Корреляционный анализ

Преподаватель: Башашина Ксения Викторовна

19.04.2015

Корреляционный анализ

Вообще, в природе, и в медицине в частности, существуют вполне определённые связи признаков. Например, связь между строением тела и предрасположенностью к тем или иным заболеваниям, связь между телосложением и темпераментом.

Наиболее простым видом связи между величинами является функциональная зависимость, когда какая-либо величина определяется как однозначная функция другой или нескольких других величин. Иными словами, *функциональная* связь – это такая связь между переменными, при которой каждому значению одной величины соответствуют строго определённые значения другой. Например, к функциональной относится зависимость между высотой местности и насыщением гемоглобина кислородом.

Однако, нередко встречаются и такие связи между величинами, которые нельзя отнести к функциональным зависимостям. К ним, например, относятся связи между урожаем и количеством осадков или между ростом отцов и сыновей. Известно, что между ростом и массой тела человека существует положительная связь, т.е. более высокие люди обычно имеют большую массу, но бывают и исключения.

Если связь между показателями проявляется не в каждом случае, а заметна лишь при многократном сопоставлении рассматриваемых признаков, то её называют **корреляционной** (от лат. correlatio – связь, соответствие).

Корреляция (Correlation) – связь между двумя или более переменными (в последнем случае корреляция называется множественной). Цель корреляционного анализа – установление наличия или отсутствия этой связи.

Корреляционная зависимость характеризуется тем, что каждому значению одной величины соответствует *множество* возможных значений другой величины. Например, при росте человека 170 см масса тела может быть 70 кг, 65 кг, 72 кг и т.д. Случайный разброс этих возможных значений объясняется влиянием большого числа дополнительных факторов, от которых отвлекаются, изучая связь между данными величинами.

Пусть сделаны измерения двух признаков X и У:

$X_1, X_2,...,X_n$ и $Y_1, Y_2,...,Y_n$.

Необходимо установить, *существует ли связь между изменениями признаков X и Y* и, если эта связь существует, то *определить её тип, глубину и достоверность*.

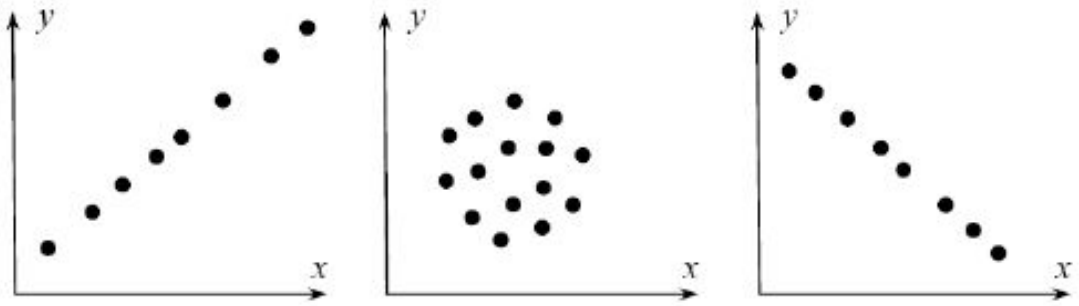
Для *качественной оценки* связи между признаками строят график.

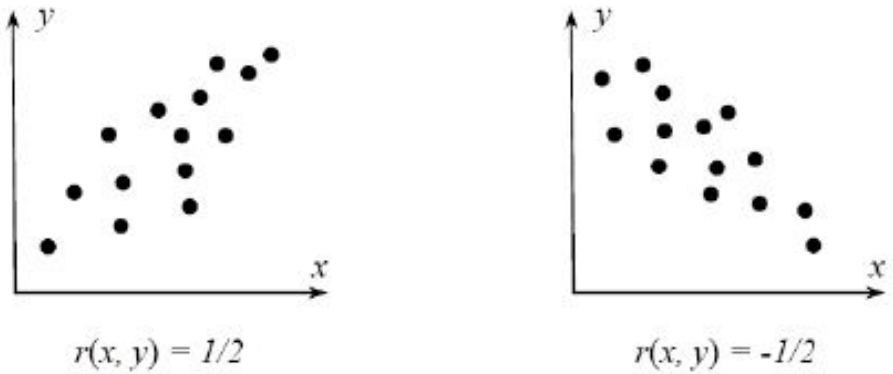
Экспериментальные графики для величин X и Y, находящихся в корреляционной зависимости, состоят из ряда точек, не укладывающихся на какую-либо определённую кривую. Каждая точка (x,y) на плоскости отображает результат одного измерения. Такой точечный график называют ***корреляционным полем***. По корреляционному полю можно качественно оценить наличие или отсутствие зависимости и указать положительна она или отрицательна.

- *Количественная оценка.*

В случае, когда имеются две переменных, значения которых измерены в цифровой шкале отношений (единицы измерений при этом не важны – например, масса тела может быть измерена в граммах, килограммах, тоннах – они не влияют на значение коэффициента корреляции), используется коэффициент линейной корреляции Пирсона r, который принимает значения от -1 до +1 (нулевое его значение свидетельствует об отсутствии корреляции).

Корреляционные поля





Величины коэффициента линейной корреляции в различных ситуациях

Проанализировав знак коэффициента корреляции, определяют *тип* корреляционной связи:

- если $r > 0$, то связь *прямая (положительная)*, т.е. при возрастании одной величины другая в среднем тоже возрастает;
- если $r < 0$, то связь *обратная (отрицательная)*, т.е. при возрастании одной величины другая имеет тенденцию в среднем убывать.
- Если статистическая связь между признаками *отсутствует*, то $r = 0$.

Величина коэффициента корреляции показывает *глубину* линейной связи между двумя выборками, т.е. характеризует степень близости зависимости величин X и Y к линейной функциональной зависимости. Графически это выражается *теснотой* или разбросанностью точек корреляционного поля.

