ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

УПРАВЛЕНИЕ ЦИФРОВЫХ ОБРАЗОВАТЕЛЬНЫХ ТЕХНОЛОГИЙ

Кафедра «Информационные технологи»

УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ

к проведению лабораторных и практических занятий по дисциплинам

«Статистические методы обработки и информации»,

«Разработка и стандартизация ПС и ИТ»

Авторы
Соболь Б.В., Пешхоев И.М.,
Остроух Е.Н., Богданова Н.Ю

Ростов-на-Дону, 2023

Аннотация

Методические указания предназначены для студентов направления для студентов направлений 09.03.02- Информационные системы и технологии, WEB-ориентированные информационно-аналитические системы и 09.03.03- Прикладная информатика; очной, очно-заочной, заочной форм обучения.

Авторы

Соболь Б.В.,
Пешхоев И.М.,
Остроух Е.Н.,
Богданова Н.Ю.

Оглавление

ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ применяется в задачах инноватики, в которых используются статистические данные, полученные либо опытным путем, либо с помощью экспертных оценок. При этом требуется оценить влияние качественного фактора (или нескольких факторов) на изучаемый процесс, что позволяет выделить из множества параметров задачи наиболее существенные, используемые при построении целевой функции.

Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ применяется для того, чтобы ответить на следующий вопрос: влияет ли некий качественный фактор *A*, имеющий несколько уровней A_1, A_2, ..., A_k, на случайную величину *X*, имеющую нормальный закон распределения вероятностей с дисперсией σ^2.

Находим критическое значение $f\left(\alpha; \nu_{\text{ÌÃ}}; \nu_{\text{ÂÃ}}\right) = f\left(\alpha; k-1; N-k\right)$ по таблице распределения Фишера-Снедекора.

Если $F_{\text{набл}} \geq f\left(\alpha; k-1; N-k\right)$, то делается вывод о том, что влияние фактора *A* значимо с уровнем значимости α.

Пример 1.1 Определить с уровнем α = 0,05 значимость различия производительности труда в трех бригадах рабочих-токарей за десять дней работы (табл. 1.1, за каждый день приведено среднее число изготовленных за час деталей на одного рабочего).

Таблица 1.1

Дни	1-я бригада	2-я бригада	3-я бригада
1	13	15	15
2	14	13	17
3	15	14	16
4	14	13	17
5	16	16	16
6	13	15	18
7	12	14	19
8	13	14	16
9	14	16	17
10	15	15	15

Используем табличный процессор EXCEL.

Решение. Введите данные о производительности труда из табл.1.1 в диапазоне A1:C11.

Войти в меню «Анализ данных», в появившемся окне выберите «Однофакторный дисперсионный анализ», укажите в строке «Входной интервал» диапазон A1:C11, установите флажок в строке «Метки», введите в строке «Альфа» уровень значимости 0,05, в строке «Выходной интервал» введите ссылку A15 и нажмите кнопку «Ok».

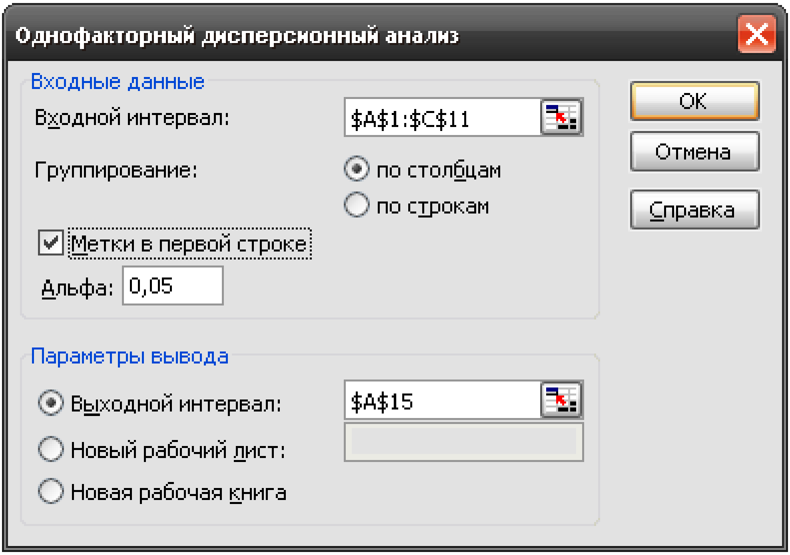

Рис.1.1

В результате расчета в диапазоне A15:G29 будут выведены итоги (рис.1.2). Значение *F*-критерия равно 14,3571, и оно больше критического значения 3,3541.

	A	B	C	D	E	F	G
15	Однофакторный дисперсионный анализ						
16							
17	ИТОГИ						
18	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
19	1 бригада	10	139	13,9	1,4333		
20	2 бригада	10	145	14,5	1,1667		
21	3 бригада	10	166	16,6	1,6		
22							
23							
24	Дисперсионный анализ						
25	*Источник*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
26	Между	40,2	2	20,1	14,3571	0,0001	3,3541
27	Внутри групп	37,8	27	1,4			
28							
29	Итого	78	29				

Рис.1.2

Вывод. Различие производительности труда в бригадах значимо с уровнем значимости 0,05.

Обратите внимание на *P-Значение* в ячейке F26. Это число 0,0001 соответствует уровню значимости, для которого наблюденное значение *F*-критерия 14,3571 является критическим. Другими словами, различие в производительности труда значимо с уровнем 0,0001.

Так как *P-Значение* меньше уровня значимости 0,05, то различие производительности труда в бригадах значимо.

Двухфакторный дисперсионный анализ без повторений

Пусть на значения случайной величины X оказывают влияние два фактора A и B, причем эти факторы друг с другом не взаимодействуют, но влияют на математическое ожидание величины X и не влияют на её дисперсию.

Имеется выборка $\{x_{ij},\ i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, l\}$ значений признака X, полученных при уровнях A_i и B_j факторов A и B. Каждой паре A_i и B_j значений факторов соответствует одно значение x_{ij}.

Необходимо проверить две гипотезы:

H_0^A : Фактор A не влияет на значение признака X;

H_0^B : Фактор B не влияет на значение признака X.

Пример 1.2 При выращивании помидоров на тридцати участках применялись пять сортов семян и шесть технологий выращивания. В табл.1.2 приведены показатели урожайности помидоров. Влияют ли факторы (сорт семян и технология выращивания) на урожайность продукции?

Таблица 1.2

	Технология 1	Технология 2	Технология 3	Технология 4	Технология 5	Технология 6
Сорт А	138	131	126	135	134	142
Сорт Б	128	144	124	128	134	126
Сорт В	130	126	129	135	136	136
Сорт Г	145	134	144	140	148	145
Сорт Д	120	140	130	125	129	127

Решение. Введите данные из табл.1.2 в программе Excel, как показано на рис.1.3.

	A	B	C	D	E	F	G
1		Технология 1	Технология 2	Технология 3	Технология 4	Технология 5	Технология 6
2	Сорт А	138	131	126	135	134	142
3	Сорт Б	128	144	124	128	134	126
4	Сорт В	130	126	129	135	136	136
5	Сорт Г	145	134	144	140	148	145
6	Сорт Д	120	140	130	125	129	127

Рис.1.3

Войти в меню «Анализ данных», выберите процедуру «Двухфакторный дисперсионный анализ без повторений», и в открывшемся окне (рис.1.4) введите в строке «Входной интервал» диапазон A1:G6, поставьте флажок в строке «Метки», выберите в строке «Альфа» уровень значимости 0,05 и нажмите «Ок».

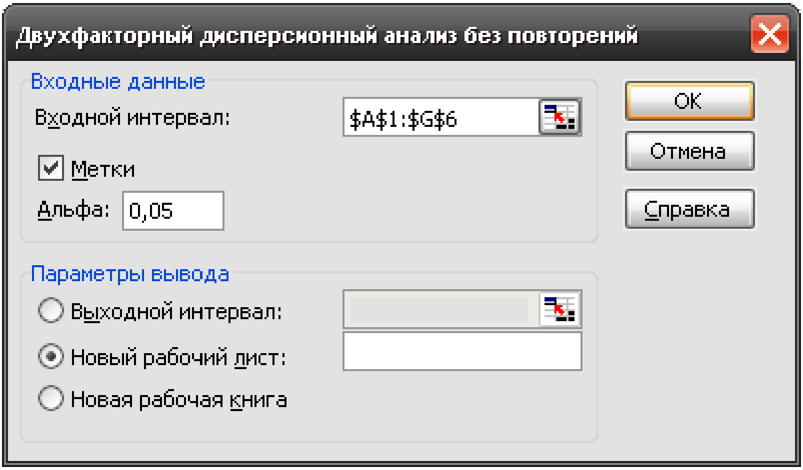

Рис.1.4

Результаты приведены в табл. 1.3.

Таблица 1.3

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ без повторений						
2							
3	*ИТОГИ*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
4	Сорт А	6	806	134,33	30,67		
5	Сорт Б	6	784	130,67	53,87		
6	Сорт В	6	792	132	18		
7	Сорт Г	6	856	142,67	24,67		
8	Сорт Д	6	771	128,5	44,3		
9							
10	Технология 1	5	661	132,2	92,2		
11	Технология 2	5	675	135	51		
12	Технология 3	5	653	130,6	61,8		
13	Технология 4	5	663	132,6	36,3		
14	Технология 5	5	681	136,2	50,2		

15	Технология 6	5	676	135,2	73,7		
16							
17							
18	Дисперсионный анализ						
19	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
20	Строки	719,47	4	179,8667	4,8525	0,0067	2,8661
21	Столбцы	116,17	5	23,2333	0,6268	0,6813	2,7109
22	Погрешность	741,33	20	37,0667			
23							
24	Итого	1576,97	29				

Вывод. Мы видим, что наблюденное значение F-критерия больше критического значения для строк (ячейки E20 и G20). В ячейке F20 *P-Значение* равно 0,0067 и оно меньше уровня значимости 0,05 (это следует из предыдущего). Фактор «Сорт семян» *влияет* на урожайность помидоров.

А в строке «Столбцы» наблюденное значение F-критерия меньше критического значения для строк (ячейки E21 и G21). (В ячейке F21 *P-Значение* 0,6813 больше уровня значимости 0,05.) Фактор «Технология выращивания» *не влияет* на урожайность помидоров.

Двухфакторный дисперсионный анализ с повторениями

Пусть на значения случайной величины X оказывают влияние два фактора A и B, причем эти факторы взаимодействуют друг с другом, влияют на математическое ожидание величины X и не влияют на её дисперсию.

Имеется выборка $\{x_{ijk}$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, q\}$ значений признака X, полученных при уровнях A_i и B_j факторов A и B. Каждой паре (A_i, B_j) значений факторов A и B соответствует q значений признака x_{ijk}, $k = 1, 2, \ldots, q$.

Необходимо проверить три гипотезы:

H_0^A : Фактор A не влияет на значение признака X;

H_0^B : Фактор B не влияет на значение признака X;

H_0^{AB} : Взаимодействие факторов *A* и *B* не влияет на значение признака *X*.

Пример 1.3 При выращивании помидоров на тридцати участках применялись пять видов удобрений и шесть технологий выращивания. Каждый участок был разбит на четыре делянки, т.е. каждой паре уровней факторов (вид удобрений, технология выращивания) соответствуют четыре значения показателя урожайности (табл.1.4). Влияют ли факторы (вид удобрений и технология выращивания) на урожайность продукции?

Таблица 1.4

	Технология 1	Технология 2	Технология 3	Технология 4	Технология 5	Технология 6
Вид А	133	142	134	140	140	144
	147	133	148	132	146	145
	137	141	127	142	136	144
	128	124	138	134	131	134
Вид Б	127	137	120	127	127	127
	130	123	128	126	122	138
	131	131	146	125	144	126
	132	141	144	124	122	122
Вид В	121	122	149	127	120	127
	128	145	131	127	129	125
	137	145	144	142	146	139
	128	131	125	141	125	148
Вид Г	131	137	123	127	123	124
	123	138	144	146	135	125
	139	136	122	122	142	139
	137	127	131	120	129	125
Вид Д	149	148	132	143	131	127
	136	141	120	128	142	135
	129	149	125	135	131	145
	135	129	146	126	130	133

Решение. Введите данные из табл.1.4 в программе *Excel*, как показано на рис.1.5.

