ДГТУ
Кафедра "Управление качеством"

ПРОВЕРКА НУЛЕВОЙ ГИПОТЕЗЫ О РАВЕНСТВЕ ДИСПЕРСИЙ И РАВЕНСТВЕ СРЕДНИХ В ДВУХ ВЫБОРКАХ

Методические указания к практическим занятиям
по дисциплине «Методы анализа данных»

Ростов-на-Дону
2021

Варианты задания

№ п/п	Вариант 1		Вариант 2		Вариант 3		Вариант 4		Вариант 5	
	У	Х	У	Х	У	Х	У	Х	У	Х
1	491	522	211,2	209,5	210,2	212	216,3	219,6	216,4	209,5
2	479	524	208,5	208,7	207,2	207,8	215,3	212,3	212	208,7
3	514	497	205,2	210,2	210,5	205,9	210,7	215,2	202,7	210,2
4	507	511	204	203,3	198,2	199,9	212,5	211,3	210,4	203,3
5	483	534	209,6	203,7	210,2	209,6	208,4	212,8	214,8	203,7
6	543	525	208,1	207,9	211	206,2	210,3	208,4	210,8	207,9
7	521	517	206,2	204,8	198,7	205,8	208,1	212,9	209	204,8
8	536	547	199	197,7	202	213,1	207,5	210,6	212,3	197,7
9	499	489	197,2	210,6	199,5	215,3	206,9	213,6	212,2	210,6
10	552	526	199,1	207,2	209,8	201,2	209,6	206,8	214,2	207,2
11	523	527	204,6	207	215,8	204,6	217,2	207,6	212,6	207
12	467	489	212,7	207,5	205,8	200,9	211,4	214,4	212,6	207,5
13	489	494	204,1	196,6	214,6	199,4	209,6	206,1	207,1	196,6
14	513	501	200,2	205,5	208	202,7	213,5	210,6	212,3	205,5
15	535	521	201,1	209,2	205,5	200	209,1	209,8	211,4	209,2
16	501	488	201,3	203,1	196,3	205,5	208	205,3	203,6	203,1
17	529	533	202,2	204,4	212,1	206,6	210,8	209,1	207	204,4
18	509	520	197,1	211	208,4	202,6	215,6	205,1	209	211
19	530	478	204,8	201,3	208,4	212,3	214,5	212,9	204,3	201,3
20	499	492	201,6	202,3	204,3	201,4	209,1	212	204,2	202,3

Рассматриваемые гипотезы имеют вид:

$$H_0 : \mu_1 = \mu_4; \quad H_1 : \mu_1 \neq \mu_4 (\alpha = 0{,}05),$$

$$H_0 : \mu_2 = \mu_4; \quad H_1 : \mu_2 \neq \mu_4 (\alpha = 0{,}05). \qquad (5.1)$$

$$H_0 : \mu_3 = \mu_4; \quad H_1 : \mu_3 \neq \mu_4 (\alpha = 0{,}05)$$

то есть на уровне значимости α=0,05 мы рассматриваем гипотезу о том, что время на регулировку μ_1 оператором группы G1 не отличается от времени на решение задачи с применением ЭС – μ_4. Проверка однородности дисперсий показала, что значения данного признака принадлежат одной совокупности. Принятие данной гипотезы позволяет проверить нулевую гипотезу равенства средних. Расчет статистик производился по формулам:

$$\bar{x}_1 = \frac{1}{n_1}\sum x_{1i} = 18{,}3 \qquad \bar{x}_4 = \frac{1}{n_4}\sum x_{4i} = 5{,}0$$

$$S_1^2 = \frac{\sum(x_{1i} - \bar{x}_1)^2}{n_1 - 1} = 9{,}82 \qquad S_4^2 = \frac{\sum(x_{4i} - \bar{x}_4)^2}{n_4 - 1} = 2{,}0$$

$$s_{1-4} = \sqrt{\frac{S_1^2 + S_4^2}{n_1 + n_4 - 2}} = 0{,}70 \qquad t_{1-4} = \frac{\bar{x}_1 - \bar{x}_4}{s_{1-4}\sqrt{\frac{1}{n_1} + \frac{1}{n_4}}} = 44{,}43.$$

где $\bar{x}_1$ – среднее значение затрат времени на регулировку молотилки оператором группы G1; $\bar{x}_4$ – среднее значение затрат времени на регулировку молотилки оператором при помощи ЭС; S_1^2, S_4^2 – дисперсии; s_{1-4} – среднее квадратическое отклонение; t – расчетное значение критерия Стьюдента.

Табличное значение критерия $t_{22;\,0,05} = 1{,}76$. Расчетное значение $|t_{1-4}| = 44{,}43$. Так как 44,43 > 1,76, то гипотезу о равенстве средних H_0 – отвергаем. При 5%-ом уровне значимости существует различие между временем, затраченным на регулировку молотилки при помощи ЭС и временем, затраченным на регулировку молотилки без ЭС по группе операторов G1.

ДГТУ
Кафедра "Управление качеством"

ПРОВЕРКА ГИПОТЕЗЫ О ВИДЕ РАСПРЕДЕЛЕНИЯ

Методические указания к практическим занятиям
по дисциплине «Методы анализа данных»

Ростов-на-Дону
2021

ВВЕДЕНИЕ

При нормальном законе распределения вероятности измеряемой величины несмещенной, состоятельной и эффективной оценкой среднего значения является *среднее арифметическое (среднее арифметическое взвешенное).*

Среднее арифметическое используется в качестве оценки среднего значения и в том случае, когда на основе анализа априорной и апостериорной информации не удается сформулировать правдоподобную гипотезу относительно закона распределения вероятности результата измерения. Хотя в этом случае среднее арифметическое может оказаться неэффективной оценкой, его стандартное отклонение все равно в $\sqrt{n}$ раз меньше стандартного отклонения результата измерения S.

Если закон распределения незначительно отличается от нормального, причем это отличие проявляется в повышенной вероятности больших отклонений от среднего значения, то используют *робастные оценки* среднего значения.

Если гипотеза о том, что результат измерения подчиняется нормальному закону распределения вероятности, отвергается и принимается гипотеза о том, что он подчиняется другому вполне определенному закону, то эффективная оценка среднего значения синтезируется *методом максимального правдо–подобия.*

Порядок проверки гипотезы о виде закона распределения с помощью критериев согласия рекомендуется такой: 1 – выбирают меру расхождения между теоретическим и эмпирическим законами распределения u. Закон распределения $P(U>u)$ должен быть известен; 2 – задают уровень значимости α; 3 – вычисляют меру расхождения для исследуемого статистического распределения $u_э$; 4 – находят табличное значение u_α, отвечающее заданному уровню значимости $P(U > u_\alpha) = \alpha$; 5 – делают вывод относительно проверяемой гипотезы о согласованности теоретического и эмпирического распределений:

если $u_э > u_\alpha$ – гипотеза отклоняется; если $u_э < u_\alpha$ – гипотеза принимается.

При большом объеме выборки (n >50) используется критерий χ^2 – Пирсона (задача 1).

Для n < 50 целесообразно использовать, например, составной критерий (задача 2).

Задача 1. ПРОВЕРКА ГИПОТЕЗЫ ПРИ БОЛЬШОМ ОБЪЕМЕ ВЫБОРКИ

Дано. Произведены измерения условной величины. Результаты измерений представлены в виде отклонений от номинального значения. Экспериментальные значения распределены по интервалам (табл. 1). Количество экспериментальных данных, попадающих в i–й интервал, приведено в табл. 2.

Т а б л и ц а 1

Экспериментальные значения распределенные по интервалам.