Глубина корреляционной связи определяется, исходя из следующих критериев:

- если $0 < |r| \leq 0{,}3$, то связь *слабая;*
- если $0{,}3 < |r| \leq 0{,}5$, то связь *умеренная;*
- если $0{,}5 < |r| \leq 0{,}7$, то связь *значительная;*
- если $0{,}7 < |r| \leq 0{,}9$, то связь *сильная;*
- если $0{,}9 < |r| < 1$, то связь *очень сильная.*
- При $|r| = 1$ связь между величинами *функциональная*.

Таким образом, чем ближе абсолютная величина **r** к единице, тем сильнее связь между признаками и теснее расположены точки на графике. Однако, для обоснованного вывода о наличии связи не достаточно анализа величины коэффициента корреляции; необходимо проверить его *достоверность*. Иными словами, требуется ответить на вопрос: является ли вычисленный по данным наблюдений коэффициент корреляции *значимым*, т.е. можно ли верить полученному значению коэффициента, учитывая случайный характер выборок значений исследуемых величин. Значимость корреляционной связи при определённом уровне доверительной вероятности можно проверить с помощью критерия Стьюдента.

В случае *линейной корреляции* между признаками X и Y алгоритм расчетов по данному методу следующий:

- Вычисляют средние арифметические значения обоих признаков:

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

- Вычисляют отклонения каждого значения x_i от $\bar{x}$:

$$\Delta x_1 = x_1 - \bar{x}, \Delta x_2 = x_2 - \bar{x}, \dots, \Delta x_n = x_n - \bar{x}$$

- Вычисляют отклонения каждого значения y_i от $\bar{y}$

$$\Delta y_1 = y_1 - \bar{y}, \Delta y_2 = y_2 - \bar{y}, \dots, \Delta y_n = y_n - \bar{y}$$

- Вычисляют сумму произведений отклонений:

$$S_1 = \sum_{i=1}^{n} \Delta x_i \cdot \Delta y_i = \Delta x_1 \cdot \Delta y_1 + \Delta x_2 \cdot \Delta y_2 + \dots + \Delta x_n \cdot \Delta y_n$$

- Вычисляют произведение сумм квадратов отклонений:

$$S_2 = \sum_{i=1}^{n}(\Delta x_i)^2 \cdot \sum_{i=1}^{n}(\Delta y_i)^2 = [(\Delta x_1)^2 + (\Delta x_2)^2 + \cdots + (\Delta x_n)^2] \cdot [(\Delta y_1)^2 + \cdots + (\Delta y_n)^2]$$

- Определяют коэффициент r линейной парной корреляции по формуле:

$$r = \frac{S_1}{\sqrt{S_2}}$$

- Оценивают тип и глубину корреляционной связи между признаками X и Y.

- Вычисляют среднюю ошибку коэффициента корреляции:

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}},$$

где n – число коррелирующих пар.

- Определяют критерий достоверности коэффициента корреляции:

$$t_r = \frac{r}{m_r}$$

- Из таблицы Стьюдента для числа степеней свободы ν = n - 2 определяют стандартные значения критериев Стьюдента, соответствующие трем порогам достоверности: 0,95; 0,99; 0,999.

- Сравнивают критерий достоверности t_r со стандартными значениями критериев Стьюдента и делают вывод о достоверности коэффициента корреляции:

- если $t_r \geq t_{st0,999}$, то достоверность коэффициента корреляции 99,9%;
- если $t_r \geq t_{st0,99}$, то достоверность коэффициента корреляции 99%;
- если $t_r \geq t_{st0,95}$, то достоверность коэффициента корреляции 95%;
- если $t_r < t_{st0,95}$, то коэффициент корреляции недостоверен, доверять ему нельзя.

Коэффициент корреляции Пирсона также может быть вычислен в программе Excel функцией КОРРЕЛ.

Отметим, что коэффициент корреляции Пирсона симметричен, то есть не зависит от перестановки переменных: $r(y, x) = r(x, y)$. Универсальных рецептов установления корреляции между немонотонно и нелинейно связанными переменными на сегодняшний день не существует.