	A	B	C	D	E	F	G
1		Техноло гия 1	Техноло гия 2	Техноло гия 3	Техноло гия 4	Техноло гия 5	Технол огия 6
2	Вид А	133	142	134	140	140	144
3		147	133	148	132	146	145
4		137	141	127	142	136	144
5		128	124	138	134	131	134
6	Вид Б	127	137	120	127	127	127
7		130	123	128	126	122	138
8		131	131	146	125	144	126
9		132	141	144	124	122	122
10	Вид В	121	122	149	127	120	127
11		128	145	131	127	129	125
12		137	145	144	142	146	139
13		128	131	125	141	125	148
14	Вид Г	131	137	123	127	123	124
15		123	138	144	146	135	125
16		139	136	122	122	142	139
17		137	127	131	120	129	125
18	Вид Д	149	148	132	143	131	127
19		136	141	120	128	142	135
20		129	149	125	135	131	145
21		135	129	146	126	130	133

Рис.1.5

Войти в меню «Анализ данных», выберите процедуру «Двухфакторный дисперсионный анализ с повторениями», и в открывшемся окне (рис.1.6) введите в строке «Входной интервал» диапазон A1:G21, в строке «Число строк для выборки» введите число 4, выберите в строке «Альфа» уровень значимости 0,05 и нажмите «Ок».

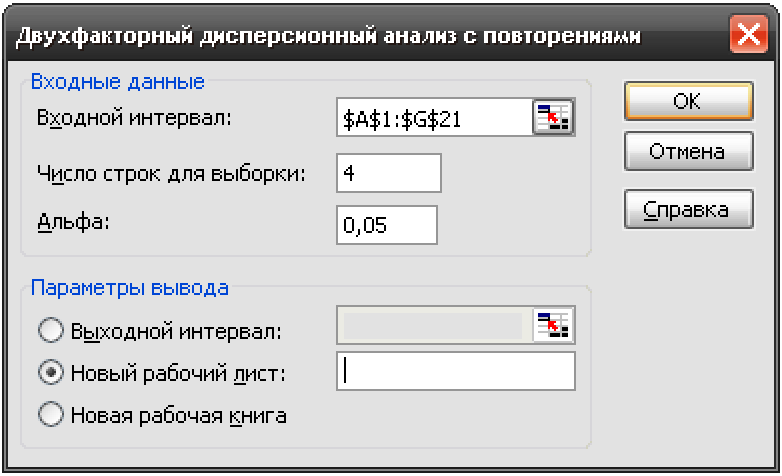

Рис.1.6

Результаты приведены в таблице 1.5.

Таблица 1.5

	A	B	C	D	E	F	G	H
1	Двухфакторный дисперсионный анализ с повторениями							
2								
3	ИТОГИ	Технология 1	Технология 2	Технология 3	Технология 4	Технология 5	Технология 6	Итого
4	*Вид А*							
5	Счет	4	4	4	4	4	4	24
6	Сумма	545	540	547	548	553	567	3300
7	Среднее	136,25	135	136,75	137	138,25	141,75	137,5
8	Дисперсия	64,92	70,00	76,92	22,67	40,25	26,92	44,09
9								
10	*Вид Б*							
11	Счет	4	4	4	4	4	4	24
12	Сумма	520	532	538	502	515	513	3120

13	Сред-нее	130	133	134,5	125,5	128,75	128,25	130
14	Дис-пер-сия	4,67	61,33	158,33	1,67	108,92	46,92	59,22
15								
16	*Вид В*							
17	Счет	4	4	4	4	4	4	24
18	Сумма	514	543	549	537	520	539	3202
19	Сред-нее	128,50	135,75	137,25	134,25	130,00	134,75	133,42
20	Дис-пер-сия	43,00	127,58	124,25	70,25	127,33	116,25	89,56
21								
22	*Вид Г*							
23	Счет	4	4	4	4	4	4	24
24	Сумма	530	538	520	515	529	513	3145
25	Сред-нее	132,50	134,50	130,00	128,75	132,25	128,25	131,04
26	Дис-пер-сия	51,67	25,67	103,33	140,92	66,25	51,58	62,48
27								
28	*Вид Д*							
29	Счет	4	4	4	4	4	4	24
30	Сумма	549	567	523	532	534	540	3245
31	Сред-нее	137,25	141,75	130,75	133,00	133,50	135,00	135,21
32	Дис-пер-сия	70,92	84,92	127,58	59,33	32,33	56,00	69,22
33								

34	*Итого*							
35	Счет	20	20	20	20	20	20	
36	Сумма	2658	2720	2677	2634	2651	2672	
37	Сред-нее	132,9	136	133,85	131,7	132,55	133,6	
38	Дис-пер-сия	49,36	67,89	102,66	64,12	70,68	73,73	
39								
40								
41	Дисперсионный ана-лиз							
42	*Ис-точ-ник вари-ации*	*SS*	*df*	*MS*	*F*	*P-Зна-чение*	*F кри-тиче-ское*	
43	Вы-борка	892,72	4	223,1792	3,0959	0,0195	2,4729	
44	Столб цы	217,17	5	43,4333	0,6025	0,6981	2,3157	
45	Взаи-мо-дей-ствие	759,58	20	37,9792	0,5268	0,9479	1,6883	
46	Внут-ри	6488	90	72,0889				
47								
48	Итого	8357,5	119					

Вывод. Сравнивая наблюденные значения *F* (ячейки F43:F45) с критическими (ячейки H43:H45), мы видим, что вид удобрений влияет на урожайность, а технология выращивания и взаимодействие этих факторов не влияют.

Задания для самостоятельной работы

1. Определить с уровнем α = 0,05 значимость различия производительности труда в двух бригадах рабочих за десять дней работы (табл. 1.6, за каждый день приведено среднее число изготовленных за час деталей на одного рабочего и число рабо-

тавших в этот день рабочих в бригаде).

Таблица 1.6

Вариант 1										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	13	14	14	15	14	14	15	14	16	16
Число работавших	7	8	6	6	8	6	7	8	7	7
2-я бригада	15	17	16	13	14	14	17	15	15	15
Число работавших	5	6	4	6	4	5	4	5	4	5
Вариант 2										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	13	14	15	13	16	13	14	16	15	14
Число работавших	6	7	6	7	6	7	8	6	6	7
2-я бригада	13	16	16	15	16	13	15	17	14	15
Число работавших	5	6	5	4	6	4	6	5	6	5
Вариант 3										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	13	14	16	13	13	14	13	14	16	13
Число работавших	7	8	7	8	7	7	8	7	6	8
2-я бригада	16	13	14	17	13	17	13	14	14	14
Число работавших	4	5	6	4	5	4	6	6	5	4
Вариант 4										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	16	15	16	16	13	16	16	14	13
Число работавших	7	7	8	6	7	7	6	7	6	6
2-я бригада	16	14	13	16	15	15	15	16	15	16
Число работавших	6	4	4	4	5	5	4	6	4	4
Вариант 5										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	13	16	14	16	13	14	16	16	15
Число работавших	6	8	7	8	8	6	6	8	7	6
2-я бригада	17	14	16	13	16	15	17	13	13	17
Число работавших	6	5	5	4	6	5	6	5	6	5
Вариант 6										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	13	13	13	14	15	16	15	14	16	14
Число работавших	7	8	6	6	7	6	7	8	7	6

2-я бригада	13	15	13	17	15	13	15	17	15	17
Число работавших	4	4	5	6	4	4	5	5	5	5
Вариант 7										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	13	16	14	16	16	13	16	15	16
Число работавших	6	6	7	8	6	6	8	7	7	7
2-я бригада	17	15	17	13	13	17	14	15	13	13
Число работавших	6	4	5	5	6	4	6	4	4	6
Вариант 8										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	16	13	15	14	13	16	15	14	14
Число работавших	6	6	8	7	8	6	8	8	8	6
2-я бригада	13	16	16	14	13	17	13	16	14	13
Число работавших	6	5	4	6	4	6	5	4	5	4
Вариант 9										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	16	13	16	14	15	16	15	13	14	16
Число работавших	8	7	8	8	8	7	6	6	8	7
2-я бригада	14	16	17	16	16	13	13	16	17	14
Число работавших	4	6	6	4	4	5	6	6	6	5
Вариант 10										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	16	13	13	15	16	14	15	13	15	14
Число работавших	6	6	7	6	7	8	8	6	6	7
2-я бригада	14	15	13	17	13	14	15	14	14	16
Число работавших	4	4	6	6	4	6	5	6	4	4

2. Определить с уровнем $\alpha = 0{,}05$ значимость различия производительности труда в трех бригадах рабочих-токарей за десять дней работы (табл. 1.7, за каждый день приведено среднее число изготовленных за час деталей на одного рабочего).

Таблица 1.7

Вариант 1										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	13	13	16	16	13	15	15	15	14	14
2-я бригада	16	16	13	13	13	15	15	16	15	16
3-я бригада	17	15	15	16	13	18	17	15	18	14
Вариант 2										

День	1	2	3	4	5	6	7	8	9	10
1-я бригада	13	13	16	14	14	16	15	15	13	16
2-я бригада	17	16	17	17	13	16	15	13	13	16
3-я бригада	17	13	16	17	16	18	13	18	16	15
Вариант 3										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	13	15	15	16	13	13	13	13	16
2-я бригада	17	16	16	14	15	15	15	14	15	17
3-я бригада	15	13	13	14	15	15	18	15	14	14
Вариант 4										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	13	13	15	14	15	13	13	14	16
2-я бригада	17	13	13	16	13	14	13	14	16	15
3-я бригада	14	15	17	18	17	18	16	16	18	15
Вариант 5										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	15	14	14	15	16	13	13	16	13	15
2-я бригада	16	14	13	13	14	14	14	16	13	15
3-я бригада	16	13	15	13	13	13	13	13	14	15
Вариант 6										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	16	14	16	16	16	16	16	16	16	16
2-я бригада	17	15	15	15	15	17	15	14	14	17
3-я бригада	13	13	13	16	14	15	18	16	15	18
Вариант 7										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	16	14	16	13	13	13	14	13	14
2-я бригада	15	17	13	16	16	15	15	13	14	16
3-я бригада	14	16	18	15	18	18	15	16	16	18
Вариант 8										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	15	14	16	15	16	15	15	14	13	16
2-я бригада	14	17	15	17	16	14	17	13	15	16
3-я бригада	15	14	17	18	18	15	17	17	15	16
Вариант 9										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	13	13	15	15	14	14	14	14	14	15

2-я бригада	15	15	17	17	14	16	15	17	15	17
3-я бригада	17	17	17	13	17	17	13	16	17	18
Вариант 10										
День	1	2	3	4	5	6	7	8	9	10
1-я бригада	14	13	16	13	15	15	14	14	15	14
2-я бригада	13	17	15	13	15	13	17	17	16	14
3-я бригада	15	13	15	17	15	15	18	16	17	15

3. При подготовке к соревнованиям двадцати спортсменов-многоборцев, имевших близкие спортивные результаты, применялись четыре рациона питания и четыре методики тренировок. В табл.1.8 приведены показатели в баллах, полученные спортсменами на соревнованиях. Влияют ли факторы (рацион питания и методика тренировок) на достижения спортсменов?

Таблица 1.8

Вариант 1				
	методика 1	методика 2	методика 3	методика 4
рацион А	1187	1139	1153	1174
рацион Б	1080	1220	1191	1067
рацион В	1101	1267	1220	1096
рацион Г	1134	1151	1254	1216
Вариант 2				
	методика 1	методика 2	методика 3	методика 4
рацион А	1121	1008	1160	1149
рацион Б	1209	1106	1169	1223
рацион В	1132	1125	1245	1283
рацион Г	1182	1145	1212	1110
Вариант 3				
	методика 1	методика 2	методика 3	методика 4
рацион А	1012	1121	1094	1181
рацион Б	1089	1081	1069	1182
рацион В	1277	1288	1225	1092
рацион Г	1233	1231	1196	1283
Вариант 4				
	методика 1	методика 2	методика 3	методика 4
рацион А	1026	1096	1037	1118
рацион Б	1218	1105	1199	1081
рацион В	1137	1110	1241	1170

рацион Г	1289	1236	1209	1170
Вариант 5				
	методика 1	методика 2	методика 3	методика 4
рацион А	1184	1019	1117	1024
рацион Б	1117	1116	1183	1163
рацион В	1205	1133	1247	1103
рацион Г	1140	1248	1233	1118
Вариант 6				
	методика 1	методика 2	методика 3	методика 4
рацион А	1070	1106	1162	1175
рацион Б	1142	1080	1186	1065
рацион В	1107	1218	1258	1265
рацион Г	1191	1298	1199	1120
Вариант 7				
	методика 1	методика 2	методика 3	методика 4
рацион А	1144	1134	1149	1185
рацион Б	1143	1239	1167	1103
рацион В	1165	1226	1197	1135
рацион Г	1119	1170	1136	1285
Вариант 8				
	методика 1	методика 2	методика 3	методика 4
рацион А	1077	1101	1177	1172
рацион Б	1200	1156	1168	1107
рацион В	1272	1193	1274	1093
рацион Г	1194	1115	1254	1108
Вариант 9				
	методика 1	методика 2	методика 3	методика 4
рацион А	1112	1130	1093	1017
рацион Б	1159	1218	1225	1162
рацион В	1115	1266	1247	1190
рацион Г	1266	1145	1223	1214
Вариант 10				
	методика 1	методика 2	методика 3	методика 4
рацион А	1093	1061	1168	1018
рацион Б	1243	1185	1095	1210
рацион В	1153	1223	1137	1181
рацион Г	1182	1241	1222	1157

4. При подготовке к соревнованиям спортсменов-многоборцев, имевших близкие спортивные результаты, применялись два рациона питания и пять методик тренировок, причем каждой паре (рацион питания, методика тренировок) соответствует четыре спортсмена. В табл.1.9 приведены показатели в баллах, полученные спортсменами на соревнованиях. Влияют ли факторы (рацион питания и методика тренировок) на достижения спортсменов?