№ интервала	Предпоследняя цифра номера зачетной книжки									
	0	1	2	3	4	5	6	7	8	9
	Нижняя и верхняя границы интервалов									
1	1;2	0;2	0;5	0;10	0;3	0;5	0;20	0;100	0;4	4;6
2	2;3	2;4	5;10	10;20	3;6	5;10	20;40	100;200	4;8	6;8
3	3;4	4;6	10;15	20;30	6;9	10;15	40;60	200;300	8;12	8;10
4	4;5	6;8	15;20	30;40	9;12	15;20	60;80	300;400	12;16	10;12
5	5;6	8;10	20;25	40;50	12;15	20;25	80;100	400;500	16;20	12;14
6	6;7	10;12	25;30	50;60	15;18	25;30	100;120	500;600	20;24	14;16
7	7;8	12;14	30;35	60;70	18;21	30;35	120;140	600;700	24;28	16;18
8	8;9	14;16	35;40	70;80	21;24	35;40	140;160	700;800	28;32	18;20
9	9;10	16;18	40;45	80;90	24;25	40;45	160;180	800;900	32;36	20;22
10	10;12	18;20	45;50	90;100	27;30	45;50	180;200	900;1000	36;40	22;24
P*	0,95	0,975	0,98	0,99	0,90	0,80	0,995	0,98	0,95	0,99

*P – доверительная вероятность.

Таблица 2

Количество экспериментальных данных, попадающих в i-й интервал

Пос–ледняя цифра № зачёт-ной книжки	Число экспериментальных данных, попадающих в i–й интервал									
	№ интервала									
	1	2	3	4	5	6	7	8	9	10
0	2	6	25	72	133	120	88	46	10	4
1	0	1	5	20	60	32	18	8	4	0
2	1	10	35	120	210	95	42	20	6	2
3	2	6	12	20	25	19	10	8	4	1
4	5	32	50	92	110	89	47	25	12	3
5	3	10	21	35	48	60	39	23	14	6
6	0	4	9	14	17	15	10	7	3	1
7	2	15	32	50	82	90	75	41	18	5
8	1	10	18	25	36	28	20	16	11	2
9	2	14	30	43	55	40	26	14	4	0

Требуется:

1. Построить гистограмму эмпирического распределения.

2. Проверить гипотезу о соответствии эмпирического распределения нормаль–ному закону (закону Гаусса).

3. Построить доверительный интервал для среднего значения измеряемой вели-чины.

Методические рекомендации по решению задачи

1. Построить гистограмму эмпирического распределения.

1.1. Определить эмпирическую (статистическую) вероятность попадания случайной измеряемой величины в i–й интервал (частость):

$$P_i = \frac{m_i}{N},$$

где m_i – число значений, попавших в i–й интервал; N – общее число экспери–ментальных данных: $N = \sum_{i=1}^{n} m_i$, где n – число интервалов.

1.2. В системе координат $X - P$ построить гистограмму. Она представляет собой фигуру, состоящую из прямоугольников. Основанием их являются отрезки, изображающие интервалы вариационного ряда (см. табл. 1), а высоты равны значениям эмпирических частот (см. табл. 2).

Определить середину i–го интервала $X_{0i} = \frac{X_i + X_{i+1}}{2}$. Например, для варианта № 1, учитывая границы первого интервала, $X_{01} = \frac{0+2}{2} = 1$.

По условию задачи ширина интервалов $h = 2$.

Примечание. Для проверки гипотезы о соответствии эмпирического распределения нормальному закону при большом объеме выборки используется критерий χ^2. Для корректного применения данного критерия нужно, чтобы в крайних интервалах число значений было $m \geq 5$. Поэтому необходимо проверить исходные данные и, если нужно, перегруппировать их.

2. Проверить гипотезу о соответствии эмпирического распределения нормальному закону (закону Гаусса).

2.1. Рассчитать среднее арифметическое значение $\bar{x}$ результатов измерений (см. табл. 2): $\bar{x} = \frac{\sum_{i=1}^{n} m_i x_{0i}}{N}$.

2.2. Рассчитать среднее квадратичное отклонение (СКО):

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_{0i} - \bar{x})^2 \times m_i}{N - 1}}.$$

2.3. Определить теоретическую вероятность попадания значений измеряемой величины в i–й интервал:

$$P_{Ti} = \frac{h}{S} \varphi(U_i),$$

где $\varphi(U_i)$ – плотность нормированного нормального распределения,

$U_i = \frac{x_{0i} - \bar{x}}{S}$; нормированная нормальная величина (ордината кривой нормированного нормального распределения), $\varphi(U) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}$.

Для наглядности промежуточные результаты необходимо представить таблицей, например, как в табл. 3.

Таблица 3

Расчетные данные для проверки гипотезы о нормальности распределения

№ интервала $i \quad \overline{1,n}$	$m_i(x_{0i} - \bar{x})^2$	U_i	$\varphi(U_i)$	P_{Ti}	$\frac{(m_i - NP_{Ti})^2}{NP_{Ti}}$
1 2 … 10					
	$\sum =$			$\sum =$	$\sum =$

2.4. Провести аппроксимирующую прямую через точки P_{Ti}.

На рисунок нанести соответствующие точки P_{Ti} и соединить плавной прямой.

2.5. Проверка гипотезы о нормальности эмпирического распределения.

2.5.1. При $n > 40...50$ для проверки гипотезы можно пользоваться *критерием Пирсона.*

Определить расчетное значение критерия согласия Пирсона χ^2:

$$\chi^2_P = \sum_{i=1}^{n} \frac{(m_i - NP_{Ti})^2}{NP_{Ti}}.$$

2.5.2. Определить теоретическое значение критерия Пирсона $\chi^2_{теор}$.

В зависимости от доверительной вероятности и числа степеней свободы по статистическим таблицам (прил. 1), находим $\chi^2_{теор}$.

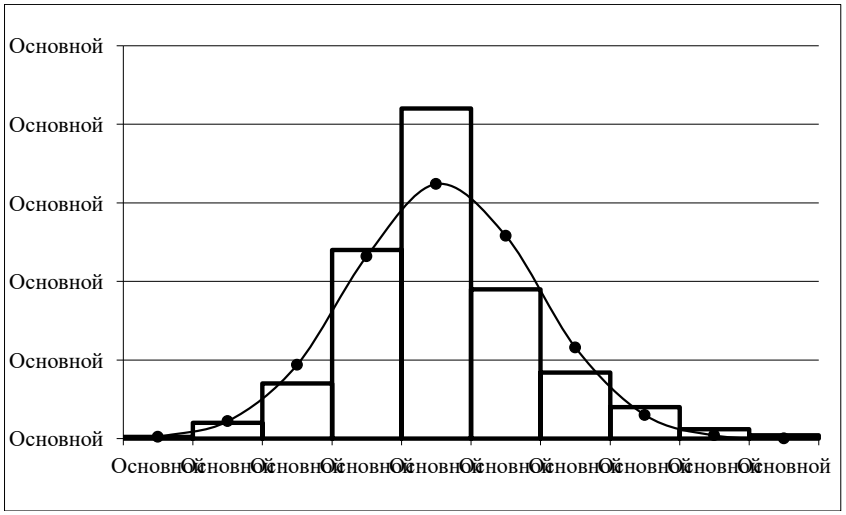
P = 0,975. Если используем уровень значимости α, то

$$\alpha = 1 - P = 1 - 0{,}975 = 0{,}025.$$

2.5.3. Сделать вывод о соответствии эмпирического распределения нормальному.

Если $\chi_P^2 > \chi_{r;\alpha}^2$ – гипотезу отвергают, если $\chi_P^2 \le \chi_{r;\alpha}^2$ – гипотезу принимают.