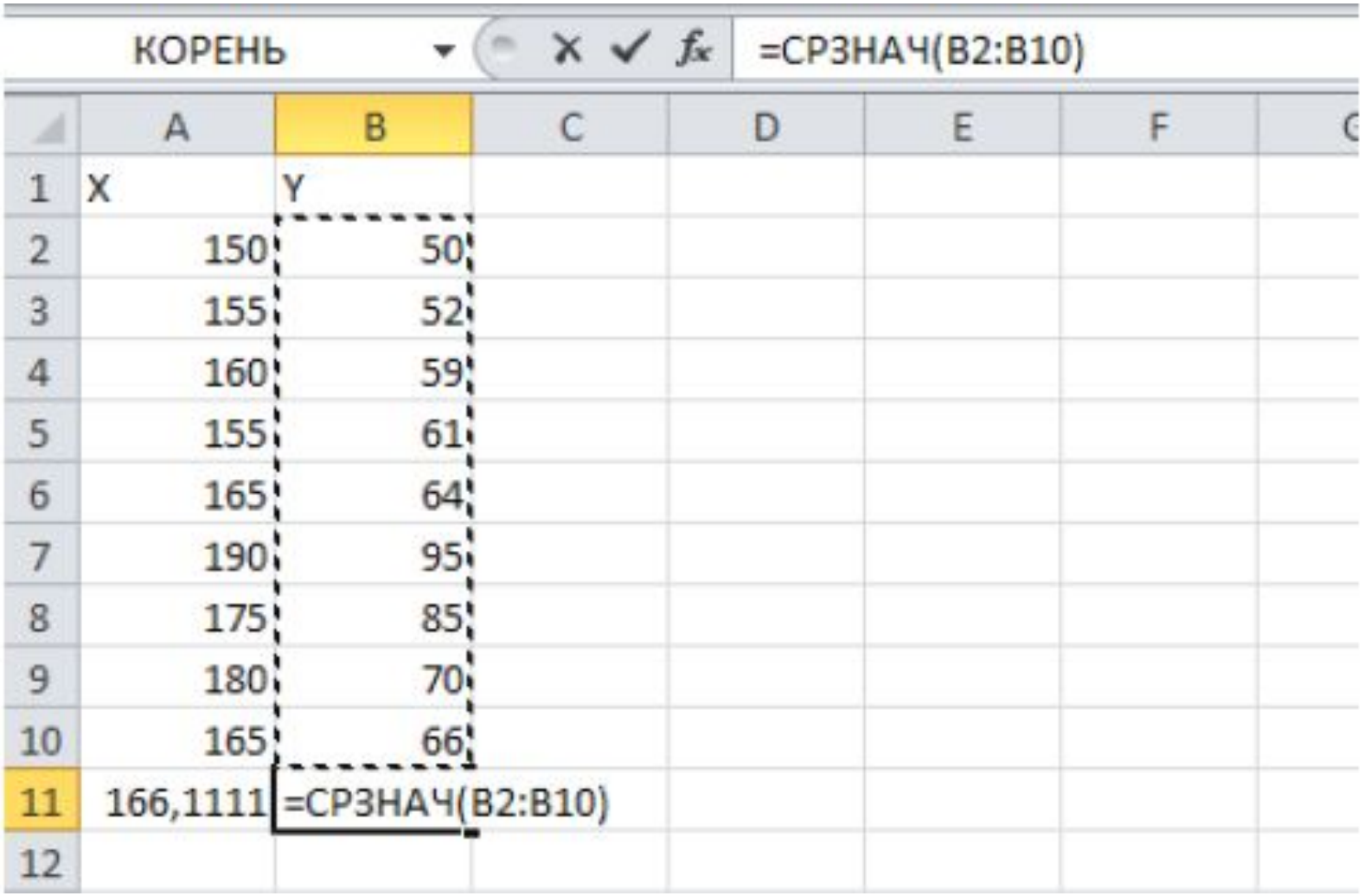
Задание

- В ходе обследования 9 пациентов среди прочих показателей измеряли их рост и вес. Результаты измерений приведены в таблице:

X_i Рост, см	150	155	160	155	165	190	175	180	165
Y_i Вес, кг	50	52	59	61	64	95	85	70	66

- Необходимо провести корреляционный анализ между весом и ростом пациентов. Построить корреляционное поле.

1. Вводим исходные данные. Вычисляем средние арифметические значения обоих признаков:



2. Вычисляем сумму произведений отклонений. Для этого вначале найдем отклонение каждого значения *x* и *y* от среднего значения. Обратите внимание на использование в формуле абсолютной ссылки.

E2 f_x =B2-B11

	A	B	C	D	E	F	G
1	X	Y		ΔX	ΔY		
2	150	50		-16,1111	-16,8889		
3	155	52		-11,1111	-14,8889		
4	160	59		-6,11111	-7,88889		
5	155	61		-11,1111	-5,88889		
6	165	64		-1,11111	-2,88889		
7	190	95		23,88889	28,11111		
8	175	85		8,888889	18,11111		
9	180	70		13,88889	3,111111		
10	165	66		-1,11111	-0,88889		
11	166,1111	66,88889	Ср_знач				
12							

3. Затем найдем произведение ΔX и ΔY.

F10 | =D10*E10

	A	B	C	D	E	F	G
1	X	Y		ΔX	ΔY	ΔX*ΔY	
2	150	50		-16,11111	-16,8889	272,0988	
3	155	52		-11,11111	-14,8889	165,4321	
4	160	59		-6,111111	-7,88889	48,20988	
5	155	61		-11,11111	-5,88889	65,4321	
6	165	64		-1,111111	-2,88889	3,209877	
7	190	95		23,888889	28,11111	671,5432	
8	175	85		8,8888889	18,11111	160,9877	
9	180	70		13,888889	3,111111	43,20988	
10	165	66		-1,111111	-0,88889	0,987654	
11	166,1111	66,88889	Ср_знач				
12							

4. И наконец, подсчитаем сумму произведений отклонений, используя функцию СУММ.

F11 f_x =СУММ(F2:F10)