Таблица 1.9

Вари-ант 1					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1047	1162	1105	1167	1196
	1108	1019	1049	1147	1080
	1002	1023	1062	1104	1130
	1053	1069	1124	1027	1116
рацион Б	1093	1070	1219	1199	1080
	1144	1149	1208	1072	1148
	1189	1102	1097	1152	1230
	1083	1205	1233	1216	1119
Вари-ант 2					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1084	1117	1062	1104	1156
	1025	1052	1109	1173	1033
	1169	1123	1103	1187	1040
	1027	1066	1051	1060	1048
рацион Б	1062	1117	1059	1216	1238
	1208	1237	1189	1198	1089
	1118	1101	1179	1222	1076
	1203	1164	1226	1205	1232
Вари-ант 3					

	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1165	1080	1182	1144	1165
	1096	1091	1111	1109	1051
	1069	1033	1174	1186	1053
	1138	1140	1001	1101	1091
рацион Б	1123	1051	1196	1111	1073
	1089	1195	1218	1112	1116
	1222	1099	1063	1234	1098
	1117	1098	1186	1070	1105
Вари-ант 1					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1065	1150	1166	1123	1133
	1150	1076	1036	1029	1160
	1098	1012	1035	1031	1038
	1083	1128	1114	1161	1181
рацион Б	1229	1143	1152	1240	1214
	1134	1109	1148	1114	1084
	1055	1209	1070	1128	1141
	1113	1181	1183	1130	1188
Вари-ант 4					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1005	1168	1028	1133	1093
	1099	1098	1111	1145	1007
	1196	1022	1068	1061	1147
	1114	1051	1015	1073	1019
рацион Б	1142	1085	1198	1200	1092
	1102	1094	1214	1200	1234
	1164	1137	1060	1217	1204

	1237	1229	1101	1053	1167
Вари-ант 5					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1089	1018	1032	1017	1116
	1005	1189	1045	1082	1110
	1025	1194	1102	1026	1055
	1146	1185	1058	1102	1087
рацион Б	1236	1168	1204	1206	1105
	1113	1128	1117	1119	1059
	1150	1204	1211	1249	1116
	1122	1234	1232	1153	1199
Вари-ант 6					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1190	1103	1147	1197	1040
	1161	1065	1052	1003	1123
	1105	1145	1031	1152	1095
	1042	1111	1087	1002	1157
рацион Б	1113	1236	1071	1128	1244
	1161	1164	1071	1219	1167
	1110	1135	1050	1188	1088
	1206	1150	1116	1160	1137
Вари-ант 7					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1168	1014	1035	1182	1124
	1118	1019	1186	1087	1118
	1050	1030	1165	1019	1127
	1051	1179	1016	1053	1076

рацион Б	1108	1110	1108	1158	1091
	1089	1119	1124	1141	1097
	1088	1230	1050	1182	1203
	1081	1072	1109	1055	1158
Вари-ант 8					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1053	1103	1171	1026	1123
	1117	1127	1012	1175	1031
	1082	1104	1095	1114	1148
	1138	1150	1170	1028	1041
рацион Б	1103	1196	1165	1114	1232
	1088	1162	1225	1201	1235
	1211	1164	1114	1112	1050
	1059	1096	1192	1226	1177
Вари-ант 9					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1040	1072	1120	1078	1147
	1116	1059	1015	1007	1143
	1198	1016	1007	1173	1056
	1086	1032	1178	1185	1016
рацион Б	1064	1103	1171	1150	1126
	1057	1177	1091	1056	1174
	1136	1199	1182	1194	1135
	1072	1128	1172	1154	1141
Вари-ант 10					
	методика 1	методика 2	методика 3	методика 4	методика 5
рацион А	1010	1155	1063	1131	1157

	1157	1073	1182	1017	1118
	1100	1115	1162	1198	1195
	1040	1060	1029	1140	1170
рацион Б	1095	1082	1058	1054	1158
	1173	1075	1243	1163	1053
	1056	1107	1133	1175	1097
	1234	1175	1075	1223	1057

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Корреляционный анализ позволяет определить наличие связей между случайными величинами, вычислив коэффициент корреляции, и оценить силу связи.

Пусть рассматриваются две случайные величины *X* и *Y*.

Если каждому значению случайной величины *X* соответствует единственное значение случайной величины *Y*, то связь между *X* и *Y* называется *функциональной зависимостью*.

Если каждому значению случайной величины *X* соответствует закон распределения вероятностей случайной величины *Y*, то связь между *X* и *Y* называется *вероятностной (стохастической) зависимостью*.

Случайные величины *X* и *Y* называются *независимыми*, если закон распределения вероятностей случайной величины *Y* не зависит от того, какое значение приняла случайная величина *X*. В противном случае величины *X* и *Y* называются *зависимыми*.

Если математическое ожидание *M*(*Y*) случайной величины *Y* зависит от того, какое значение *x* приняла случайная величина *X*, то такая зависимость называется *корреляционной*. В этом случае говорят об условном математическом ожидании $\bar{y}_x = M\left(Y \mid X = x\right)$.

Уравнением регрессии Y по X называется функция $\bar{y}_x = \varphi\left(x\right)$, выражающая зависимость математического ожидания случайной величины *Y* от значения случайной величины *X*. График этой функции называется *линией регрессии Y по X*.

Линейное уравнение регрессии Y по X имеет вид $\bar{y}_x = \beta_0 + \beta_1 x$.

Основной задачей корреляционного анализа является выявление *корреляционной связи* между случайными переменными *X* и *Y*.

Числовой характеристикой *линейной корреляционной связи* между случайными величинами *X* и *Y* является *коэффициент корреляции*

$$\rho = \frac{\operatorname{cov}\left(X, Y\right)}{\sigma_x \sigma_y} = \frac{M\left(X - \bar{x}, Y - \bar{y}\right)}{\sigma_x \sigma_y},$$

где $\bar{x}$, $\bar{y}$ — математические ожидания, а σ_x, σ_y — среднеквадратические отклонения случайных величин *X* и *Y*.

Свойства коэффициента корреляции:

1) $|\rho| \leq 1$.

Если $\rho < 0$, то при возрастании одной из случайных величин, условное математическое ожидание другой убывает.

Если $\rho > 0$, то возрастание одной из случайных величин ведет к возрастанию условного математического ожидания другой.

2) Если случайные величины X и Y независимы, то $\rho = 0$. (Обратное утверждение неверно).

3) Если $|\rho| = 1$, то между X и Y существует линейная зависимость

$$\bar{y}_x = \beta_0 + \beta_1 x.$$

Справедливо и обратное утверждение: если между X и Y существует линейная зависимость, то $|\rho| = 1$.

Для выявления *нелинейной корреляционной зависимости* между X и Y используются *корреляционные отношения*.

Корреляционное отношение Y по X определяется формулой:

$$\eta_{yx} = \sqrt{\frac{D\left[M\left(Y \mid X\right)\right]}{\sigma_y}}.$$

Корреляционное отношение X по Y определяется аналогично:

$$\eta_{xy} = \sqrt{\frac{D\left[M\left(X \mid Y\right)\right]}{\sigma_x}}.$$

Свойства корреляционного отношения:

1) $\eta_{yx} \neq \eta_{xy}$. Корреляционное отношение не симметрично.

2) $0 \leq |\rho| \leq \eta_{yx} \leq 1$.

3) Если случайные величины X и Y независимы, то $\eta_{yx} = \eta_{xy} = 0$. Обратное утверждение неверно.

Если $\eta_{yx} = 0$, то Y некоррелирована с X, но при этом может быть *X* коррелирована с *Y*: $\eta_{xy} \neq 0$. Возможны случаи, когда $\eta_{yx} = 0$ и $\eta_{xy} = 1$.

4) Условие $\eta_{yx} = \eta_{xy} = 1$ равносильно существованию функциональной связи между *X* и *Y*.

Если $|\rho| = \eta_{yx} = \eta_{xy} = 1$, то между *X* и *Y* существует *линейная функциональная зависимость.*

Если $|\rho| < \eta_{yx} = \eta_{xy} = 1$, то между *X* и *Y* существует *нелинейная функциональная зависимость.*

Условие $0 < |\rho| = \eta_{yx} = \eta_{xy} < 1$ равносильно существованию *линейной корреляционной зависимости* между *X* и *Y*.

Если корреляция между *X* и *Y* *нелинейна*, то $|\rho| < \min(\eta_{yx}, \eta_{xy})$.

Выборочный коэффициент корреляции и проверка его значимости

Выборочный коэффициент корреляции позволяет оценить тесноту *корреляционной линейной связи* между признаками *X* и *Y*.

Пусть в результате наблюдений получена таблица парных значений (x_i, y_i), $i = 1, \ldots, n$.

При описании тесноты линейной корреляционной связи между признаками *X* и *Y* обычно пользуются следующими выводами о силе связи в зависимости от значения *значимого коэффициента корреляции*:

$0 < \lvert r_в \rvert < 0{,}2$	Очень слабая зависимость
$0{,}2 \le \lvert r_в \rvert < 0{,}4$	Слабая зависимость
$0{,}4 \le \lvert r_в \rvert < 0{,}7$	Средняя зависимость
$0{,}7 \le \lvert r_в \rvert < 0{,}9$	Сильная зависимость
$0{,}9 \le \lvert r_в \rvert < 1$	Очень сильная зависимость

Корреляционное поле — это графическое изображение на координатной плоскости точек выборки (x_i, y_i).

Корреляционная таблица представляет выборку из парных значений
(x_i, y_i) в сгруппированном виде.

Общий вид корреляцион- ной таблицы представлен ниже (табл.2.1).

Таблица 2.1

x_i \ y_j	y_1^*	y_2^*	…	y_l^*	n_{xi}	Групповая средняя $\overline{y}_i$
x_1^*	$n_{1,1}$	$n_{1,2}$	…	$n_{1,l}$	n_{x1}	$\overline{y}_1$
x_2^*	$n_{2,1}$	$n_{2,2}$	…	$n_{2,l}$	n_{x2}	$\overline{y}_2$
…	…	…	…	…	…	…
x_m^*	$n_{m,1}$	$n_{m,2}$	…	$n_{m,l}$	n_{xm}	$\overline{y}_m$
n_{yj}	n_{y1}	n_{y2}	…	n_{yl}		
Групповая средняя $\overline{x}_j$	$\overline{x}_1$	$\overline{x}_2$	…	$\overline{x}_l$	–	–

Корреляционное отношение и проверка гипотезы об отсутствии корреляционной зависимости между признаками

По таблице распределения Фишера-Снедекора (или с помощью функции FРАСПОБР() программы *Excel*) находим критическое значение $F_{\alpha;m-1;n-m}$ значение *F*-критерия Фишера-Снедекора с уровнем значимости α с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы.

Если $F > F_{\alpha;m-1;n-m}$, то принимается гипотеза о наличии корреляционной зависимости между *Y* и *X*.

Если $F > F_{\alpha;l-1;n-l}$, то принимается гипотеза о наличии корреляционной зависимости между *X* и *Y*.

Для вычисления критического значения $F_{\alpha;m-1;n-m}$ в программе Excel предназначена функция

FРАСПОБР(вероятность;степени_свободы1;степени_свободы2)

Вероятность — уровень значимости α.

Степени_свободы1 — число степеней свободы $m-1$.
Степени_свободы2 — число степеней свободы $n-m$.

Пример 2.1 Задана таблица парных значений (x_i, y_i) (табл. 2.2).