2.5.4. Результат представить в виде рисунка, например:



Гистограмма и кривая нормального распределения

3. Построить доверительный интервал для среднего значения:

$$\overline{X} - t_P S_{\overline{x}} \le X \le \overline{X} + t_P S_{\overline{x}},$$

где t_P – коэффициент распределения Стьюдента при заданной доверительной вероятности и числа степеней свободы (приложение 2); $S_{\overline{x}}$ – среднее квадратическое отклонение среднего значения: $S_{\overline{x}} = \frac{S}{\sqrt{n}}$.

Примечание. Если при заданном P в таблице отсутствует теоретическое значение критерия *t*, то для нахождения искомого значения *t* необходимо использовать интерполяционную формулу $y = y_0 + \frac{x - x_0}{x_1 - x_0}(y_1 - y_0)$.

ЗАДАЧА 2. ПРОВЕРКА ГИПОТЕЗЫ ПРИ МАЛОМ ОБЪЕМЕ ВЫБОРКИ

Дано. Произведены измерения условной величины (табл. 4).

Таблица 4

Значения условной величины

	Номер варианта																			
№ замера	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	71	10	20	12	13	15	6,6	6,2	8	4,6	50	20	37	33	43	150	109	167	100	148
2	59	11	21	18	20	19	2,7	3,8	8	3,5	7	39	23	44	42	169	148	152	152	225
3	76	14	9	9	11	16	6,1	6,4	7	5,3	39	29	44	27	43	127	124	95	127	134
4	74	8	8	17	18	16	5,3	4,1	8	6,5	20	69	30	33	38	131	173	133	222	9
5	81	7	8	24	10	23	4,9	5,5	6	6,4	69	42	38	54	30	186	152	146	87	258
6	84	18	15	17	13	15	6,0	5,0	7	7,1	49	68	48	39	38	129	127	82	188	166
7	77	30	17	14	10	20	0,6	4,8	7	2,4	24	38	45	42	36	165	156	157	129	225
8	80	16	12	19	12	20	8,1	4,5	5	4,7	50	69	48	32	31	138	100	65	140	286
9	76	14	9	13	19	16	5,5	5,6	6	6,3	62	30	14	25	42	98	105	189	122	105
10	74	6	17	16	13	13	4,8	3,9	6	3,7	49	42	12	36	22	160	162	153	53	99
11	71	20	10	17	20	14	6,2	2,6	7	4,2	34	29	26	42	36	137	111	152	86	182
12	73	10	12	21	13	17	4,5	3,7	8	3,0	21	54	29	29	50	144	173	199	204	214
13	84	12	14	14	14	13	6,5	5,8	6	2,8	38	38	48	47	40	131	132	165	188	191
14	76	15	6	5	11	14	6,1	3,4	6	3,8	10	40	34	31	19	161	160	120	88	101
15	82	20	15	19	18	8	4,7	3,9	7	4,1	27	38	32	42	49	106	111	154	168	45
16	85	13	10	13	13	15	4,4	4,5	9	2,5	58	30	48	39	55	142	74	148	148	228
17	60	17	23	25	15	13	4,9	4,8	5	4,3	42	40	33	49	28	100	158	104	132	240
18	82	20	14	17	16	17	9,3	5,0	5	4,5	36	48	30	34	39	143	162	175	90	193
19	72	17	8	19	15	12	2,5	5,5	6	5,2	42	28	27	36	43	119	176	131	206	171
20	86	8	9	23	12	15	6,0	3,1	8	4,6	32	52	27	52	48	150	143	136	103	65
21	72	18	10	10	14	13	9,3	5,0	6	4,6	17	31	23	60	35	153	112	163	132	–41
22	72	11	9	27	17	14	7,7	4,3	7	4,6	45	47	28	43	33	117	135	110	201	114
23	84	8	12	12	14	10	3,7	4,6	6	6,6	22	63	32	33	29	157	171	72	186	103
24	92	15	18	19	18	14	3,9	6,6	6	4,9	34	50	26	38	26	114	143	95	194	103
25	70	21	13	29	14	13	3,7	6,3	8	4,8	50	33	40	43	39	125	64	58	161	135
26	68	20	12	15	9	15	6,8	3,5	6	4,4	44	60	30	38	35	123	186	149	146	299
27	78	14	21	18	19	22	6,7	5,0	8	7,4	57	33	23	30	31	146	162	123	108	64
28	67	14	17	15	14	14	4,7	7,7	6	6,0	30	45	30	39	34	137	156	194	159	223
29	79	15	6	24	19	13	3,9	5,1	6	7,9	68	44	38	45	36	143	157	204	187	–50
30	76	24	9	8	6	16	4,9	4,4	5	4,2	22	21	49	39	44	107	145	42	49	36
*P**	0,95	0,9	0,95	0,95	0,99	0,9	0,95	0,99	0,9	0,95	0,9	0,95	0,9	0,99	0,95	0,9	0,95	0,99	0,9	0,95

*P** – доверительная вероятность.

Требуется:

1. Построить гистограмму эмпирического распределения.

2. Проверить гипотезу о соответствии эмпирического распределения нормальному закону (закону Гаусса).

3. Построить доверительный интервал для среднего значения измеряемой величины.

При 10...15 < N < 40...50 можно пользоваться составным критерием d. В этом случае рассчитывается критерий $d_{\text{расч.}}$

$$d = \frac{\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}_n\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}$$

и проверяется выполнение условия

$$d_{min} \le d_{\text{расч}} \le d_{max},$$

где d_{min} и d_{max} зависят от вероятности P*, с которой принимается решение, и находятся по статистическим таблицам (прил. 3).

Если это условие соблюдается, то дополнительно проверяются «хвосты» эмпирического закона распределения вероятности.

При $10 \le n \le 20$ считается допустимым отклонение одного из независимых значений результата измерения x_i от среднего арифметического больше, чем на $2{,}5S_x$.

При $20 < n \le 50$ – допускается отклонение двух значений, что соответствует доверительной вероятности P** $\approx 0{,}98$.

Несоблюдение хотя бы одного из двух условий достаточно для того, чтобы гипотеза о нормальности закона распределения вероятности результата измерения была отвергнута. В противном случае гипотеза принимается с вероятностью $P \ge P^* + P^{**} - 1$.

При $n < 10...15$ гипотеза о том, что результат измерения подчиняется нормальному закону распределения вероятности, не проверяется. Решение принимается на основании анализа априорной информации.

От принятой гипотезы относительно закона распределения вероятности результата измерения зависит вид оценки его среднего значения.

Методические рекомендации по решению задачи

Для удобства решения задачи необходимые исходные данные и промежуточные результаты вычислений целесообразно представить в табличном виде, например, табл. 5.

Таблица 5

Расчетные данные для проверки гипотезы о нормальности распределения

№ замера $i = \overline{1,n}$	x_i	$\bar{x}$	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
1 2 … n				
			$\sum =$	$\sum =$

Рекомендуемая литература

1. Громыко Г.Л. Общая теория статистики: Практикум.– М.:ИНФРА–М, 2000.– 139 с.

2. Маркин Н.С. Основы теории обработки результатов измерений. М.: Изд–во стандартов, 1991.– 268 с.