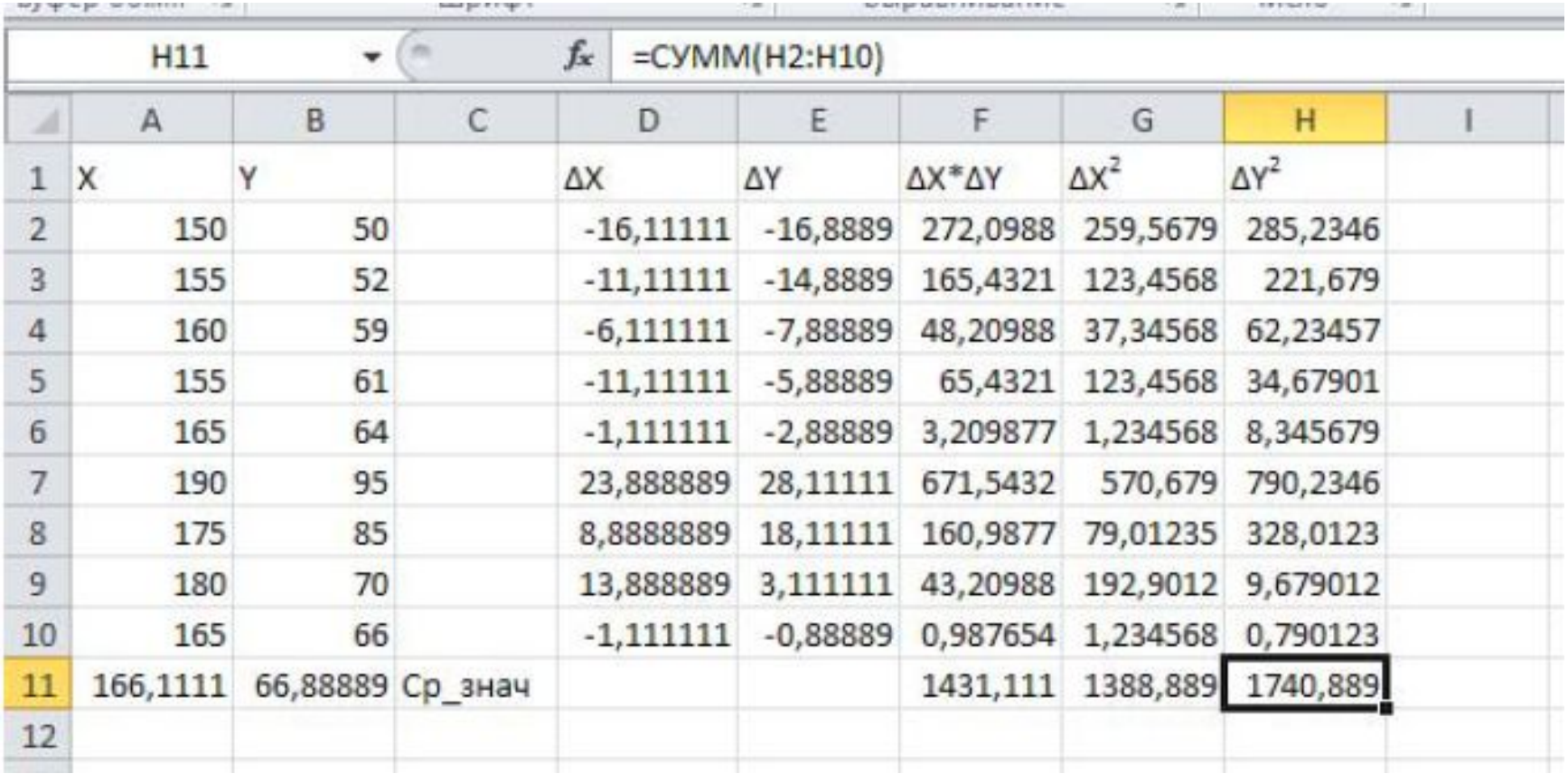
	A	B	C	D	E	F	G
1	X	Y		ΔX	ΔY	ΔX*ΔY	
2	150	50		-16,11111	-16,8889	272,0988	
3	155	52		-11,11111	-14,8889	165,4321	
4	160	59		-6,111111	-7,88889	48,20988	
5	155	61		-11,11111	-5,88889	65,4321	
6	165	64		-1,111111	-2,88889	3,209877	
7	190	95		23,888889	28,11111	671,5432	
8	175	85		8,8888889	18,11111	160,9877	
9	180	70		13,888889	3,111111	43,20988	
10	165	66		-1,111111	-0,88889	0,987654	
11	166,1111	66,88889	Ср_знач			1431,111	
12							

5. Вычисляем произведение сумм квадратов отклонений. Найдем ΔX^2 и ΔY^2.

H10 | f_x | =E10*E10

	A	B	C	D	E	F	G	H	I
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2	
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346	
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679	
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457	
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901	
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679	
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346	
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123	
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012	
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123	
11	166,1111	66,88889	Ср_знач			1431,111			
12									

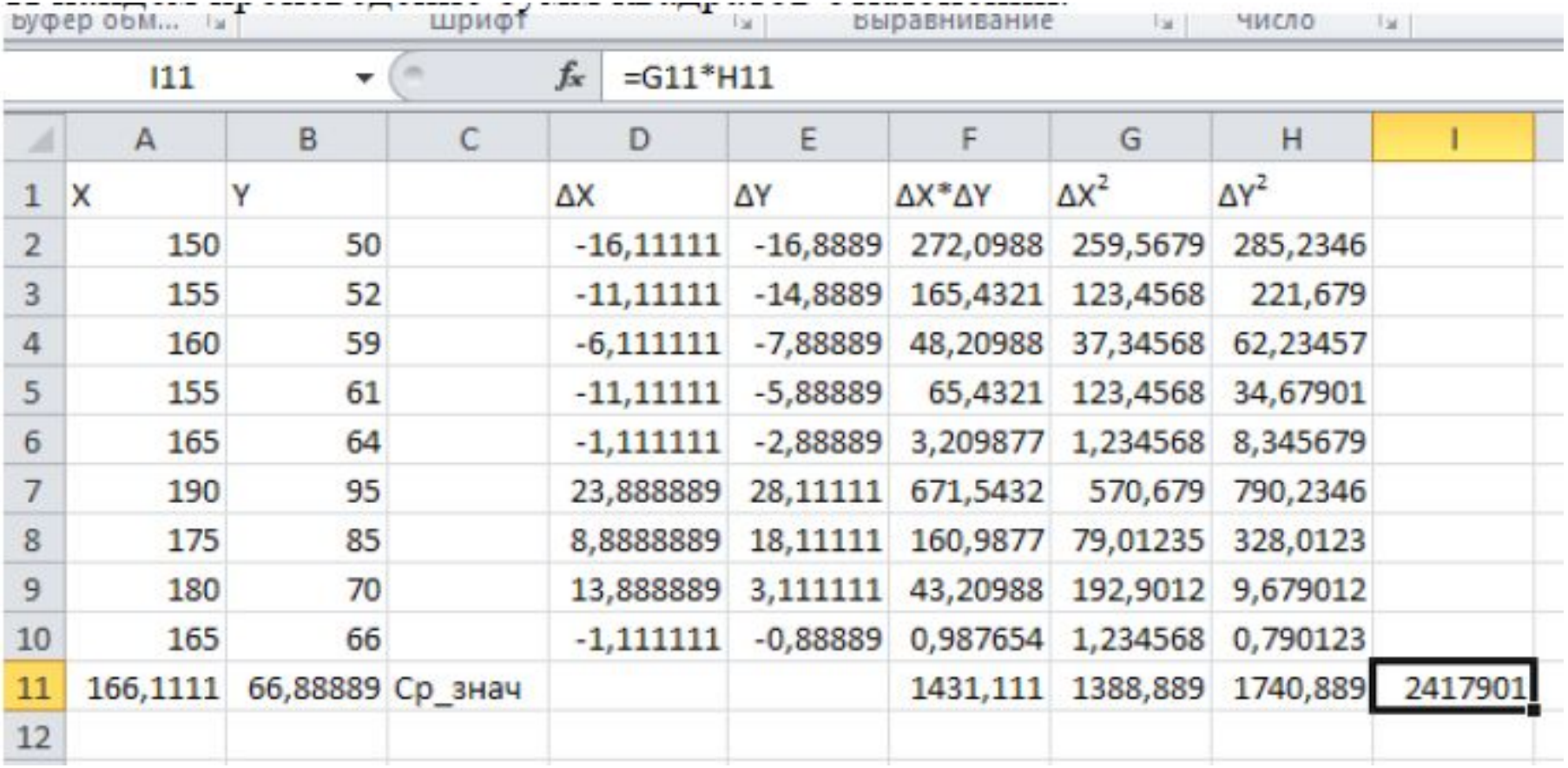
Затем подсчитаем сумму в каждом получившемся столбце.