1) Вычислить выборочный коэффициент корреляции.

2) Проверить значимость коэффициента корреляции для уровня
α = 0,01.

3) Построить корреляционное поле.

Таблица 2.2

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	0,43	2,66	11	0,32	2,52	21	0,52	3,54
2	0,25	2,09	12	0,79	4,11	22	0,38	2,27
3	0,52	2,89	13	0,70	4,52	23	0,32	2,04
4	0,63	4,13	14	0,95	4,90	24	0,34	2,48
5	0,01	0,26	15	0,60	3,81	25	0,28	2,14
6	0,91	5,01	16	0,98	5,68	26	0,91	5,04
7	0,41	2,44	17	0,75	4,03	27	0,90	4,97
8	0,45	3,10	18	0,69	4,08	28	0,60	3,19
9	0,05	0,89	19	0,46	2,57	29	0,45	2,45
10	0,44	2,29	20	0,63	3,49	30	0,84	4,98

Решение. 1) Введите исходные данные в диапазоне A1:C31 (рис. 2.1, показаны первые десять строк из тридцати).

Войти в меню «Анализ данных», и в появившемся окне выберем функцию «Корреляция». Появится окно, показанное на рис. 2.2. В этом окне укажем входной интервал B1:C31, группирование «по столбцам», поставим флажок в строке «Метки в первой строке», в параметрах вывода выберем «Новый рабочий лист» и нажмем кнопку «ОК». Получим таблицу, показанную на рис. 2.3.

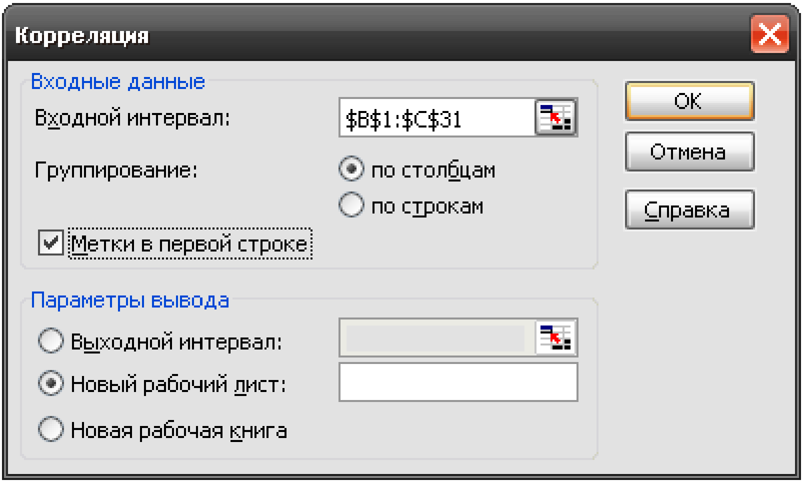

Рис. 2.2

Функция «Корреляция» из пакета «Анализ данных» предназначена для вычисления корреляционной матрицы, которая содержит коэффициенты корреляции r_{ij} между всевозможными парами (X_i, X_j), где X_1, X_2, …, X_k — признаки, между которыми исследуется статистическая связь. Функция «Корреляция» выдает симметричную корреляционную матрицу, диагональные элементы которой равны единице.

	A	B	C
1		x	y
2	x	1	
3	y	0,975718	1

Рис. 2.3

На рис. 2.3 коэффициент корреляции содержится в ячейке B3 .

2) Для проверки значимости коэффициента корреляции вычислим значение $T_{набл}$.

Для этого введите в ячейку D7 формулу

=D3*КОРЕНЬ((30-2)/(1-D3^2)).

Получим $T_{набл}$ = 23,57.

В ячейку D9 введите формулу =СТЬЮДРАСПОБР(0,01;28), получим значение критерия Стьюдента 2,763262442.

Так как $T_{набл} = 23{,}57 > t_{кр}(\alpha;\ k) = 2{,}7$, то делаем вывод: коэффициент корреляции значим с уровнем доверия 99% $(1 - 0{,}01 = 0{,}99)$.

Значение выборочного коэффициента корреляции положительно и близко к единице, что означает *очень сильную прямую* связь между рассматриваемыми признаками.

3) Выделим диапазон ячеек B1:C31 и с помощью мастера диаграмм построим диаграмму «Точечная». Получим график, изображенный на рис. 2.4.

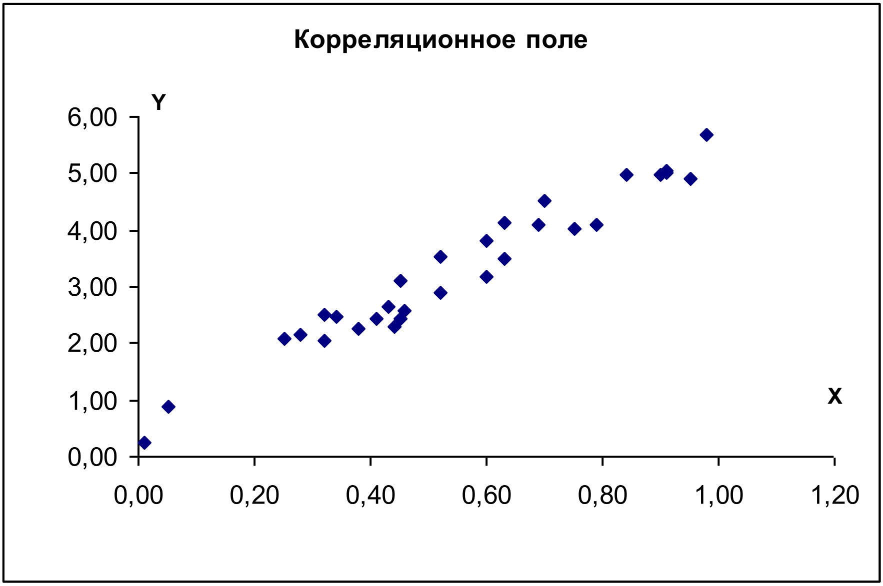

Рис. 2.4

Из этого графика можно сделать вывод о том, что между X и Y есть линейная корреляционная зависимость, так как точки расположены близко к некоторой воображаемой прямой. Этот вывод подтверждается значением выборочного коэффициента корреляции.

Замечание. Мы не можем в данном примере вычислить выборочное корреляционное отношение, так как данные не сгруппированы в корреляционную таблицу.

Пример 2.2 Дана корреляционная таблица (табл. 2.3).

Таблица 2.2

y_j / x_i	9	13	17	21	25
22,5	2	1			
27,5	3	6	4		
32,5		3	11	7	
37,5		1	2	6	2
42,5				1	1

1) Вычислить значение выборочного коэффициента корреляции и проверить значимость для уровня α = 0,05.

2) Вычислить выборочные корреляционные отношения η^2_{yx} и η^2_{yx}, и проверить гипотезу об отсутствии корреляционной зависимости между *Y* и *X*.

Решение. 1) Введите корреляционную таблицу в программе *Excel* в диапазоне A1:F6.

В ячейках G1, H1 и A7, A8 введите обозначения, как показано на рис. 2.5.

В ячейку G2 введите формулу =СУММ(B2:F2) и затем протянем маркером заполнения ячейку G2 до G6. Получим суммы частот n_{xi}.

В ячейку G7 введем формулу =СУММ(G2:G6). Получим объем выборки $n = 50$.

В ячейку B7 введите формулу =СУММ(B2:B6) и затем протяните маркером заполнения ячейку B7 до F7. Получим суммы частот n_{yj}.

В ячейку H2 введите формулу

=СУММПРОИЗВ(B2:F2;B$1:F$1)/G2

и затем протяните маркером заполнения ячейку H2 до H6. В диапазоне H2:H6 получим групповые средние $\overline{y}_i$.

В ячейку B8 введите формулу

=СУММПРОИЗВ(B2:B6;$A2:$A6)/B7

и затем протяните маркером заполнения ячейку B8 вправо до F2. В диапазоне B8:F8 получим групповые средние $\overline{x}_j$.

	A	B	C	D	E	F	G	H
1		9	13	17	21	25	n_{xi}	Группов. ср. Yx
2	22,5	2	1				3	10,33
3	27,5	3	6	4			13	13,31
4	32,5		3	11	7		21	17,76
5	37,5		1	2	6	2	11	20,27
6	42,5				1	1	2	23,00
7	n_{yj}	5	11	17	14	3	50	
8	Групповая ср. Xy	25,50	29,32	31,91	35,36	39,17		
9	Коэфф. корреляции	0,74	7,62	2,01				
10		0,56	6,32	2,58				
11	Корр. отнош. Y по X	0,75						
12		0,55	6,20	2,58				
13	Корр. отнош. X по Y	0,74						

Рис. 2.5

Теперь в ячейку А9 введите текст «Коэфф. корреляции», а в В9 — формулу (2.5) для вычисления выборочного коэффициента корреляции:

=(50*СУММПРОИЗВ(A2:A6;G2:G6;H2:H6)-
СУММПРОИЗВ(A2:A6;G2:G6)*СУММПРОИЗВ(B1:F1;B7:F7))/
(КОРЕНЬ(50*СУММПРОИЗВ(A2:A6^2;G2:G6)-
СУММПРОИЗВ(A2:A6;G2:G6)^2)*КОРЕНЬ(50*
СУММПРОИЗВ(B1:F1^2;B7:F7)-
СУММПРОИЗВ(B1:F1;B7:F7)^2))

и нажмите комбинацию клавиш Ctrl + Shift + Enter.

Получим значение $r = 0{,}74$.

Замечание. В приведенной формуле используются операции с массивами. Например, A2:A6^2 означает, что содержимое каждой ячейки диапазона A2:A6 будет возводиться в квадрат.

Для проверки значимости коэффициента корреляции вычислим значение $T_{набл}$. Для этого введите в ячейку С9 формулу

=B9*КОРЕНЬ((50-2)/(1-B9^2))

Получим $T_{набл} = 7{,}62$.

В ячейку D9 введите формулу =СТЬЮДРАСПОБР(0,05;48), получим значение критерия Стьюдента 2,01.

Так как $T_{набл}$ = > $t_{кр}(\alpha; k)$ делаем вывод: коэффициент корреляции значим с уровнем доверия 95%.

Значение выборочного коэффициента корреляции положительно и близко к единице, что означает *сильную прямую* связь между рассматриваемыми признаками.

2) Введите в ячейку B10 формулу

=СУММПРОИЗВ((H2:H6-СУММПРОИЗВ(B1:F1;B7:F7)/50)^2;G2:G6)/
СУММПРОИЗВ((B1:F1-СУММПРОИЗВ(B1:F1;B7:F7)/50)^2;B7:F7)

и нажмите комбинацию клавиш Ctrl + Shift + Enter.

В ячейку B11 введите формулу =КОРЕНЬ(B10). Получим значение η_{yx} = 0,749585091 ≈ 0,77.

Введите в ячейку C10 формулу =B10*45/4, а в ячейку D10 формулу =FРАСПОБР(0,05;4;45).

В ячейке C10 получим значение статистики F = 6,32, в ячейке D10 критическое значение $F_{\alpha;m-1;n-m} = F_{0,05;5-1;50-5} = 2,58$. Следовательно, можно утверждать, что между Y и X есть *корреляционная зависимость.*

Так как выборочное корреляционное отношение η_{yx} = 0,75 почти совпадает с выборочным коэффициентом корреляции $r_в$ = 0,74, мы можем заключить, что между Y и X есть *линейная корреляционная зависимость.*

Введите в ячейку B12 формулу для вычисления η_{xy}:

=СУММПРОИЗВ((B8:F8-СУММПРОИЗВ(A2:A6;G2:G6)/50)^2;B7:F7)/
СУММПРОИЗВ((A2:A6-СУММПРОИЗВ(A2:A6;G2:G6)/50)^2;G2:G6)

и нажмите комбинацию клавиш Ctrl + Shift + Enter.

В ячейку B13 введите формулу =КОРЕНЬ(B12).

Получим значение η_{xy} = 0,74.

Для проверки значимости введите в ячейку C12 формулу =B12*45/4. Получим значение 6,2.

Критическое значение то же самое $F_{\alpha;l-1;n-l} = F_{0,05;5-1;50-5} = 2,58$. Так как $F > F_{0,05;5-1;50-5}$, можно утверждать, что между X и Y есть *корреляционная зависимость.*

Мы видим, что $\eta_{xy} \neq \eta_{yx}$, но их значения близки значению коэффициента корреляции $r_в$ = 0,74, поэтому мы можем заключить, что между X и Y есть *линейная корреляционная зависимость*.