ПРИЛОЖЕНИЕ 1

Значения χ_α^2, удовлетворяющие условию $P(\chi^2>\chi_\alpha^2) = \alpha$

Число степеней свободы k	Уровень значимости α					
	0,005	0,010	0,025	0,050	0,100	0,200
5	16,70	15,10	12,80	11,10	9,24	7,29
6	18,50	16,80	14,40	12,60	10,60	8,56
7	20,30	18,50	16,00	14,10	12,00	9,80
8	22,00	20,10	17,50	15,50	13,40	11,00
9	23,60	21,70	19,00	16,90	14,70	12,20
10	25,20	23,20	20,50	18,30	16,00	13,40
15	32,80	30,60	27,50	25,00	22,30	19,30
20	40,0	37,6	34,2	31,4	28,4	25,0

ПРИЛОЖЕНИЕ 2

Значения коэффициента распределения Стьюдента

Число измерений n	При доверительной вероятности P				
	0,90	0,95	0,98	0,99	0,999
10	1,83	2,26	2,82	3,25	4,78
11	1,81	2,23	2,76	3,17	4,59
12	1,80	2,20	2,72	3,11	4,44
13	1,78	2,18	2,68	3,06	4,32
14	1,77	2,16	2,65	3,01	4,22
15	1,76	2,15	2,62	2,98	4,14
16	1,75	2,13	2,60	2,95	4,07
17	1,75	2,12	2,58	2,92	4,02
18	1,74	2,11	2,57	2,90	3,97
19	1,73	2,10	2,55	2,88	3,92
20	1,73	2,09	2,54	2,86	3,88
∞	1,65	1,96	2,33	2,58	3,29

ПРИЛОЖЕНИЕ 3

Граничные значения d

N	P^*=0,90		P^*=0,95		P^*=0,99	
	d_{min}	d_{max}	d_{min}	d_{max}	d_{min}	d_{max}
11	0,7409	0,8899	0,8899	0,9073	0,6675	0,9359
16	0,7452	0,8733	0,8733	0,8884	0,6829	0,9137
21	0,7495	0,8631	0,8631	0,8768	0,6950	0,9001
26	0,7530	0,8570	0,8570	0,8686	0.7040	0,8901
31	0,7559	0,8511	0,8511	0,8625	0,7110	0,8827
36	0,7583	0,8468	0,8468	0,8578	0,7167	0,8769
41	0,7604	0,8436	0,8436	0,8540	0,7216	0,8722
46	0,7621	0,8409	0,8409	0,8508	0,7256	0,8682
51	0,7636	0,8385	0,8385	0,8481	0,7291	0,8648

ДГТУ

Кафедра "Управление качеством"

Методические указания для выполнения практической работы по дисциплине «Методы анализа данных»

Задача 3. Выявление корреляционной связи, оценка ее значимость при различных уровнях доверительной вероятности.

Варианты задания

№	Параметры	Значения	α
1	Y	5,7; 5,7; 8,0; 8,0; 8,0; 12,0; 12,0; 12,0; 12,0; 14,3; 14,3; 15,6; 15,6; 15,6; 15,6; 17,8; 20,0; 20,0; 20,0; 20,0; 20,0; 25,0; 25,0; 25,0; 25,0; 25,0; 25,0; 25,0; 34,1; 34,1; 34,1; 34,1; 34,1; 34,1; 34,1; 34,1; 38,0; 38,0; 38,0; 40,3; 40,3	0,05 0,01
	X	83; 85; 66; 73; 70; 62; 61; 59; 57; 51; 49 ; 45; 42; 40; 48; 40; 37; 36 ;32; 28; 24; 23; 26; 20; 21; 20; 24; 15; 18; 17; 16; 17; 15; 10; 13; 10; 9; 10; 8; 7	
2	Y	13; 13; 13; 13; 13; 14,7; 14,7; 15,3; 15,3; 15,3; 15,3; 15,3; 15,3; 17,9; 17,9; 17,9; 17,9; 17,9; 17,9; 17,9; 19,1; 19,1; 19,1; 20; 20; 20; 23,4; 23,4; 23,4; 27,8; 27,8; 27,8; 30; 30; 30; 30; 34,5; 34,5; 36; 36	0,1 0,05
	X	3; 5; 7; 10; 5; 16; 18; 23; 26; 26; 28; 29; 30; 30; 35; 40; 37; 36; 32; 28; 40; 45; 50; 53; 56; 60; 67; 70; 78; 80; 90; 85 100; 100; 92; 110; 130; 120; 140; 150	
3	Y	17,5; 17,5; 17,5; 17,5; 17,5; 17,5; 17,5; 17,5; 20,4; 20,4; 20,4; 20,4; 23,7; 23,7; 25,1; 25,1; 25,1; 27,3 ; 27,3; 27,3; 27,3; 29,9; 29,9; 29,9; 29,9; 32; 32; 32; 32; 34,5; 34,5; 34,5; 34,5; 36,4; 36,4; 38; 38; 38; 38; 38	0,1 0,01
	X	100; 102; 110; 108; 110; 110; 90; 86; 85; 80; 78; 76; 70; 67; 60; 56; 53; 50; 45; 42; 40; 40; 39; 38; 37; 35; 30; 30; 29; 28; 23; 20; 18 ;18; 16; 10; 10; 7; 9; 5	
4	Y	33; 33; 35,1; 35,1; 35,1; 37,3; 37,3; 37,3; 37,3; 37,3; 37,3; 37,3; 38,4; 38,4; 38,4; 38,4; 39,1; 39,1; 39,1; 40,7; 40,7; 40,7; 40,7; 40,7; 40,7; 40,7; 43,5; 43,5; 43,5; 43,5; 43,5; 43,5; 45,6; 45,6; 48,9; 48,9; 48,9; 50,2; 50,2; 50,2	0,05 0,01
	X	17,5; 17,5; 17,5; 17,5; 17,5; 17,5; 17,5; 17,5; 20,4; 20,4; 20,4; 20,4; 23,7; 23,7; 25,1; 25,1; 25,1; 27,3 ; 27,3; 27,3; 27,3; 29,9; 29,9; 29,9; 29,9; 32; 32; 32; 32; 34,5; 34,5; 34,5; 34,5; 36,4; 36,4; 38; 38; 38; 38; 38	
5	Y	14,8; 14,8; 16,9; 18,4; 18,4; 18,4; 18,4; 18,4; 19,2; 19,2; 19,2; 19,2; 19,2; 21,7; 21,7; 21,7; 21,7; 21,7; 21,7; 21,7; 21,7; 21,7; 23,4; 25; 25; 25; 25; 27,5; 27,5; 27,5; 29,2; 29,2; 29,2; 29,2; 31,1; 31,1; 31,1; 31,1; 31,1; 31,1	0,05 0,1
	X	7; 3; 13; 9; 10; 15; 19; 13; 20; 25; 35; 30; 30; 34; 35; 37; 40; 35; 45; 45; 60; 55; 60; 65; 56; 65; 67; 80; 65; 90; 95; 78; 85; 85; 100; 115; 120; 110; 115; 100	

6	Y	44,1; 46,7; 46,7; 46,7; 46,7; 46,7; 46,7; 46,7; 46,7; 46,7; 46,7; 49; 49; 52,3; 52,3; 52,3; 52,3; 52,3; 56,4; 56,4; 56,4; 56,4; 56,4; 59,2; 59,2; 59,2; 64; 67,8; 67,8; 67,8; 69,5; 69,5; 69,5; 69,5; 69,5; 71,6; 71,6; 71,6; 71,6; 71,6	0,1 0,01
	X	120; 110; 115; 110; 115; 100; 95; 90; 100; 85; 95; 80; 90; 70; 75; 80; 71; 80; 56; 55; 50; 50; 60; 50; 40; 45; 35; 34; 30; 25; 25; 20; 19; 15; 13; 13; 10; 9; 7; 3	

Количественная оценка силы связи между исследуемыми факторами определяется посредством коэффициента корреляции *r* по формуле:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\cdot(Y_i - \overline{Y})}{n\cdot\sigma_X\cdot\sigma_Y}\ , \qquad (1)$$

где n – число пар значений исследуемых факторов; $\overline{X}$, $\overline{Y}$, σ_X, σ_Y – средние значения и среднеквадратические отклонения соответственно входного и выходного факторов.