H11 | =СУММ(H2:H10)

	A	B	C	D	E	F	G	H	I
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2	
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346	
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679	
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457	
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901	
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679	
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346	
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123	
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012	
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123	
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	
12									

И найдем произведение сумм квадратов отклонений.

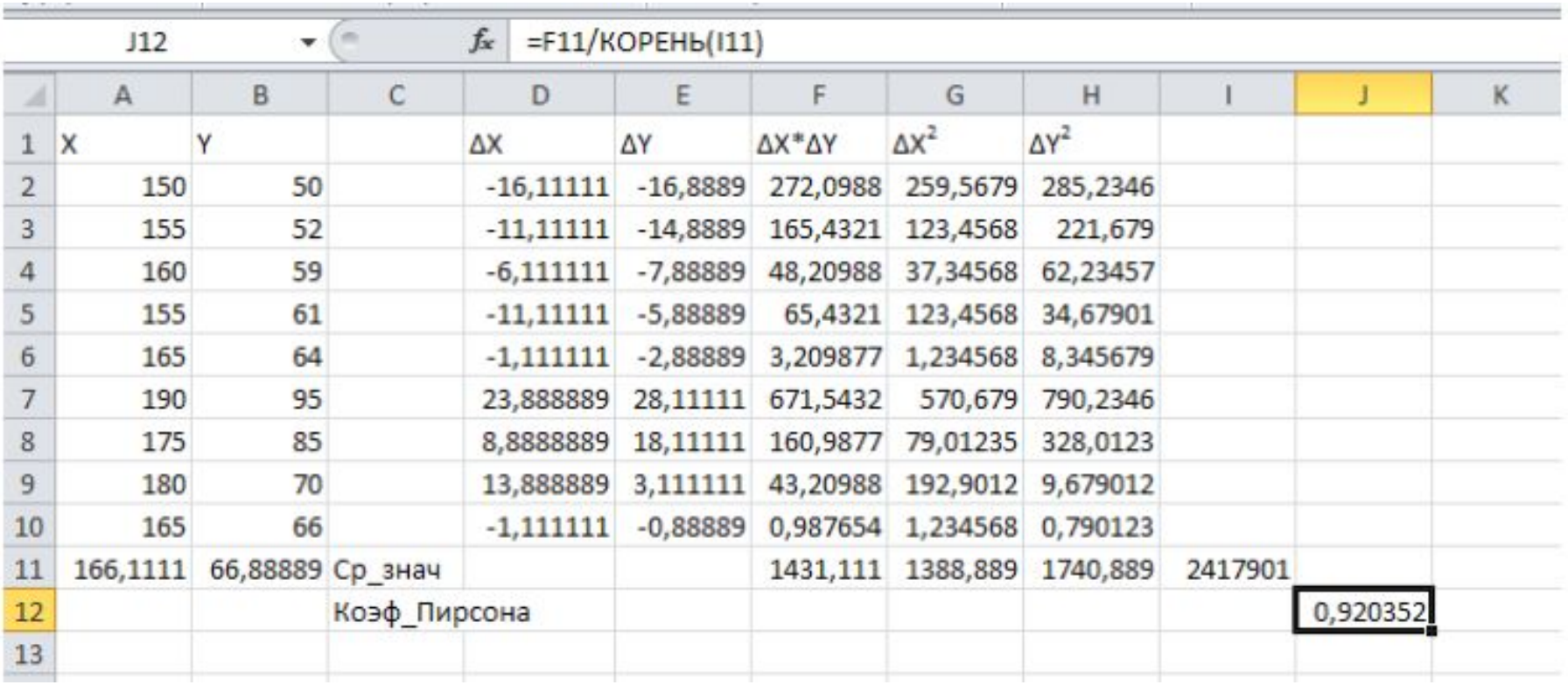


Буфер обм... | Шрифт | Выравнивание | Число

I11 | fx | =G11*H11

	A	B	C	D	E	F	G	H	I
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2	
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346	
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679	
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457	
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901	
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679	
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346	
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123	
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012	
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123	
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901
12									

6. Определяем коэффициент r линейной парной корреляции, используя приведенную формулу.



J12 =F11/КОРЕНЬ(I11)

	A	B	C	D	E	F	G	H	I	J	K
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2			
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346			
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679			
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457			
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901			
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679			
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346			
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123			
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012			
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123			
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901		
12			Коэф_Пирсона							0,920352	
13											

7. Оцениваем тип и глубину корреляционной связи между признаками X и У.

F17

	A	B	C	D	E	F	G	H	I	J
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2		
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346		
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679		
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457		
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901		
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679		
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346		
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123		
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012		
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123		
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901	
12			Коэф_Пирсона							0,920352
13				Связь						
14					прямая	(тип связи)				
15					сильная	(глубина связи)				
16										

8. Вычисляем среднюю ошибку коэффициента корреляции. Обратите внимание, что вид формулы в строке формул соответствует приведенной выше.

J17 | =КОРЕНЬ(((1-J12*J12)/(9-2)))

	A	B	C	D	E	F	G	H	I	J	K
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2			
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346			
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679			
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457			
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901			
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679			
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346			
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123			
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012			
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123			
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901		
12			Коэф_Пирсона							0,920352	
13				Связь							
14					прямая	(тип связи)					
15					сильная	(глубина связи)					
16											
17			Ср_ошибка							0,147818	
18											

9. Определяем критерий достоверности коэффициента корреляции.