Пример 2.3. 1) Смоделировать выборку (x_i, y_i) объема n = 100, где x_i — значения нормально распределенной случайной величины с математическим ожиданием a = 2 и среднеквадратическим отклонением σ = 1, а $y_i = (x_i - 2)^3$.

2) Вычислить значение выборочного коэффициента корреляции и проверить значимость для уровня α = 0,05.

3) Вычислить выборочные корреляционные отношения η^2_{yx} и η^2_{yx}, и проверить гипотезу об отсутствии корреляционной зависимости между Y и X.

Решение. 1) Введите на рабочем листе Лист1 в ячейку A1 формулу =СЛЧИС(), в ячейку B1 — формулу =НОРМОБР(A1;2;1), в ячейку C1 — формулу =(A1-2)^3.

Выделите ячейки A1:C1 и протяните маркером заполнения вниз до строки A100:C100.

Выделите диапазон B1:C100 скопируйте в буфер, перейдите на Лист2, щелкните правой кнопкой мыши в ячейке A2, в контекстном меню выберите «Специальная вставка», в появившемся окне в разделе «Вставить» выбереите «значения» и нажмите «ОК».

Замечание. Функция СЛЧИС() пересчитывается при каждом обновлении листа, т.е. все значения диапазона A1:C100 рабочего листа Лист1 будут изменяться при изменениях на этом листе (при вводе формул и т.п.). Поэтому мы на листе Лист2 *зафиксировали* полученные случайные значения.

На рабочем листе Лист2 мы имеем искомую парную выборку (x_i, y_i) в диапазоне A2:B101 (рис. 2.6, показана только часть).

	A	B	C
1	X	Y	Коэфф. коррел
2	3,937518	7,273399	0,827244738
3	-0,09897	-9,24742	$T_{набл}$
4	3,60475	4,132589	14,57582176
5	1,36283	-0,25868	$t_{кр}(\alpha;\ k)$
6	3,904107	6,903575	1,984467404
7	2,529041	0,14807	
8	2,490822	0,118242	
9	1,040602	-0,88307	
10	1,59425	-0,0668	
11	0,561686	-2,97551	

Рис. 2.6

2) В ячейку C2 введите формулу =КОРРЕЛ(A2:A101;B2:B101). Получим значение 0,827244738.

Для проверки значимости коэффициента корреляции вычислите значение $T_{набл}$. Для этого введите в ячейку C4 формулу

=C2*КОРЕНЬ((100-2)/(1-C2^2)).

Получим $T_{набл}$ = 14,57582176.

В ячейку C6 введите формулу =СТЬЮДРАСПОБР(0,05;98).

Получим значение критерия Стьюдента 1,9845.

Так как $T_{набл} = > t_{кр}(\alpha;\ k)$ делаем вывод: коэффициент корреляции значим с уровнем доверия 95%.

Значение выборочного коэффициента корреляции показывает *сильную прямую корреляционную зависимость* между рассматриваемыми признаками.

На самом деле между *Y* и *X* существует функциональная связь. Но эта связь *нелинейная*, поэтому значение выборочного коэффициента корреляции не равно единице.

3) Для вычисления корреляционного отношения необходимо построить корреляционную таблицу. Для этого создадим программу-макрос по следующему алгоритму:

Выполните команду меню «Сервис — Макрос — Редактор *Visual Basic*», в открывшемся окне выполните команду меню «Insert-Module» и введите текст программы *KorTab*(*x*, *y*, *k*) на языке *Visual Basic*:

```
Option Explicit: Option Base 0
Function KorTab(x, y, k)
 Application.Volatile (False)
 Dim i, j, m, n, ns, ns1 As Integer:   Dim nxi(), nyj() As Integer
 Dim nij(), xi(), yj(), axj(), ayi() As Variant
 Dim minx, maxx, miny, maxy, hx, hy, s As Variant
 n = Application.Count(x)
 ReDim nij(k + 2, k + 2), xi(k), yj(k), axj(k), ayi(k), nxi(k), nyj(k)
 minx = Application.WorksheetFunction.Min(x):   maxx = Application.WorksheetFunction.Max(x)
 miny = Application.WorksheetFunction.Min(y):   maxy = Application.WorksheetFunction.Max(y)
 hx = (maxx - minx) / (k - 1): hy = (maxy - miny) / (k - 1)
 xi(0) = minx - hx / 2: yj(0) = miny - hy / 2
 For i = 1 To k:   xi(i) = xi(i - 1) + hx: yj(i) = yj(i - 1) + hy: Next i
 For i = 1 To k: For j = 1 To k:  For m = 1 To n
 If (xi(i - 1) < x(m)) And (x(m) <= xi(i)) And (yj(j - 1) < y(m)) And (y(m) <= yj(j)) Then nij(i, j) = nij(i, j) + 1
 Next m:  Next j:  Next i
 For i = 1 To k:  xi(i) = xi(i) - hx / 2: yj(i) = yj(i) - hy / 2:  Next i
 For j = 1 To k:  s = 0: nyj(j) = 0:  For i = 1 To k: nyj(j) = nyj(j) + nij(i, j)
      s = s + xi(i) * nij(i, j):   Next i:   axj(j) = s / nyj(j):  Next j
 For i = 1 To k: s = 0: nxi(i) = 0:  For j = 1 To k:
  nxi(i) = nxi(i) + nij(i, j): s = s + yj(j) * nij(i, j): Next j:  ayi(i) = s / nxi(i):  Next i
 ns = 0: ns1 = 0:  For i = 1 To k: ns = ns + nyj(i): ns1 = ns1 + nxi(i):  Next i
 For i = 1 To k:  nij(i, 0) = xi(i): nij(i, k + 1) = nxi(i): nij(i, k + 2) = ayi(i)
  nij(0, i) = yj(i): nij(k + 1, i) = nyj(i): nij(k + 2, i) = axj(i)
 Next i: nij(k + 1, k + 1) = ns: nij(k + 1, k + 2) = ns1
 KorTab = nij
End Function
```

Для контроля правильности текст программы приведен ниже на рис.2.8.

Для вызова программы необходимо: 1) выделить диапазон ячеек, содержащий (k + 3) строки, (k + 3) столбца; 2) в строке формул ввести:

=KorTab(Диапазон_X; Диапазон_Y;Число_интервалов_k)

и, удерживая нажатыми клавиши Ctrl и Shift, нажать Enter.

Диапазон_X — диапазон ячеек, содержащий значения X;

Диапазон_Y — диапазон ячеек, содержащий значения Y;

Число_интервалов_k — число интервалов группировки выборки.

Программа строит квадратную матрицу частот $[n_{ij}]$ порядка k, находит середины интервалов x_i^* и y_j^*, условные средние $\overline{x}_j$ и $\overline{y}_i$, частоты n_{xi} и n_{yj}. Результаты выводятся в массив размера (k + 3) строки, (k + 3) столбца.

После ввода текста программы перейдите на Лист2, выделите диапазон ячеек D1:P13, в строке формул введите

=KorTab(A2:A101;B2:B101;10)

и нажатмите комбинацию клавиш Ctrl + Shift + Enter.

Мы получим корреляционную таблицу (рис.2.7). Здесь первый параметр соответствует массиву значений показателя X, второй параметр — массиву значений показателя Y, третий параметр — числу интервалов группировки выборки.

В строке E1:N1 получены середины y_j^* интервалов группировки показателя Y, а в столбце D2:D11 — середины x_i^* интервалов группировки показателя X, в диапазоне E2:N11 (выделен рамкой) — частоты n_{ij}, в строке E12:N12 — частоты n_{yj}, в строке E13:N13 — условные средние $\overline{x}_j$, в столбце O2:O11 — частоты n_{xi}, в столбце P2:P11 — условные средние $\overline{y}_i$, в ячейках O12, P12 выведены для контроля суммы $\sum_{i=1}^{k} n_{xi} = \sum_{j=1}^{k} n_{yj} = n.$

	D	E	F	G	H	I	J	K	L	M	N	O	P
1	0,00	-21,01	-17,21	-13,40	-9,60	-5,80	-2,00	1,80	5,60	9,40	13,21	0	0
2	-0,76	1	1	0	0	0	0	0	0	0	0	2	-19,11
3	-0,19	0	0	1	2	0	0	0	0	0	0	3	-10,87
4	0,38	0	0	0	0	1	6	0	0	0	0	7	-2,54
5	0,95	0	0	0	0	0	11	0	0	0	0	11	-2,00
6	1,52	0	0	0	0	0	17	5	0	0	0	22	-1,14
7	2,09	0	0	0	0	0	0	18	0	0	0	18	1,80
8	2,66	0	0	0	0	0	0	16	0	0	0	16	1,80
9	3,23	0	0	0	0	0	0	12	0	0	0	12	1,80
10	3,79	0	0	0	0	0	0	1	6	1	0	8	5,60
11	4,36	0	0	0	0	0	0	0	0	0	1	1	13,21
12	0,00	1	1	1	2	1	34	52	6	1	1	100	100
13	0,00	-0,76	-0,76	-0,19	-0,19	0,38	1,13	2,50	3,79	3,79	4,36	0	0

Рис.2.7

Вычислим корреляционное отношение η_{yx}. Для этого введите в ячейку Q2 формулу

=СУММПРОИЗВ((P2:P11-СУММПРОИЗВ(E1:N1;E12:N12)/100)^2;O2:O11)/
СУММПРОИЗВ((E1:N1-СУММПРОИЗВ(E1:N1;E12:N12)/100)^2;E12:N12),

а в ячейку Q3 — формулу =КОРЕНЬ(Q2). Получим значение 0,969389943.

Вычислим корреляционное отношение η_{xy}. Для этого введите в Q4 формулу

=СУММПРОИЗВ((E13:N13-СУММПРОИЗВ(D2:D11;O2:O11)/100)^2;E12:N12)/
СУММПРОИЗВ((D2:D11-СУММПРОИЗВ(D2:D11;O2:O11)/100)^2;O2:O11)

и в ячейку Q5 введите формулу =КОРЕНЬ(Q4). Получим значение 0,902441781.

Полученные значения $\eta_{yx} = 0{,}97$ и $\eta_{xy} = 0{,}90$ близки к единице, что подтверждает наличие нелинейной корреляционной связи между признаками. Значимость можно не проверять, так как значения близки к единице и объем выборки большой (n = 100).

Замечание. При построении корреляционной таблицы мы фактически заменили выборочные значения показателей средними значениями интервалов группировки. Это привело к небольшому искажению картины — *нелинейная функциональная зависимость* между исходными показателями соответствует *сильной нелинейной корреляционной зависимости* между сгруппированными показателями. По этой же причине выборочные характеристики, построенные по корреляционной таблице, будут только приближенно совпадать с выборочными характеристиками, полученными для исходной выборки.

```
(General)                                    KorTab

   Option Explicit: Option Base 0
Function KorTab(x, y, k)
 Application.Volatile (False)
 Dim i, j, m, n, ns, ns1 As Integer:   Dim nxi(), nyj() As Integer
 Dim nij(), xi(), yj(), axj(), ayi() As Variant
 Dim minx, maxx, miny, maxy, hx, hy, s As Variant
 n = Application.Count(x)
 ReDim nij(k + 2, k + 2), xi(k), yj(k), axj(k), ayi(k), nxi(k), nyj(k)
 minx = Application.WorksheetFunction.Min(x):  maxx = Application.WorksheetFunction.Max(x)
 miny = Application.WorksheetFunction.Min(y):  maxy = Application.WorksheetFunction.Max(y)
 hx = (maxx - minx) / (k - 1): hy = (maxy - miny) / (k - 1)
 xi(0) = minx - hx / 2: yj(0) = miny - hy / 2
 For i = 1 To k:   xi(i) = xi(i - 1) + hx: yj(i) = yj(i - 1) + hy:   Next i
 For i = 1 To k: For j = 1 To k:  For m = 1 To n
 If (xi(i - 1) < x(m)) And (x(m) <= xi(i)) And (yj(j - 1) < y(m)) And (y(m) <= yj(j)) Then nij(i, j) = nij(i, j) + 1
 Next m:  Next j:  Next i
 For i = 1 To k:  xi(i) = xi(i) - hx / 2: yj(i) = yj(i) - hy / 2:  Next i
 For j = 1 To k:  s = 0: nyj(j) = 0:  For i = 1 To k: nyj(j) = nyj(j) + nij(i, j)
       s = s + xi(i) * nij(i, j):   Next i:   axj(j) = s / nyj(j):  Next j
 For i = 1 To k: s = 0: nxi(i) = 0:  For j = 1 To k:
  nxi(i) = nxi(i) + nij(i, j): s = s + yj(j) * nij(i, j): Next j:  ayi(i) = s / nxi(i):  Next i
 ns = 0: ns1 = 0:  For i = 1 To k: ns = ns + nyj(i): ns1 = ns1 + nxi(i):  Next i
 For i = 1 To k:  nij(i, 0) = xi(i): nij(i, k + 1) = nxi(i): nij(i, k + 2) = ayi(i)
  nij(0, i) = yj(i): nij(k + 1, i) = nyj(i): nij(k + 2, i) = axj(i)
 Next i: nij(k + 1, k + 1) = ns: nij(k + 1, k + 2) = ns1
 KorTab = nij
End Function
```

Рис.2.8

ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

1. Сформировать с помощью формулы =СЛЧИС() в диапазоне A1:A100 массив значений *X*, а в диапазоне B1:B100 с помощью той же формулы — массив значений *Y.* В ячейку C1 введите формулу =*k**A1+СЛЧИС(), затем протяните маркером заполнения ячейку C1 до C100. Этот массив значений обозначим *Z*. С помощью выборочного коэффициента корреляции с уровнем доверия 0,95 выяснить:
А) Есть ли между *Y* и *X* линейная корреляционная зависимость?
Б) Есть ли между *Y* и *Z* линейная корреляционная зависимость?
В) Есть ли между *Z* и *X* линейная корреляционная зависимость?