Для удобства расчета коэффициента корреляции можно использовать формулу:

$$r = \frac{n\cdot\sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{\sqrt{n\cdot\sum_{i=1}^{n}X_i^2 - (\sum_{i=1}^{n}X_i)^2}\cdot\sqrt{n\cdot\sum_{i=1}^{n}Y_i^2 - (\sum_{i=1}^{n}Y_i)^2}}. \qquad (2)$$

Проверка значимости коэффициента корреляции.

Объем выборки значительный (n>100).

Коэффициент корреляции r, определенный по выборочным данным, может не совпадать с действительным значением, соответствующим генеральной совокупности (в силу малой представительности выборки).

Ошибка выборочного коэффициента парной корреляции определяется как:

$$S_r = \frac{1-r^2}{\sqrt{n-1}}. \qquad (3)$$

При n > 100 можно предположить, что коэффициент корреляции распределен нормально. Тогда справедливо выражение:

$$P\{r - X_r\cdot S_r \le r \le r + X_r\cdot S_r\},$$

где P – вероятность.

При P = 0,9 X_r = 1,653; P = 0,95 X_r = 1,96; P = 0,99 X_r = 2,576.

Объем выборки от 30 до 100. При малых n гипотеза о нормальном распределении коэффициента корреляции, как правило, не подтверждается. В этом случае для оценки значимости r применяют t-критерий Стьюдента.

$$t_{расч.} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (4)$$

Для числа степеней свободы $\nu = n - 2$ и уровня значимости α по статис-тическим таблицам находят теоретическое значение t-критерия Стьюдента ($t_{теор.}$).

Если $t_{расч.} \geq t_{теор.}$, то предположение о нулевом значении коэффициента корреляции не подтверждается.

Если $t_{расч.} < t_{теор.}$, то считается, что величина r незначимо отличается от нуля.

Доверительный интервал для оценки истинного значения коэффициента корреляции в генеральной совокупности (ρ) определяется как

$$r - t_{теор.} S_r \leq \rho \leq r + t_{теор.} S_r. \qquad (5)$$

Объем выборки n < 30.

Для оценки значимости r при малом объеме выборки целесообразно использовать z-преобразование Фишера. Статистика z определяется по формуле:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \qquad (6)$$

Распределение z асимптотически приближается к нормальному. Вариация z выражается формулой, которая распределена по нормальному закону со средним μ_z и дисперсией σ_z^2:

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \qquad \sigma_z^2 = \frac{1}{n-3}$$

Область принятия гипотезы о нулевой корреляции имеет вид:

$$-z_{\alpha/2} \leq \frac{\sqrt{n-3}}{2}\ln\left(\frac{1+r}{1-r}\right) < z_{\alpha/2}, \qquad (7)$$

где z – стандартная, нормально распределенная случайная величина. Если расчетное значение окажется вне этого интервала, то это будет признаком наличия статистической корреляции с уровнем значимости α.

Для $\alpha = 0{,}05$ $z_{\alpha/2} = 1{,}96$; $\alpha = 0{,}02$ $z_{\alpha/2} = 2{,}32$;

$\alpha = 0{,}01$ $z_{\alpha/2} = 2{,}58$; $\alpha = 0{,}1$ $z_{\alpha/2} = 1{,}64$.

ДГТУ

Кафедра управление качеством

Методические указания для выполнения практической работы по дисциплине «Методы анализа данных»

Задача 4. Проверка гипотезы о линейности корреляционной связи. Определение корреляционного отношения.

В случае линейной зависимости r = η. Если связь – нелинейная, то r < η. Это позволяет использовать η в качестве меры линейности связи между переменными X и У. Корреляционное отношение η определяется как

$$\eta_{yx} = \sqrt{\frac{\sum (y - \bar{y})^2 - (y - \overline{y_x})^2}{\sum (y - \bar{y})^2}}, \qquad (8)$$

где $\sum (y - \bar{y})^2$ – сумма квадратов отклонений индивидуальных значений у от общей средней арифметической $\bar{y}$; $\sum (y - \overline{y_x})^2$ – сумма квадратов отклонений вариант от групповых средних $\overline{y_x}$, соответствующих определенным, фиксированным значениям независимой переменной х.

Корреляционное отношение можно рассчитать и по такой формуле

$$\eta_{yx} = \sqrt{\frac{\sum (\overline{y_x} - \bar{y})^2}{\sum (y - \bar{y})^2}}, \qquad (9)$$

где $\sum (\overline{y_x} - \bar{y})^2$ – сумма квадратов отклонений групповых средних $\overline{y_x}$ от общей средней $\bar{y}$.

Значения η распределены на отрезке [0; 1]: $0 \le \eta \le 1$. Чем ближе η к 1, тем теснее связь между переменными X и У.

Проверка значимости корреляционного отношения осуществляется с помощью F-критерия Фишера. Его значение рассчитывается по формуле:

$$F_{расч.} = \frac{\eta^2 (n - m)}{(1 - \eta^2)(m - 1)}, \qquad (10)$$

где *n* – объем выборки; *m* – число групп.

Критическое значение *F* определяется по таблицам распределения Фишера (приложение А) по уровню значимости *a* и числу степеней свободы: $F_{теор.}(\alpha;\nu_1;\nu_2)$, где $\nu_1 = m - 1$; $\nu_2 = n - m$;

Расчетное значение $F_{расч.}$ необходимо сравнить с критическим $F_{кр}$.

По общему правилу проверки статистических гипотез:

- если $F_{расч.} < F_{кр}$, нулевую гипотезу о том, что η незначим, нельзя отклонить;
- если $F_{расч.} \geq F_{кр}$, нулевая гипотеза отклоняется в пользу альтернативной. Коэффициент η значимо отличается от нуля.

Пример расчета корреляционного отношения приведен в таблице 2

Таблица 2 – Методика расчета корреляционного отношения

X	**Y**	$\overline{Y_x}$	$Y - \overline{Y}$	$(Y - \overline{Y})^2$	$\overline{Y_x} - \overline{Y}$	$(\overline{Y_x} - \overline{Y})^2$	$Y - \overline{Y_x}$	$(Y - \overline{Y_x})^2$
3	66	66	31,2	975,36	31,2	975,36	0,00	0,00
5	59	70,7	24,2	587,13	35,9	1288,63	–11,67	136,11
5	66		31,2	975,36	35,9	1288,63	–4,67	21,78
5	87		52,2	2728,05	35,9	1288,63	16,33	266,78
7	80	80	45,2	2045,82	45,2	2045,82	0,00	0,00

Продолжение таблицы 2

9	31	35,7	–3,8	14,21	0,9	0,81	–4,67	21,78
9	31		–3,8	14,21	0,9	0,81	–4,67	21,78
9	45		10,2	104,67	0,9	0,81	9,33	87,11
11	24	43,0	–10,8	115,98	8,2	67,75	–19,00	361,00
11	31		–3,8	14,21	8,2	67,75	–12,00	144,00
11	31		–3,8	14,21	8,2	67,75	–12,00	144,00
11	31		–3,8	14,21	8,2	67,75	–12,00	144,00
11	52		17,2	296,90	8,2	67,75	9,00	81,00
11	66		31,2	975,36	8,2	67,75	23,00	529,00
11	66		31,2	975,36	8,2	67,75	23,00	529,00
13	17	30,4	–17,8	315,75	–4,4	19,41	–13,36	178,59
13	24		–10,8	115,98	–4,4	19,41	–6,36	40,50
13	24		–10,8	115,98	–4,4	19,41	–6,36	40,50
13	31		–3,8	14,21	–4,4	19,41	0,64	0,40
13	31		–3,8	14,21	–4,4	19,41	0,64	0,40
13	31		–3,8	14,21	–4,4	19,41	0,64	0,40