J18 =J12/J17

	A	B	C	D	E	F	G	H	I	J	K
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2			
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346			
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679			
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457			
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901			
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679			
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346			
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123			
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012			
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123			
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901		
12			Коэф_Пирсона							0,920352	
13				Связь							
14					прямая	(тип связи)					
15					сильная	(глубина связи)					
16											
17			Ср_ошибка							0,147818	
18			Критерий достоверности							6,22624	
19											

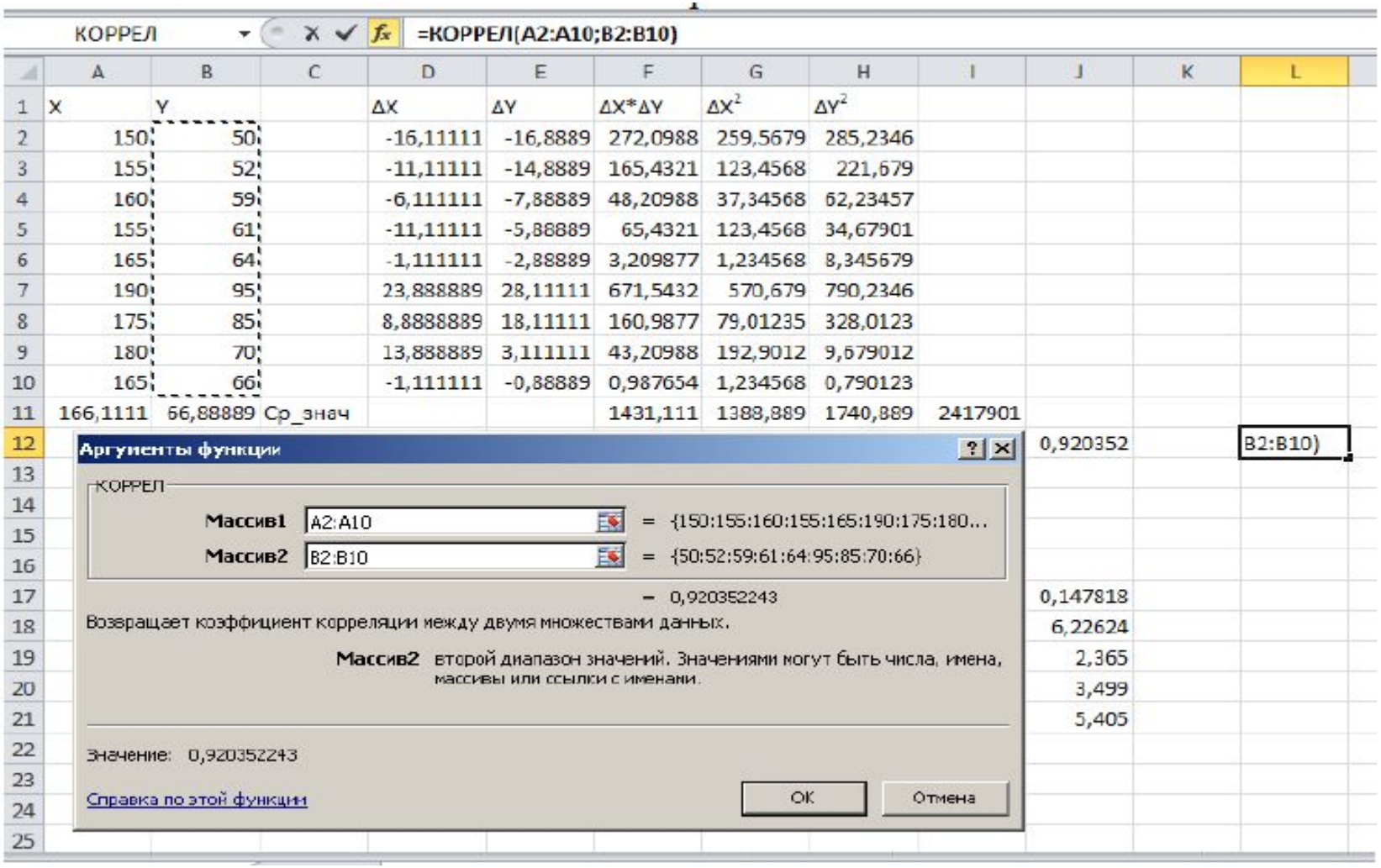
10. Из таблицы Стьюдента для числа степеней свободы v = n - 2 определяем стандартные значения критериев Стьюдента, соответствующие трем порогам достоверности: 0,95; 0,99; 0,999.

	A	B	C	D	E	F	G	H	I	J	K
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2			
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346			
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679			
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457			
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901			
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679			
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346			
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123			
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012			
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123			
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901		
12			Коэф_Пирсона							0,920352	
13				Связь							
14					прямая	(тип связи)					
15					сильная	(глубина связи)					
16											
17			Ср_ошибка							0,147818	
18			Критерий достоверности							6,22624	
19			Критерий Стьюдента					0,999		2,365	
20								0,99		3,499	
21								0,95		5,405	

11. Сравниваем критерий достоверности t_r со стандартными значениями критериев Стьюдента и делаем вывод о достоверности коэффициента корреляции.