Варианты значений параметра *k* даны в табл. 2.4.

Таблица 2.4

Вари ант	*k*	Вари ант	*k*	Вари ант	*k*	Вари ант	*k*
1	0,5	6	1,6	11	3,1	16	4,6
2	0,2	7	1,9	12	3,4	17	4,9
3	0,7	8	2,2	13	3,7	18	5,2
4	1,1	9	2,5	14	4	19	5,5
5	1,3	10	2,8	15	4,3	20	5,8

2. 1) Смоделировать выборку (x_i, y_i) объема $n = 100$, где x_i — значения нормально распределенной случайной величины с математическим ожиданием $a = 3$ и среднеквадратическим отклонением $\sigma = 2$, а $y_i = (x_i + 1)^2$ (указание: введите в ячейку A1 формулу =НОРМОБР(СЛЧИС();3;е), затем протяните эту ячейку маркером автозаполнения вниз до A100. В ячейку B1 введите формулу =(A1+1)^2, протяните B1 вниз до B100).

2) Вычислить значение выборочного коэффициента корреляции и проверить значимость для уровня $\alpha = 0,05$.

3) Вычислить выборочные корреляционные отношения η^2_{yx} и η^2_{yx}, и проверить гипотезу об отсутствии корреляционной зависимости между *Y* и *X*.

3. По данным корреляционной таблицы (даны пары (x_i, y_j) и частоты n_{ij}) определить наличие и характер (линейной или нелинейной)

корреляционной зависимости между *Y* и *X*, между *X* и *Y*. Варианты таблиц приведены ниже в табл. 2.5.

Таблица 2.5

Вариант 1					
X \ Y	4	6	8	10	12
4	4				
5,5	1	9	7		
7	5	5	5	4	3
8,5		9	6	4	6
10			7	5	4
Вариант 2					
X \ Y	3	5	7	9	11
3	2				
4,5	4	5	4		
6	2	5	5	6	4
7,5		7	6	7	7
9			4	5	7
Вариант 3					
X \ Y	1	3	5	7	9
4	2				
5,5	1	5	4		
7	4	5	3	7	6
8,5		5	6	5	4
10			7	4	3
Вариант 4					
X \ Y	1	3	5	7	9
2	2				
3,5	5	9	7		
5	4	7	7	4	2
6,5		8	7	3	4
8			3	5	5
Вариант 5					
X \ Y	3	5	7	9	11
4	1				
5,5	3	8	6		

Вариант 6					
X \ Y	3	5	7	9	11
1	5				
2,5	4	4	6		
4	3	5	5	7	2
5,5		5	4	7	7
7			4	4	5
Вариант 7					
X \ Y	3	5	7	9	11
3	1				
4,5	4	9	6		
6	2	8	5	7	5
7,5		6	6	7	4
9			7	7	4
Вариант 8					
X \ Y	1	3	5	7	9
0	1				
1,5	3	7	4		
3	1	7	5	4	3
4,5		5	7	4	6
6			5	4	3
Вариант 9					
X \ Y	4	6	8	10	12
1	3				
2,5	2	5	5		
4	4	8	3	5	2
5,5		8	3	5	4
7			3	3	6
Вариант 10					
X \ Y	3	5	7	9	11
2	1				
3,5	4	6	5		

7	3	9	4	6	6
8,5		5	3	3	3
10			7	3	3

5	3	9	6	6	3
6,5		9	6	3	7
8			6	6	7

РЕГРЕССИОННЫЙ АНАЛИЗ

Проанализировав связи между случайными переменными, и удостоверившись, что корреляция имеет место, определим уравнение регрессии, т.е., функциональную связь между этими переменными, используя метод наименьших квадратов.

Парная регрессия

Парная регрессия Пусть при изучении зависимости показателя Y от значений показателя X в результате наблюдений получена таблица парных значений

(x_i, y_i), $i = 1, \ldots, n$.

Линейное уравнение регрессии имеет вид

$$\bar{y}_x = b_0 + b_1 x \qquad (3.1)$$

Коэффициенты b_1 и b_0 определяются по формулам:

$$b_1 = \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i - n \cdot \sum_{i=1}^{n} x_i \cdot y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \cdot \sum_{i=1}^{n} x_i^{\,2}},$$

$$b_0 = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i - b_1 \cdot \sum_{i=1}^{n} x_i\right) = \bar{y} - b_1 \cdot \bar{x}, \qquad (3.2)$$

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i. \qquad (3.3)$$

2.1.1 Проверка значимости уравнения парной регрессии

Для проверки значимости модели используется F-критерий (критерий Фишера):

$$F = \frac{\sum_{i=1}^{n}\left(y_{\text{теор},\, i} - \overline{y}\right)^2 (n-2)}{\sum_{i=1}^{n}\left(y_i - y_{\text{теор},\, i}\right)^2} \qquad (3.4)$$

Уравнение (3.1) значимо, если

$$F > F_{\alpha;1;n-2}, \qquad (3.5)$$

где $F_{\alpha;1;n-2}$ значение *F*-критерия Фишера-Снедекора с уровнем значимости α с $k_1 = m - 1 = 2 - 1 = 1$ и $k_2 = n - 2$ степенями свободы.

2.1.2 Проверка значимости и построение доверительного интервала для коэффициентов парной регрессии

Значимость коэффициентов регрессии проверяется с помощью t- критерия Стьюдента:

$$t_{b_0} = \frac{|b_0|}{s_{b_0}},\ s_{b_0}^2 = \frac{s_{\text{î ñò}}^2 \sum x_i^2}{n\sum\left(x_i - \overline{x}\right)^2}, \qquad (3.6)$$

$$t_{b_1} = \frac{|b_1|}{s_{b_1}},\ s_{b_1}^2 = \frac{s_{\text{î ñò}}^2}{n\sum\left(x_i - \overline{x}\right)^2}, \qquad (3.7)$$

$$s_{\text{î ñò}}^2 = \frac{\sum_{i=1}^{n}\left(y_i - y_{\text{òåî ð},\, i}\right)^2}{n-2},\ y_{\text{òåî ð},i} = \overline{y}_{x_i} = b_0 + b_1 x_i\,. \qquad (3.8)$$

Параметр модели признается статистически значимым, если

$$t \geq t_{\text{кр}}(\alpha;\ n - 2), \qquad (3.9)$$

где α — уровень значимости; $n - 2$ — число степеней свободы.

Доверительный интервал для генерального коэффициента регрессии β$_i$, имеет вид

$$b_i - t_{\text{êð}}(\alpha;n-2)\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}} \le \beta_i \le b_i + t_{\text{êð}}(\alpha;n-2)\frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}. \quad (3.10)$$

Здесь значение $t_{кр}(\alpha;\ n-2)$ определяется с помощью функции СТЬЮДРАСПОБР(α;n – 2) программы *Excel* или по таблице распределения Стьюдента, а значения s_x и s_y вычисляются по формулам

$$s_x = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2},\ s_y = \sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}. \quad (3.11)$$

2.1.3 Построение доверительного интервала для условного математического ожидания и индивидуальных значений зависимой переменной

Доверительный интервал для условного математического ожидания зависимой переменной $M_x(Y)$ имеет вид

$$y_x - t_{\text{êð}}(\alpha;n-2)s_{y_x} \le M_x(Y) \le y_x + t_{\text{êð}}(\alpha;n-2)s_{y_x}, \quad (3.12)$$

где s_{y_x} — оценка дисперсии для групповой средней y_x, вычисляется по формуле

$$s_{y_x}^2 = s_{\text{î ñò}}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \quad (3.13)$$

Доверительный интервал для индивидуального значения зависимой переменной y_0, соответствующего значению x_0 $(y_0 = b_0 + b_1x_0)$ имеет вид

$$y_{x_0} - t_{\text{êð}}(\alpha;n-2)s_{y_0} \le y_0 \le y_{x_0} + t_{\text{êð}}(\alpha;n-2)s_{y_0}, \quad (3.14)$$

где s_{y_0} вычисляется по формуле

$$s_{y_x}^2 = s_{\text{ост}}^2 \left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum\limits_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \right) \qquad (3.15)$$

Если уравнение парной регрессии имеет вид $\bar{y}_x = b_1 x$, то доверительный интервал для условного математического ожидания зависимой переменной $M_x(Y)$ имеет вид

$$y_x - t_{\text{кр}}\left(\alpha; n-2\right) s_{y_x} \le M_x\left(Y\right) \le y_x + t_{\text{кр}}\left(\alpha; n-2\right) s_{y_x}$$

, (3.16)

Множественная линейная регрессия

Пусть при изучении зависимости переменной Y от нескольких переменных X_1, X_2, …, X_p получены наблюденные значения y_i и $x_{i,1}$, $x_{i,2}$, …, $x_{i,p}$, i = 1, 2, …, n; p — число объясняющих переменных.

Уравнение *множественной линейной регрессии* имеет вид:

$$\bar{y}_x = b_0 + b_1 x_1 + b_2 x_2 \ldots + b_p x_p \qquad (3.16)$$

Для определения коэффициентов b_0, b_1, …, b_p нужно решить систему линейных уравнений (систему нормальных уравнений метода наименьших квадратов):

$$\begin{pmatrix} n & \sum x_{i1} & \sum x_{i2} & \ldots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1} x_{i2} & \ldots & \sum x_{i1} x_{ip} \\ \sum x_{i2} & \sum x_{i2} x_{i1} & \sum x_{i2}^2 & \ldots & \sum x_{i2} x_{ip} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \sum x_{ip} & \sum x_{ip} x_{i1} & \sum x_{ip} x_{i2} & \ldots & \sum x_{ip}^2 \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \ldots \\ b_p \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \sum x_{i2} y_i \\ \ldots \\ \sum x_{ip} y_i \end{pmatrix} \qquad (3.17)$$

2.1.4 Проверка адекватности уравнения множественной регрессии и значимости коэффициентов модели

Процедура «Регрессия» из пакета «Анализ данных» программы *Excel* вычисляет коэффициенты регрессии

Нелинейные модели

Наиболее часто встречающиеся виды нелинейной регрессии:

1) степенная $\bar{y}_x = b_0 \cdot x^{b_1}$;

2) логарифмическая $\bar{y}_x = b_0 + b_1 \cdot \ln x$;

3) экспоненциальная $\bar{y}_x = b_0 \cdot e^{b_1 x}$;

4) полиномиальная $\bar{y}_x = b_0 + b_1 \cdot x + b_2 \cdot x^2 ... + b_k \cdot x^k$;

5) гиперболическая $\bar{y}_x = b_0 + b_1 \cdot \frac{1}{x}$;

6) степенная множественная $\bar{y}_x = b_0 \cdot x_1^{b_1} \cdot x_2^{b_2} ... \cdot x_k^{b_k}$.

Первые четыре вида уравнений регрессии можно построить в программе *Excel* с помощью функции «Диаграмма–Добавить линию тренда», если уже построена диаграмма (типа «Точечная», «Графики» и «Гистограмма») по имеющейся на рабочем листе таблице исходных данных.

Пример 3.1. 1) Построить уравнение регрессии $\bar{y}_x = b_0 + b_1 x$ по данным из табл. 3.1.