13	31		–3,8	14,21	–4,4	19,41	0,64	0,40
13	31		–3,8	14,21	–4,4	19,41	0,64	0,40
13	38		3,2	10,44	–4,4	19,41	7,64	58,31
13	38		3,2	10,44	–4,4	19,41	7,64	58,31
13	38		3,2	10,44	–4,4	19,41	7,64	58,31
15	24	33,3	–10,8	115,98	–1,4	2,06	–9,33	87,11
15	38		3,2	10,44	–1,4	2,06	4,67	21,78
17	10	19,3	–24,8	613,51	–15,4	238,27	–9,33	87,11
17	17		–17,8	315,75	–15,4	238,27	–2,33	5,44
17	17		–17,8	315,75	–15,4	238,27	–2,33	5,44
17	17		–17,8	315,75	–15,4	238,27	–2,33	5,44
17	24		–10,8	115,98	–15,4	238,27	4,67	21,78
17	31		–3,8	14,21	–15,4	238,27	11,67	136,11
19	3	7,7	–31,8	1009,28	–27,1	734,55	–4,67	21,78
19	3		–31,8	1009,28	–27,1	734,55	–4,67	21,78
19	17		–17,8	315,75	–27,1	734,55	9,33	87,11
21	17	17	–17,8	315,75	–17,8	315,75	0,00	0,00
n=39				14978,92		11532,38		3446,55

Среднее значение $\overline{Y}$ = 34,8.

Подставляя в формулу (8) значения из таблицы 2 получаем:

$$\eta_{yx} = \sqrt{\frac{\sum (y - \overline{y})^2 - (y - \overline{y_x})^2}{\sum (y - \overline{y})^2}} = \sqrt{\frac{14978{,}92 - 3446{,}55}{14978{,}92}} = 0{,}877.$$

По формуле (9) имеем (проверка):

$$\eta_{yx} = \sqrt{\frac{\sum (\overline{y_x} - \overline{y})^2}{\sum (y - \overline{y})^2}} = \sqrt{\frac{11532{,}38}{14978{,}92}} = 0{,}877.$$

Для проверки гипотезы значимости связи определим значение $F_{расч}$.

$$F_{расч.} = \frac{\eta^2 (n - m)}{(1 - \eta^2)(m - 1)} = \frac{0{,}877^2 (39 - 10)}{(1 - 0{,}877^2)(10 - 1)} = 10{,}734.$$

В нашем примере n = 39 и m = 10. Тогда $\nu_1 = 10 - 1 = 9$; $\nu_2 = 39 - 10 = 29$.

По приложению А находим, что $F_{кр.}(0{,}05; 9; 29) = 2{,}2$.

Так как $F_{расч.} \geq F_{кр}$, можно считать, что, коэффициент η значимо отличается от нуля.

Задания для самостоятельной работы

№ п/п	X	Y
1	7	4
2	7	1
3	7	5
4	7	9
5	7	9
6	7	15
7	7	13
8	7	13
9	13	20
10	13	25
11	17	29
12	17	30
13	17	30
14	17	25
15	17	35
16	21	37
17	21	40
18	21	40
19	21	45
20	21	45
21	25	55
22	27	55
23	27	60
24	27	65
25	27	56
26	27	65
27	29	67
28	29	75
29	32	80
30	32	90
31	32	100
32	32	78
33	32	80

34	32	85
35	35	100
36	35	115
37	35	120
38	35	110
39	38	140
40	38	135

ДГТУ

Кафедра "Управление качеством"

Методические указания для выполнения практической работы по дисциплине «Методы анализа данных»

Задача 5. Расчет коэффициентов множественной корреляции

Таблица 1 – Задания для выполнения практической работы

	Вариант 1	Вариант 2	Вариант 3	Вариант 4	Вариант 5
Объем выборки, n	15	22	18	34	65
r_{12}	0,6	0,5	0,65	0,7	0,64
r_{13}	0,3	0,3	0,41	0,27	0,31
r_{23}	- 0,2	-0,3	-0,35	-0,33	-0,36

Множественная корреляция

При изучении сложных явлений необходимо учитывать более двух случайных факторов. Правильное представление о природе связи между этими факторами можно получить только в том случае, если подвергнуть исследованию сразу все рассматриваемые случайные факторы. Совместное изучение трех и более случайных факторов позволит исследователю установить более или менее обоснованные предположения о причинных зависимостях между изучаемыми факторами. Простой формой множественной связи является линейная зависимость между тремя признаками. Случайные факторы обозначаются как X_1, X_2 и X_3. Парные коэффициенты корреляции между X_1 и X_2 обозначается как r_{12}, соответственно между X_1 и X_3 – r_{13}, между X_2 и X_3 – r_{23}. В качестве меры тесноты линейной связи трех факторов используют множественные коэффициенты корреляции, обозначаемые R_{123}, R_{213}, R_{312} и частные коэффициенты корреляции, обозначаемые $r_{12.3}$, $r_{13.2}$, $r_{23.1}$.

8.3.1 Множественная линейная корреляция

Множественный коэффициент корреляции $R_{1.23}$ трех факторов – это показатель тесноты линейной связи между одним из факторов (индекс перед точкой) и совокупностью двух других факторов (индексы после точки).

Значения коэффициента R всегда находятся в пределах от 0 до 1. При приближении R к единице степень линейной связи трех признаков увеличивается.

Между коэффициентом множественной корреляции, например R_{213}, и двумя коэффициентами парной корреляции r_{12} и r_{23} существует соотношение: каждый из парных коэффициентов не может превышать по абсолютной величине R_{213}. Формулы для вычисления множественных коэффициентов корреляции при известных значениях коэффициентов парной корреляции r_{12}, r_{13} и r_{23} имеют вид:

$$R_{1\cdot 23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}};$$

$$R_{2\cdot 13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}};$$

$$R_{3\cdot 12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

Квадрат коэффициента множественной корреляции R^2 называется *коэффициентом множественной детерминации.* Он показывает долю вариации зависимой переменной под воздействием изучаемых факторов.

Значимость множественной корреляции оценивается по *F*-критерию:

$$F = \frac{R^2}{1 - R^2}\left(\frac{n-1}{k-1}\right),$$

где n – объем выборки; k – число факторов. В нашем случае $k = 3$.

Нулевая гипотеза о равенстве нулю множественного коэффициента корреляции (H_0:R=0) принимается, если $F_{\text{расч.}} < F_т$, и отвергается, если $F_{\text{расч.}} \geq F_{\text{т}}$.

Теоретическое значение *F*-критерия определяется для $\nu_1 = k - 1$ и $\nu_2 = n - k$ степеней свободы и принятого уровня значимости α (прил. 1).

Пример вычисления коэффициента множественной корреляции. При изучении взаимосвязи между факторами были получены коэффициенты парной корреляции (n =15): r_{12}==0,6; r_{13} = 0,3; r_{23} = - 0,2.

Необходимо выяснить зависимость признака X_2 от признака X_1 и X_3, т. е. рассчитать коэффициент множественной корреляции:

$$R_{2\cdot13} = \sqrt{\frac{(0,6)^2 + (0,3)^2 - 2\times0,6\times0,3\times(-0,2)}{1-(0,2)^2}} = 0,74;$$

$$F = \frac{(0,74)^2}{1-(0,74)^2}\left(\frac{15-3}{2}\right) = 7,33.$$

Табличное значение F-критерия при $\nu_1 = 2$ и $\nu_2 = 15 - 3 = 12$ степенях свободы при $\alpha = 0,05$ $F_{0,05} = 3,89$ и при $\alpha = 0,01$ $F_{0,01} = 6,93$.