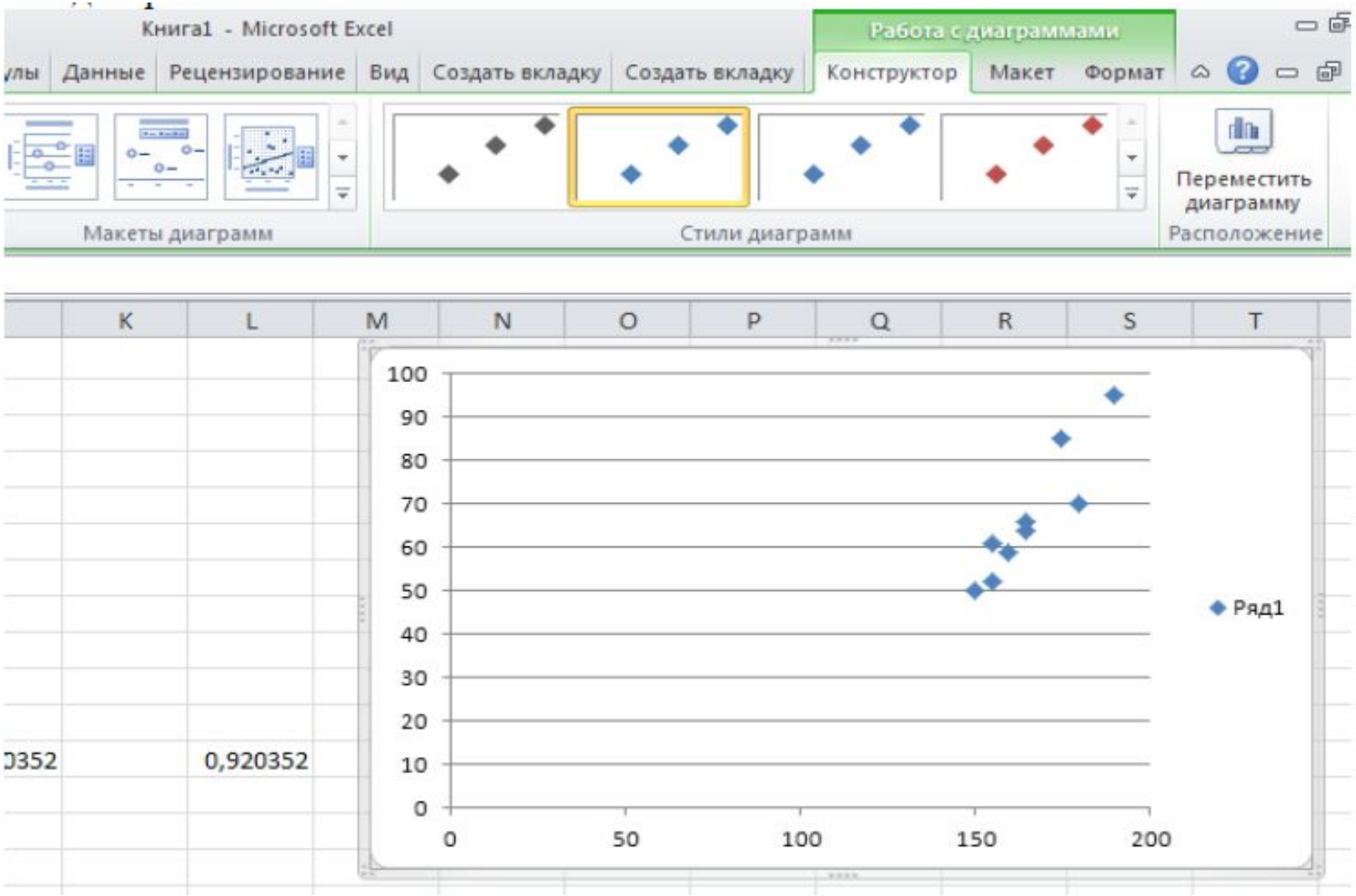
	A	B	C	D	E	F	G	H	I	J	K
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2			
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346			
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679			
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457			
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901			
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679			
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346			
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123			
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012			
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123			
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901		
12			Коэф_Пирсона							0,920352	
13				Связь							
14					прямая	(тип связи)					
15					сильная	(глубина связи)					
16											
17			Ср_ошибка							0,147818	
18			Критерий достоверности							6,22624	
19			Критерий Стьюдента					0,999		2,365	
20								0,99		3,499	
21								0,95		5,405	
22				Достоверность коэффициента корреляции 99,9 %							

13. Вычислим коэффициент корреляции Пирсона с помощью функции КОРРЕЛ. В качестве массивов 1 и 2 выберем наши массивы X и Y.