Таблица 3.1

i	x_i	y_i	*i*	x_i	y_i	*i*	x_i	y_i
1	0,43	3,61	11	0,32	3,55	21	0,52	3,65
2	0,25	3,34	12	0,79	4,42	22	0,38	3,87
3	0,52	3,41	13	0,70	3,54	23	0,32	3,83
4	0,63	4,29	14	0,95	4,43	24	0,34	3,35
5	0,01	3,43	15	0,60	3,98	25	0,28	3,87
6	0,91	4,42	16	0,98	4,17	26	0,91	3,70

7	0,41	3,81	17	0,75	4,08	27	0,90	3,61
8	0,45	3,39	18	0,69	3,71	28	0,60	3,90
9	0,05	4,00	19	0,46	3,83	29	0,45	3,44
10	0,44	3,92	20	0,63	3,72	30	0,84	3,72

2) Проверить значимость коэффициентов регрессии для уровня
$\alpha = 0{,}05$.

3) Построить доверительный интервал для генерального коэффициента регрессии β_{yx} с уровнем значимости $\alpha = 0{,}05$.

4) Проверить значимость уравнения регрессии с уровнем значимости $\alpha = 0{,}07$.

5) Построить доверительные границы для уравнения регрессии с уровнем значимости $\alpha = 0{,}05$ и построить соответствующие графики.

Решение. 1) Запишем исходные данные в диапазон A1:B31 (табл. 3.2). В ячейках A1:C1 и D2:D13 введем для наглядности указанные обозначения.

В ячейку E2 введем формулу

=(СУММ(A2:A31)*СУММ(B2:B31)-
30*СУММПРОИЗВ(A2:A31;B2:B31))/
(СУММ(A2:A31)^2-30*СУММКВ(A2:A31))

В ячейку E3 введем формулу

=СРЗНАЧ(B2:B31)-E2*СРЗНАЧ(A2:A31)

В ячейках E2, E3 получим коэффициенты уравнения регрессии
b_1 = 4,94675; b_0 = 0,56. Уравнение регрессии имеет вид

$$\overline{y}_x = 0{,}56 + 4{,}94675x$$

Чтобы выполнить задания 2) и 3) в ячейки E4:E13 вводим формулы, как показано в табл. 3.2, являющейся частью табл. 3.3. Для просмотра формул в программе *Excel* надо выполнить команду «Параметры» и поставить флажок в строке «формулы».

Таблица 3.2

	D	E
4	xср=	=СРЗНАЧ(A2:A31)
5	yср=	=СРЗНАЧ(B2:B31)
6	sост=	=СУММКВРАЗН(C2:C31;B2:B31)/28
7	sb1=	=E6/(30*СУММ((A2:A31-E4)^2))
8	sb0=	=E6*СУММКВ(A2:A31)/(30*СУММ((A2:A31-E4)^2))
9	tb1=	=ABS(E2)/КОРЕНЬ(E7)
10	tb0=	=ABS(E3)/КОРЕНЬ(E8)
11	tкрит=	=СТЬЮДРАСПОБР(0,05;28)
12	F=	=28*СУММ((C2:C31-E52)^2)/СУММКВРАЗН(C2:C31;B2:B31)
13	Fкрит=	=FРАСПОБР(0,05;1;28)

Результаты вычислений приведены в табл.3.3.

Таблица 3.3

	A	B	C	D	E
1	*x*	*y*	yтеор		
2	0,43	3,61	3,727561	b1=	0,598907286
3	0,25	3,34	3,619757	b0=	3,47
4	0,52	3,41	3,781462	xср=	0,55
5	0,63	4,29	3,847342	yср=	3,80
6	0,01	3,43	3,47602	sост=	0,081533115
7	0,91	4,42	4,015036	sb1=	0,02100934
8	0,41	3,81	3,715582	sb0=	1,51990504
9	0,45	3,39	3,739539	tb1=	4,131934193
10	0,05	4,00	3,499976	tb0=	2,814654319
11	0,44	3,92	3,73355	tкрит=	2,048407115
12	0,32	3,55	3,661681	F=	164,447372
13	0,79	4,42	3,943167	Fкрит=	4,195971707
14	0,70	3,54	3,889266	Дов.инт. Для betayx	
15	0,95	4,43	4,038992	нижн.гран.	0,177612121
16	0,60	3,98	3,829375	верх.гран.	1,020202452
17	0,98	4,17	4,05696		
18	0,75	4,08	3,919211		

19	0,69	3,71	3,883277		
20	0,46	3,83	3,745528		
21	0,63	3,72	3,847342		
22	0,52	3,65	3,781462		
23	0,38	3,87	3,697615		
24	0,32	3,83	3,661681		
25	0,34	3,35	3,673659		
26	0,28	3,87	3,637725		
27	0,91	3,70	4,015036		
28	0,90	3,61	4,009047		
29	0,60	3,90	3,829375		
30	0,45	3,44	3,739539		
31	0,84	3,72	3,973113		

2) Проверка значимости коэффициентов регрессии. Так как выполнены неравенства

$$t_{b_1} = 33,45 > t_{\text{êð}}\left(0,05;28\right) = 2,05;$$

$$t_{b_0} = 0,45 < t_{\text{êð}}\left(0,05;28\right) = 2,05$$

то делаем вывод о том, что коэффициент b_1 значим, а коэффициент b_0 незначим.

Замечание. Отсюда следует, что параметр b_0 следует исключить из модели, т.е. искать уравнение регрессии в виде $\overline{y}_x = b_1 x$, где b_1 вычисляется по формуле

$$b_1 = \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\left(\sum_{i=1}^{n} x_i\right)^2}.$$

3) В ячейку E15 вводим следующую формулу

=E2-E11*СТАНДОТКЛОНП(B2:B31)*
КОРЕНЬ(1-КОРРЕЛ(A2:A31;B2:B31)^2)/
(СТАНДОТКЛОНП(A2:A31)*КОРЕНЬ(28))

В ячейку E16 вводим формулу

=E2+E11*СТАНДОТКЛОНП(B2:B31)*
КОРЕНЬ(1-КОРРЕЛ(A2:A31;B2:B31)^2)/
(СТАНДОТКЛОНП(A2:A31)*КОРЕНЬ(28))

Доверительный интервал для генерального коэффициента регрессии β_{yx} с уровнем значимости α = 0,05 имеет вид

$4{,}51 < \beta_{yx} < 5{,}38$.

4) Проверка значимости уравнения регрессии (ячейки E12, E13):

$F = 54{,}3 > F_{\text{крит}} = 4{,}2$.

Уравнение значимо с уровнем доверия 1 – 0,05 = 0,95 = 95%.

5) Построим доверительные границы для уравнения регрессии. Для этого построим таблицу значений для нижней и верхней границ по формулам (3.12) и (3.14). Мы будем применять эти формулы к одним и тем же значениям переменной x, которые выберем так, чтобы они не выходили за пределы интервала (min x_i; max x_i). Выберем значения x = 0,50; 0,51, …, 0,60.

Таблицу для вычисления доверительных границ построим в диапазоне F1:M12 (табл. 3.5).

В диапазоне F1:M1 введите обозначения.

В F2, F3 введем числа 0,50 и 0,51, затем выделим ячейки F2:F3 и маркером заполнения протянем вниз до F12.

В G2 введем формулу

=E$6*КОРЕНЬ((1/30+(F2-E$4)^2/(29*ДИСП(A$2:A$31))))

и маркером заполнения протянем ячейку G2 вниз до G12.

В H2 введем формулу

=E$6*КОРЕНЬ((1+1/30+(F2-E$4)^2/(29*ДИСП(A$2:A$31))))

и маркером заполнения протянем ячейку H2 вниз до H12.

В ячейках диапазона I2:M2 введем формулы, указанные в табл. 3.4. Затем выделим диапазон I2:M2 и маркером заполнения протянем вниз до строки 12.

Таблица 3.4

	I	**J**	**K**	**L**	**M**
1	Yx	Y01	Yx1	Yx2	Y02

2	=E$3+E$2 *F2	=I2-E$11*H2	=I2-E$11*G2	=I2+E$11*G2	=I2+E$11*H2

В результате получим таблицу 3.5 для построения доверительных границ.

Чтобы построить график уравнения регрессии и графики доверительных границ, выделим диапазон F1:F12, затем, удерживая нажатой клавишу Ctrl, выделим диапазон I1:M12 и с помощью «Мастера диаграмм» построим диаграмму «Точечная» (рис.3.1).

Здесь Yx1 и Yx2 обозначают нижнюю и верхнюю доверительные границы для $M_x(Y)$, вычисляемые по формуле (3.12), а Y01 и Y02 обозначают нижнюю и верхнюю доверительные границы для индивидуальных значений y_0, вычисляемые по формуле (3.14).

Таблица 3.5

	F	**G**	**H**	**I**	**J**	**K**	**L**	**M**
1	x	Syx	Sy0	Yx	Y01	Yx1	Yx2	Y02
2	0,50	0,016	0,086	3,037	2,860	3,004	3,069	3,214
3	0,51	0,016	0,086	3,086	2,909	3,054	3,118	3,263
4	0,52	0,016	0,086	3,136	2,959	3,104	3,168	3,312
5	0,53	0,016	0,086	3,185	3,008	3,153	3,217	3,362
6	0,54	0,016	0,086	3,235	3,058	3,203	3,266	3,411
7	0,55	0,015	0,086	3,284	3,107	3,252	3,316	3,461
8	0,56	0,016	0,086	3,333	3,157	3,302	3,365	3,510
9	0,57	0,016	0,086	3,383	3,206	3,351	3,415	3,560
10	0,58	0,016	0,086	3,432	3,256	3,400	3,464	3,609
11	0,59	0,016	0,086	3,482	3,305	3,450	3,514	3,659
12	0,60	0,016	0,086	3,531	3,354	3,499	3,564	3,708

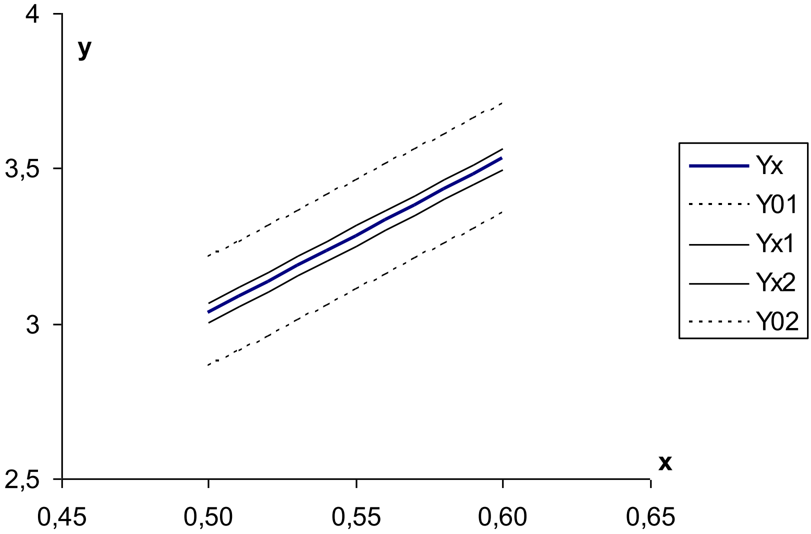

Рис. 3.1

Решение с помощью процедуры «Регрессия» из пакета «Анализ данных».

Выполните команду меню «Сервис—Анализ данных», выберите процедуру «Регрессия» и в появившемся окне (рис.3.2) в строке «Входной интервал Y:» введите диапазон B1:B31, в строке «Входной интервал X:» введите диапазон A1:A31, поставьте флажок в строках «Метки» и «Уровень надежности», в «Параметрах выхода» укажите «Новый рабочий лист» и нажмите «Ок».

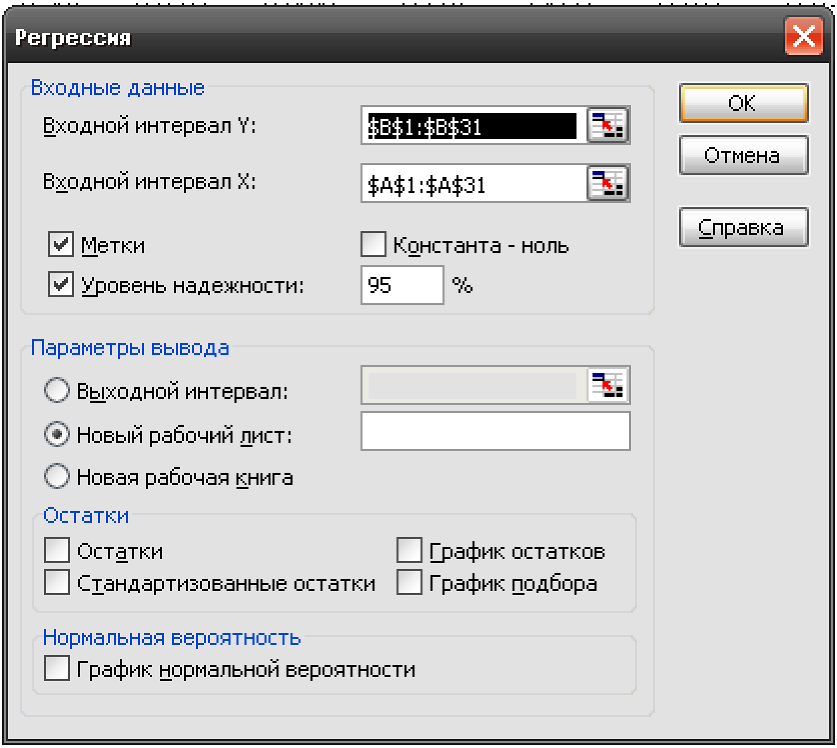

Рис.3.2

Результаты будут выведены на отдельный лист. Ниже приведены результаты.