Таким образом, взаимосвязь между признаками $R_{2.13} = 0,74$ значима на 1%-м уровне значимости $F_{\text{расч}} > F_{0,01}$.

Судя по коэффициенту множественной детерминации $R^2 = (0,74)^2 = 0,55$, вариация признака X_2 на 55% связана с действием изучаемых факторов, а 45% вариации ($1\text{-}R^2$) не может быть объяснено влиянием этих переменных.

8.3.2 Частная линейная корреляция

Частный коэффициент корреляции – это показатель, измеряющий степень коррелированности двух факторов при учете остальных факторов.

Математическая статистика позволяет установить корреляцию между двумя факторами при постоянном значении третьего, не ставя специального эксперимента, а используя парные коэффициенты корреляции r_{12}, r_{13}, r_{23}.

Частные коэффициенты корреляции рассчитывают по формулам:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\cdot r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}; \quad r_{13\cdot2} = \frac{r_{13} - r_{12}\cdot r_{23}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{23}^2\right)}};$$

$$r_{23\cdot1} = \frac{r_{23} - r_{12}\cdot r_{13}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{13}^2\right)}};$$

Цифры перед точкой указывают, между какими факторами изучается зависимость, а цифра после точки – влияние какого фактора исключается (элиминируется). Ошибку и критерий значимости частной корреляции определяют по тем же формулам, что и парной корреляции:

$$S_{r_{12\cdot3}} = \sqrt{\frac{1-r_{12.3}^2}{n-2}}; \quad t = \frac{r}{s_r}. \qquad (8.12)$$

Теоретическое значение t-критерия определяется для степеней свободы $\nu = n - 2$ и принятого уровня значимости α.

Нулевая гипотеза о равенстве нулю частного коэффициента корреляции (H_0: $r = 0$) принимается, если $t_{\text{расч}} < t_{\text{т}}$, и отвергается, если $t_{\text{расч}} \geq t_{\text{т}}$.

Частные коэффициенты могут принимать значения, заключенные между –1 и +1. Частные *коэффициенты детерминации* находят путем возведения в

квадрат частных коэффициентов корреляции:

$$D_{12.3} = r^2_{123}; \quad d_{13.2} = r^2_{132}; \quad d_{231} = r^2_{231}.$$

Определение степени частного воздействия отдельных факторов на результативный признак при исключении (элиминировании) связи его с другими признаками, искажающими эту корреляцию, часто представляет большой интерес. Иногда бывает, что при постоянном значении элиминируемого фактора нельзя подметить его статистического влияния на изменчивость других факторов.

Чтобы уяснить технику расчета частного коэффициента корреляции, рассмотрим пример.

Таблица 2 – Задания для выполнения практической работы

	Вариант 1	Вариант 2	Вариант 3	Вариант 4	Вариант 5
Объем выборки, n	15	22	18	34	65
R_{xy}	0,69	0,75	0,65	0,7	0,64
R_{xz}	0,35	0,39	0,48	0,57	0,51
R_{yz}	0,29	0,3	0,55	0,33	0,44

Имеются три параметра *X*, *Y* и *Z*. Для объема выборки $n = 180$ определены парные коэффициенты корреляции

$$R_{xy} = 0{,}799;\ r_{xz} = 0{,}57;\ r_{yz} = 0{,}507.$$

По формуле (8.12) определим частные коэффициенты корреляции:

$$r_{xy\cdot z} = \frac{0{,}799 - 0{,}570 \times 0{,}507}{\sqrt{\left(1 - 0{,}570^2\right)\left(1 - 0{,}507^2\right)}} = 0{,}720;$$

$$r_{xz\cdot y} = \frac{0{,}570 - 0{,}799 \times 0{,}507}{\sqrt{\left(1 - 0{,}799^2\right)\left(1 - 0{,}507^2\right)}} = 0{,}318;$$

$$r_{yz\cdot x} = \frac{0{,}507 - 0{,}799 \times 0{,}570}{\sqrt{\left(1 - 0{,}799^2\right)\left(1 - 0{,}570^2\right)}} = 0{,}105.$$

Частный коэффициент корреляции между фактором *X* и *Y* с постоянным значением фактора *Z* ($r_{xyz} = 0{,}720$) показывает, что лишь незначительная часть взаимосвязи этих признаков в общей корреляции ($r_{xy} = 0{,}799$) обусловлена влиянием третьего фактора (*Z*). Аналогичное заключение необходимо сделать

и в отношении частного коэффициента корреляции между фактором X и фактором Z с постоянным значением фактора Y (r_{xzy} = 0,318 и r_{xz} = 0,57). Напротив, частный коэффициент корреляции между факторами Y и Z с постоянным значением фактора X r_{yzx} = 0,105 значительно отличается от общего коэффициента корреляции r_{yz} = 0,507. Из этого видно, что если подобрать объекты с одинаковым значением фактора X, то связь между факторами Y и Z у них будет очень слабой, так как значительная часть в этой взаимосвязи обусловлена варьированием фактора X.

При некоторых обстоятельствах частный коэффициент корреляции может оказаться противоположным по знаку парному.

Например, при изучении взаимосвязи между факторами X, Y и Z были получены парные коэффициенты корреляции (при n = 100): r_{xy} = 0,6; r_{xz}= 0,9; r_{yz} = 0,4.

Частные коэффициенты корреляции при исключении влияния третьего фактора:

$$r_{xy\cdot z} = \frac{0,6 - 0,9 \times 0,4}{\sqrt{\left(1 - 0,9^2\right)\left(1 - 0,4^2\right)}} = 0,60;$$

$$r_{xz\cdot y} = \frac{0,9 - 0,6 \times 0,4}{\sqrt{\left(1 - 0,6^2\right)\left(1 - 0,4^2\right)}} = 0,90;$$

$$r_{yz\cdot x} = \frac{0,4 - 0,6 \times 0,9}{\sqrt{\left(1 - 0,6^2\right)\left(1 - 0,9^2\right)}} = -0,40.$$

Из примера видно, что значения парного коэффициента и частного коэффициента корреляции разнятся в знаке.

Метод частной корреляции дает возможность вычислить частный коэффициент корреляции второго порядка. Этот коэффициент указывает на взаимосвязь между первым и вторым фактором при постоянном значении третьего и четвертого. Определение частного коэффициента второго порядка ведут на основе частных коэффициентов первого порядка по формуле:

$$r_{12\cdot 34} = \frac{r_{12\cdot 4} - r_{13\cdot 4} \cdot r_{23\cdot 4}}{\sqrt{\left(1 - r_{13\cdot 4}^2\right)\left(1 - r_{23\cdot 4}^2\right)}},$$

где $r_{12.4}$, r_{134}, r_{234} — частные коэффициенты первого порядка, значение которых определяют по формуле частного коэффициента, используя коэффициенты парной корреляции r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34}.

1. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Методические указания к практической работе по дисциплине «Методы анализа данных»

Для характеристики формы связи при изучении корреляционной зависимости между выходным параметром и переменным фактором, при обработке результатов однофакторных экспериментов, используются уравнения приближенной регрессии [1]. Задача ставится следующим образом: по данной выборке объёма n найти уравнение приближенной регрессии и оценить допускаемую при этом ошибку. В качестве метода приближения обычно выбирают метод наименьших квадратов (МНК).

Суть метода заключается в том, что вид зависимости и значения коэффициентов описывающего ее уравнения должны обеспечивать минимальную сумму квадратов отклонений (Ф) ординат экспериментальных точек от ординат этой зависимости [2]:

$$\Phi = \sum_{i=1}^{n} (y_i - \tilde{y})^2 = min, \tag{1.1}$$

где y_i – рассчитанное по уравнению регрессии значение выходного параметра, а $\tilde{y}$ – экспериментальное значение выходного параметра, полученное при том же значении переменного фактора x_i.