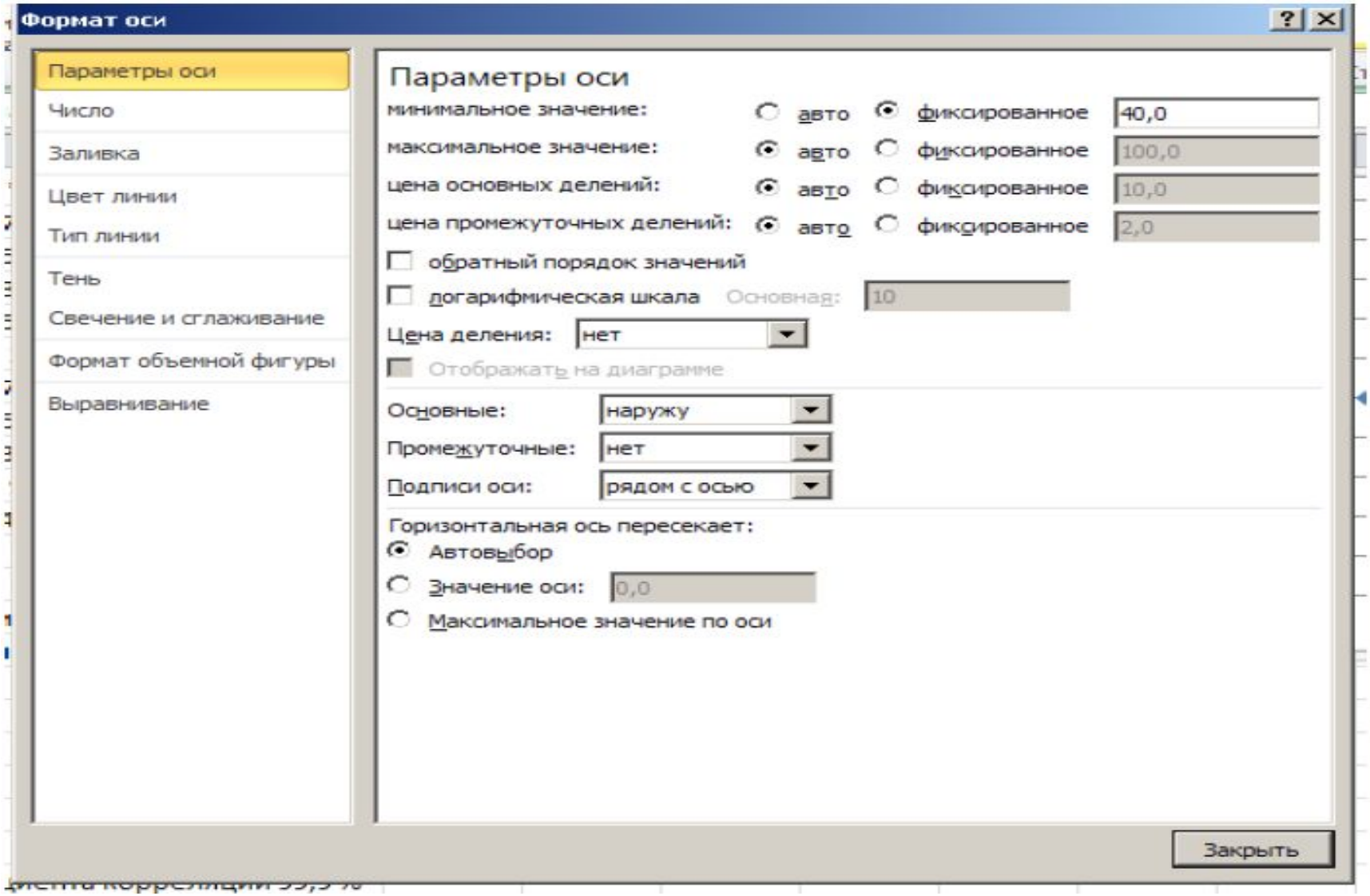


	A	B	C	D	E	F	G	H	I	J	K	L
1	X	Y		ΔX	ΔY	ΔX*ΔY	ΔX^2	ΔY^2				
2	150	50		-16,11111	-16,8889	272,0988	259,5679	285,2346				
3	155	52		-11,11111	-14,8889	165,4321	123,4568	221,679				
4	160	59		-6,111111	-7,88889	48,20988	37,34568	62,23457				
5	155	61		-11,11111	-5,88889	65,4321	123,4568	34,67901				
6	165	64		-1,111111	-2,88889	3,209877	1,234568	8,345679				
7	190	95		23,888889	28,11111	671,5432	570,679	790,2346				
8	175	85		8,8888889	18,11111	160,9877	79,01235	328,0123				
9	180	70		13,888889	3,111111	43,20988	192,9012	9,679012				
10	165	66		-1,111111	-0,88889	0,987654	1,234568	0,790123				
11	166,1111	66,88889	Ср_знач			1431,111	1388,889	1740,889	2417901			
12										0,920352		B2:B10)
13												
14												
15												
16												
17										0,147818		
18										6,22624		
19										2,365		
20										3,499		
21										5,405		
22												
23												
24												
25												

14. Построим корреляционное поле, используя вкладку Диаграммы, и выбрав тип диаграммы Точечная.

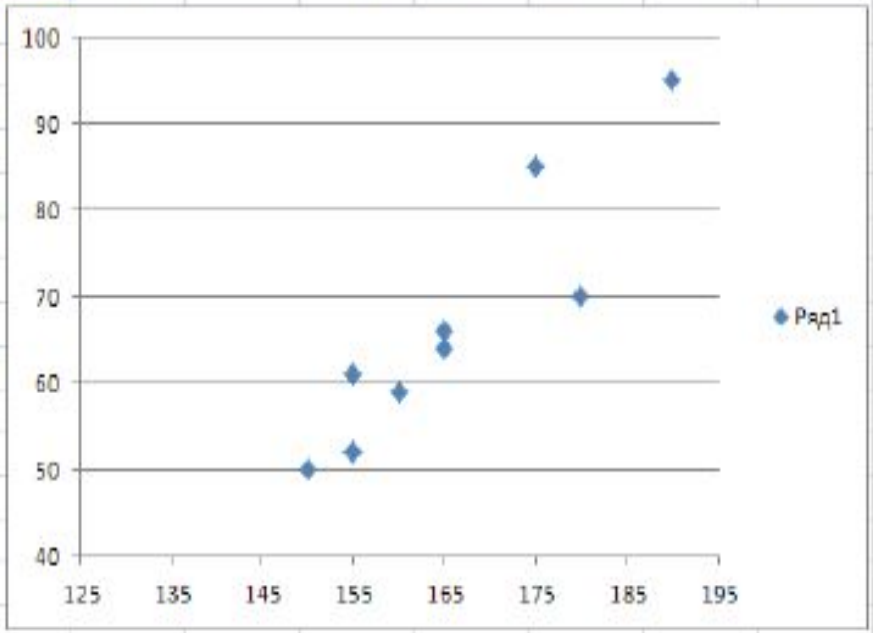


Т.к. диаграмма смещена в правый верхний угол, поместим ее в центр координатной плоскости. Для этого изменим минимальные значения осей X и Y.



M19

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	$\Delta X^*\Delta Y$	ΔX^2	ΔY^2												
2	272,0988	259,5679	285,2346												
3	165,4321	123,4568	221,679												
4	48,20988	37,34568	62,23457												
5	65,4321	123,4568	34,67901												
6	3,209877	1,234568	8,345679												
7	671,5432	570,679	790,2346												
8	160,9877	79,01235	328,0123												
9	43,20988	192,9012	9,679012												
10	0,987654	1,234568	0,790123												
11	1431,111	1388,889	1740,889	2417901											
12					0,920352		0,920352								
13															
14	(тип связи)														
15	(глубина связи)														
16															
17					0,147818										
18					6,22624										
19			0,999		2,365										
20			0,99		3,499										
21			0,95		5,405										
22	фициента корреляции 99,9 %														



Спасибо за внимание!