Таблица 3.6

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,4821							
5	R-квадрат	0,2324							
6	Нормированный R-квадрат	0,2050							
7	Стандартная ошибка	0,2855							
8	Наблюдения	30							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	1	0,6914	0,6914	8,479 7	0,0070			

13	Остаток	28	2,2829	0,0815					
14	Итого	29	2,9743						
15									
16		*Коэффи-циенты*	*Стан-дартная ошибка*	*t-стати-стика*	*P-Зна-чение*	*Ниж-ние 95%*	*Верх-ние 95%*	*Ниж-ние 95,0%*	*Верх-ние 95,0%*
17	Y-пересече-ние	3,4700	0,1246	27,8459	0,000	3,2148	3,7253	3,2148	3,7253
18	x	0,5989	0,2057	2,9120	0,007	0,1776	1,0202	0,1776	1,0202

Пример 3.2. Построить уравнение регрессии $\overline{y}_x = b_0 + b_1 x$ по данным корреляционной таблицы (табл. 3.6).

Таблица 3.7

x_i \ y_j	9	13	17	21	25
22,5	2	1			
27,5	3	6	4		
32,5		3	11	7	
37,5		1	2	6	2
42,5				1	1

2) Проверить значимость коэффициентов регрессии для уровня
α = 0,07.

3) Проверить значимость уравнения регрессии и построить доверительный интервал для него.

Задания для самостоятельной работы

1. Построить уравнение регрессии $\overline{y}_x = b_0 + b_1 x$ по данным из таблицы 3.6. Проверить значимость коэффициентов регрессии для уровня α = 0,05. Построить доверительный интервал для генерального коэффициента регрессии β_{yx} с уровнем значимости α = 0,05. Проверить значимость уравнения регрессии с уровнем значимости α = 0,05. Построить доверительные границы для уравнения регрессии с уровнем значимости α = 0,05 и построить соответствующие графики.

Таблица 3.8

вариант 1														
X	5,9	7,5	5,4	5,8	6,0	7,0	5,4	7,3	5,0	6,0	6,6	5,6	7,1	7,4
Y	-5,0	0,3	11,1	12,6	-5,4	14,9	0,1	7,4	5,4	0,7	13,7	12,1	0,5	0,1
вариант 2														
X	7,6	5,3	5,7	5,1	5,4	5,1	5,1	6,3	6,1	5,3	5,6	5,9	7,2	5,8
Y	-7,6	0,8	-5,4	5,6	11,6	-4,6	5,2	-5,6	12,4	0,2	-5,1	12,5	0,5	12,3
вариант 3														
X	6,1	8,0	7,4	7,1	7,1	7,9	7,4	5,4	7,1	5,8	7,3	5,1	5,3	7,4
Y	12,7	16,8	0,1	14,9	14,6	16,8	8,1	0,1	-6,6	0,5	-6,8	-4,6	11,4	15,7
вариант 4														
X	8,0	7,4	7,9	5,4	5,8	7,3	5,7	5,6	5,6	6,4	7,2	7,9	6,2	5,1
Y	8,3	0,1	0,2	11,4	0,5	8,2	11,9	11,5	0,0	0,1	15,1	16,7	6,5	-4,4
вариант 5														
X	7,5	5,7	6,4	7,4	7,0	7,3	6,5	7,7	5,0	6,7	7,3	6,4	6,8	5,5
Y	7,7	0,5	0,8	8,0	0,3	-7,0	-6,2	0,9	-5,0	-6,6	0,2	-5,6	-5,9	0,9
вариант 6														
X	6,6	5,2	7,1	5,3	7,2	7,6	5,4	5,1	5,6	6,8	7,7	5,8	7,4	5,2
Y	-6,5	-4,5	0,1	10,6	7,8	8,0	0,7	0,1	12,0	14,3	15,5	11,9	8,1	11,3
вариант 7														
X	5,3	6,7	6,9	7,0	6,1	5,9	6,7	5,8	7,5	6,1	5,2	5,4	7,5	5,9
Y	5,8	7,1	7,1	0,1	7,0	-5,7	14,1	11,6	7,8	-5,8	5,9	5,9	0,3	12,6

вариант 8														
X	6,8	5,8	7,6	6,2	5,1	6,7	5,0	6,2	7,9	5,2	5,8	7,1	7,8	5,1
Y	14,0	6,2	15,9	12,6	5,2	13,7	0,6	-5,5	8,6	10,6	11,9	7,5	8,7	0,0
вариант 9														
X	6,3	5,8	5,7	7,0	6,5	7,1	7,4	5,7	7,4	5,5	5,6	7,0	5,9	6,7
Y	13,4	6,4	0,0	-6,8	13,7	7,6	7,8	0,5	0,4	-4,6	0,2	7,1	-5,5	7,3
Вариант 10														
X	6,8	6,3	6,4	7,7	7,6	5,7	6,5	6,6	6,3	5,6	7,3	7,7	5,8	6,2
Y	0,7	-5,9	0,8	8,2	0,5	11,9	-6,4	14,0	0,7	5,7	7,4	7,7	0,3	0,4
Вариант 11														
X	7,8	7,2	7,1	7,7	6,1	7,5	5,4	6,8	6,9	5,3	6,5	7,7	5,9	6,1
Y	0,0	14,8	-6,3	16,0	12,5	0,0	5,9	7,7	7,5	11,3	7,4	15,7	6,3	-5,2
Вариант 12														
X	7,2	7,6	5,4	6,5	7,9	5,5	6,6	5,7	7,7	7,6	6,8	6,0	7,6	7,0
Y	0,8	15,8	11,3	7,3	-7,3	0,2	7,0	-4,7	0,4	8,4	7,5	0,4	0,1	14,7
Вариант 13														
X	5,8	7,2	5,8	6,2	5,0	7,7	6,6	5,8	7,8	7,4	6,1	5,4	7,6	6,3
Y	6,4	-6,2	6,5	-6,1	0,7	15,8	0,8	0,2	0,6	-7,0	12,5	0,6	15,8	-6,3
Вариант 14														
X	7,1	7,4	5,2	5,4	5,3	5,1	6,1	7,8	7,6	6,4	7,1	5,9	7,2	5,6
Y	-6,9	15,0	0,6	0,1	-5,0	5,8	6,1	0,8	7,8	6,9	-6,4	0,6	0,7	11,6
Вариант 15														

X	5,0	7,3	6,7	6,9	6,4	5,3	5,6	6,1	6,9	6,1	7,5	7,6	7,1	5,1
Y	1,0	14,6	13,8	14,0	1,0	-4,6	-4,9	0,7	13,9	0,4	-6,6	15,3	8,0	5,3

2. Построить уравнение регрессии $\overline{y}_x = b_0 + b_1 x_1 + b_2 x_2$ по данным из таблицы 3.9. Проверить значимость коэффициентов регрессии для уровня α = 0,05. Построить доверительный интервал для генерального коэффициента регрессии β_{yx} с уровнем значимости α = 0,05. Проверить значимость уравнения регрессии с уровнем значимости α = 0,05.

Таблица 3.9

вариант 1															
X1	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5	8,0
X2	2,4	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9	8,4	8,9	9,4
Y	-1	0	0	0	0	1	2	2	3	3	4	4	5	5	5
вариант 2															
X1	0,9	1,4	1,9	2,4	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9
X2	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5
Y	2	3	3	4	5	5	5	6	6	7	8	7	8	8	9
вариант 3															
X1	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5	8,0
X2	1,1	1,6	2,1	2,6	3,1	3,6	4,1	4,6	5,1	5,6	6,1	6,6	7,1	7,6	8,1
Y	1	2	2	3	3	4	5	5	5	6	6	7	7	8	8
вариант 4															
X1	0,9	1,4	1,9	2,4	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9
X2	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5	8,0	8,5	9,0
Y	-1	0	0	1	1	1	3	2	3	3	4	5	5	6	6
вариант 5															
X1	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5	8,0
X2	0,1	0,6	1,1	1,6	2,1	2,6	3,1	3,6	4,1	4,6	5,1	5,6	6,1	6,6	7,1
Y	3	4	4	4	5	5	6	6	7	7	9	9	9	10	10
вариант 6															
X1	0,6	1,1	1,6	2,1	2,6	3,1	3,6	4,1	4,6	5,1	5,6	6,1	6,6	7,1	7,6
X2	1,9	2,4	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9	8,4	8,9

Y	-2	-1	0	0	1	1	1	2	2	3	4	4	4	5	6
вариант 7															
X1	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5
X2	1,7	2,2	2,7	3,2	3,7	4,2	4,7	5,2	5,7	6,2	6,7	7,2	7,7	8,2	8,7
Y	-2	-1	0	0	0	1	1	3	3	3	3	4	4	5	5
вариант 8															
X1	0,3	0,8	1,3	1,8	2,3	2,8	3,3	3,8	4,3	4,8	5,3	5,8	6,3	6,8	7,3
X2	0,9	1,4	1,9	2,4	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9
Y	0	0	0	1	1	2	3	3	4	4	5	5	5	6	6
вариант 9															
X1	0,9	1,4	1,9	2,4	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9
X2	2,1	2,6	3,1	3,6	4,1	4,6	5,1	5,6	6,1	6,6	7,1	7,6	8,1	8,6	9,1
Y	-1	0	0	1	1	2	2	2	3	3	4	5	5	5	6
вариант 10															
X1	0,2	0,7	1,2	1,7	2,2	2,7	3,2	3,7	4,2	4,7	5,2	5,7	6,2	6,7	7,2
X2	0,8	1,3	1,8	2,3	2,8	3,3	3,8	4,3	4,8	5,3	5,8	6,3	6,8	7,3	7,8
Y	0	1	1	2	1	3	3	3	4	4	5	5	5	6	6
вариант 11															
X1	2,7	3,2	3,7	4,2	4,7	5,2	5,7	6,2	6,7	7,2	7,7	8,2	8,7	9,2	9,7
X2	1,7	2,2	2,7	3,2	3,7	4,2	4,7	5,2	5,7	6,2	6,7	7,2	7,7	8,2	8,7
Y	6	6	6	7	7	7	9	9	9	10	10	11	11	12	12
вариант 12															
X1	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9	8,4	8,9	9,4	9,9
X2	1,9	2,4	2,9	3,4	3,9	4,4	4,9	5,4	5,9	6,4	6,9	7,4	7,9	8,4	8,9

Y	6	6	7	7	7	8	9	8	9	10	10	11	11	12	12
вариант 13															
X1	1,3	1,8	2,3	2,8	3,3	3,8	4,3	4,8	5,3	5,8	6,3	6,8	7,3	7,8	8,3
X2	2,8	3,3	3,8	4,3	4,8	5,3	5,8	6,3	6,8	7,3	7,8	8,3	8,8	9,3	9,8
Y	-2	-1	-1	0	1	1	1	3	3	4	4	4	5	6	5
вариант 14															
X1	1,1	1,6	2,1	2,6	3,1	3,6	4,1	4,6	5,1	5,6	6,1	6,6	7,1	7,6	8,1
X2	1,8	2,3	2,8	3,3	3,8	4,3	4,8	5,3	5,8	6,3	6,8	7,3	7,8	8,3	8,8
Y	0	0	1	2	2	2	3	4	5	5	5	6	6	7	7

ЛИТЕРАТУРА

1 Вадзинский Р. Статистические вычисления в среде Excel. Библиотека пользователя. — Спб.: Питер, 2008. — 608с.

2 Соболь Б.В., Борисова Л.В., Иваночкина Т.А. Пешхоев И.М. Практикум по статистике в Excel. Ростов н/Д: Феникс, 2010. — 381с.