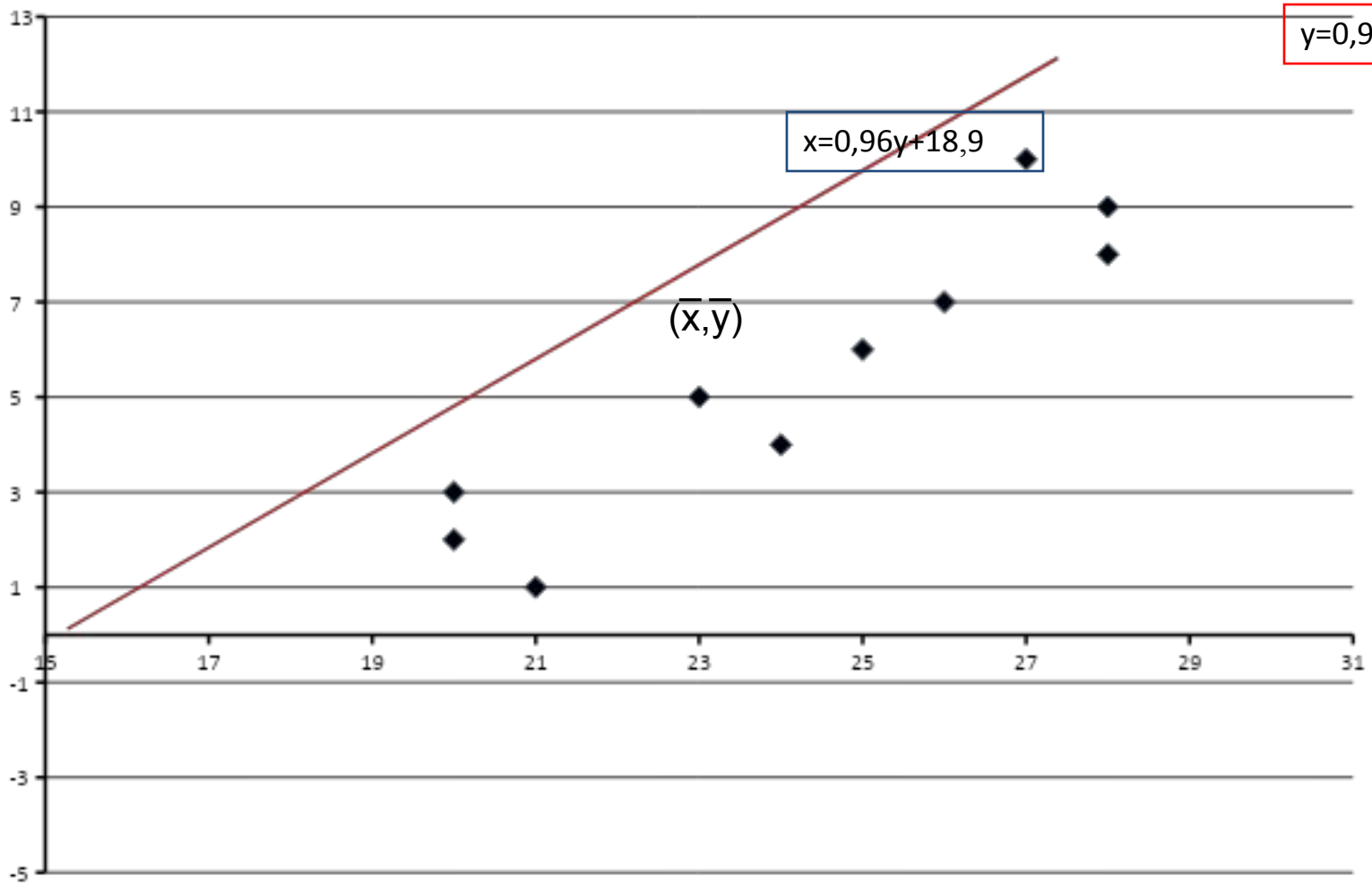
Построение линий регрессии:

$$b_{y/x} = \frac{R(X,Y)\cdot\sigma(Y)}{\sigma(X)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{79}{87,6} = 0,9$$

$$b_{x/y} = \frac{R(X,Y)\cdot\sigma(X)}{\sigma(Y)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{79}{82,5} = 0,96$$

Уравнение регрессии:

$$\begin{cases} y = b_{y/x}\cdot x - (b_{y/x}\cdot\bar{x} - \bar{y}) = 0,9\cdot x - (0,9\cdot 24,2 - 5,5) = 0,9\cdot x - 16,3 \\ x = b_{x/y}\cdot y - (b_{x/y}\cdot\bar{y} - \bar{x}) = 0,96\cdot y - (0,96\cdot 5,5 - 24,2) = 0,96\cdot y + 18,9 \end{cases}$$


y=0,9x-16,3
x=0,96y+18,9
(x̄,ȳ)
13
11
9
7
5
3
1
-1
-3
-5
15
17
19
21
23
25
27
29
31

Контрольные вопросы.

1. Что такое корреляция?
2. Дисперсия суммы случайных величин. Корреляционный момент.
3. Теорема сложения дисперсий.
4. Параметрический коэффициент корреляции .
5. Проверка параметрического коэффициента корреляции на достоверность.
6. Построение линий регрессии.
7. Ранговый коэффициент корреляции.
8. Проверка рангового коэффициента корреляции на достоверность.