Задача определения коэффициентов уравнения регрессии по МНК сводится к определению минимума функции многих переменных [1]. Если:

$$y = f(x, b_0, b_1, \dots, b_R), \tag{1.2}$$

и требуется выбрать коэффициенты $b_0, b_1, \dots, b_R$ таким образом, чтобы:

$$\Phi = \sum_{i=1}^{n} [\tilde{y}_i - f(x_i, b_0, b_1, \dots, b_R]^2 = \min \tag{1.3}$$

то необходимым условием минимума $\Phi(b_0, b_1, \dots, b_R)$ будет являться выполнение равенств:

$$\frac{\partial \Phi}{\partial b_0} = 0, \qquad \frac{\partial \Phi}{\partial b_1} = 0, \dots, \frac{\partial \Phi}{\partial b_R} = 0 \tag{1.4}$$

Т.е. минимум данной функции будет в точке, где её частные производные равны нулю.

Условие (1.4) можно записать в виде:

$$\begin{gathered}\sum_{i=1}^{n} 2[\tilde{y}_i - f(x_i, b_0, b_1, \dots, b_R)]\frac{\partial f(x_i)}{\partial b_0} = 0,\\ \sum_{i=1}^{n} 2[\tilde{y}_i - f(x_i, b_0, b_1, \dots, b_R)]\frac{\partial f(x_i)}{\partial b_1} = 0,\\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots,\\ \sum_{i=1}^{n} 2[\tilde{y}_i - f(x_i, b_0, b_1, \dots, b_R)]\frac{\partial f(x_i)}{\partial b_R} = 0\end{gathered} \qquad (1.5)$$

После преобразования:

$$\begin{gathered}\sum_{i=1}^{n} \tilde{y}_i \frac{\partial f(x_i)}{\partial b_0} - \sum_{i=1}^{n} f(x_i, b_0, b_1, \dots, b_R)\frac{\partial f(x_i)}{\partial b_0} = 0,\\ \sum_{i=1}^{n} \tilde{y}_i \frac{\partial f(x_i)}{\partial b_1} - \sum_{i=1}^{n} f(x_i, b_0, b_1, \dots, b_R)\frac{\partial f(x_i)}{\partial b_1} = 0,\\ \sum_{i=1}^{n} \tilde{y}_i \frac{\partial f(x_i)}{\partial b_R} - \sum_{i=1}^{n} f(x_i, b_0, b_1, \dots, b_R)\frac{\partial f(x_i)}{\partial b_R} = 0,\end{gathered} \qquad (1.6)$$

Система уравнений (1.6) имеет столько же уравнений, сколько неизвестных коэффициентов $b_0, b_1, \dots, b_R$ входит в уравнение регрессии, и называется системой нормальных уравнений.

При изучении зависимости выходного параметра от одного переменного фактора необходимо построить эмпирическую линию регрессии для определения вида уравнения регрессии [1]. Для этого весь диапазон изменения x на поле корреляции разбивается на k равных интервалов Δx. Все точки, попавшие в данный интервал Δx_j, относят к его середине x_j. Для этого подсчитывают частные средние $\bar{y}_j$ для каждого интервала:

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{ji}}{n_j}, \qquad (1.7)$$

где y_{ji} – экспериментальные значения выходного параметра, попавшие в интервал Δx_j, а n_j – количество значений y_{ji}.

Затем последовательно соединяют точки $(x_j; \bar{y}_j)$ отрезками прямой. Полученная ломаная называется эмпирической линией регрессии y по x. По виду эмпирической линии регрессии можно подобрать уравнение регрессии $y = f(x)$.

Для линейной зависимости $y = ax + b$ условие (1.4) будет иметь вид:

$$\begin{cases} \dfrac{\partial \Phi}{\partial b} = \sum_{i=1}^{n} [y_i - (ax_i + b)] = 0 \\ \dfrac{\partial \Phi}{\partial a} = \sum_{i=1}^{n} [y_i - (ax_i + b)] x_i = 0 \end{cases} \quad (1.8)$$

Для определения коэффициентов a и b линейного уравнения будем иметь систему линейных уравнений (1.9):

$$\begin{cases} bn + a\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b\sum_{i=1}^{n} x_i + a\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i \times y_i) \end{cases} \quad (1.9)$$

Решение системы уравнений (1.9) относительно a и b дает следующие формулы для их расчета:

$$b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x)^2} \quad (1.10)$$

$$a = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \quad (1.11)$$

Аналогичным образом, с помощью МНК можно получить формулы для расчета коэффициентов нелинейных зависимостей (1.12) – (1.18) [2]:

Логарифмическая зависимость $y = a\ln x + b$, $x_i > 0$, x ≠ 0

$$\begin{cases} a = \dfrac{\sum y_i \ln x_i - \frac{1}{n}\sum y_i \sum \ln x_i}{\sum (\ln x_i)^2 - \frac{1}{n}\sum (\ln x_i)^2} \\ b = \dfrac{\sum y_i a \sum \ln x_i}{n} \end{cases} \quad (1.12)$$

Экспоненциальная функция $y = be^{ax}$, все y_i и $x_i > 0$, $y_i \neq 0$

$$\begin{cases} a = \dfrac{\sum (\ln y_i)\, x_i - \frac{1}{n}\sum x_i \sum \ln y_i}{\sum x_i^2 - \frac{1}{n}\sum (x_i)^2} \\ b = exp\left[\dfrac{\sum \ln y_i - a\sum x_i}{n}\right] \end{cases} \quad (1.13)$$

Степенная функция $y = ax^b$, $x_i \neq 0$, $y_i \neq 0$, все y_i и $x_i > 0$

$$\begin{cases} b = \dfrac{\sum \ln x_i \ln y_i - \frac{1}{n}\sum \ln x_i \sum \ln y_i}{\sum (\ln x_i)^2 - \frac{1}{n}[\sum \ln x_i]^2} \\ a = exp\left[\dfrac{\sum \ln y_i - b\sum \ln x_i}{n}\right] \end{cases} \quad (1.14)$$

Дробно-линейная функция $y = x/(ax + b)$, $y_i \neq 0, x_i \neq 0$

$$\begin{cases} a = \dfrac{\sum \frac{x_i^2}{y_i} - \frac{1}{n}\sum x_i \sum \frac{x_i}{y_i}}{\sum x_i^2 - \frac{1}{n}\sum x_i^2} \\ b = \dfrac{\sum \frac{x_i}{y_i} - a\sum x_i}{n} \end{cases} \quad (1.15)$$

Гиперболическая функция $y = a/x + b$, $x_i > 0$

$$\begin{cases} b = \dfrac{\sum x_i^2 y_i - \frac{1}{n}\sum x_i \sum x_i^2}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\ a = \dfrac{\sum x_i^2 - b\sum x_i}{n} \end{cases} \quad (1.16)$$

Дробно-рациональная функция $y = 1/(ax + b)$, $y_i \neq 0$

$$\begin{cases} a = \dfrac{\sum \frac{x_i}{y_i} - \frac{1}{n}\sum x_i \sum \frac{1}{y_i}}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\ b = \dfrac{\sum \frac{1}{y_i} - a\sum x_i}{n} \end{cases} \quad (1.17)$$

Квадратичная (параболическая) функция $y = ax^2 + bx + c$

$$\begin{cases} cn + b\sum x_i + a\sum x_i^2 = \sum y_i \\ c\sum x_i + b\sum x_i^2 + a\sum x_i^3 = \sum y_i x_i \\ c\sum x_i^2 + b\sum x_i^3 + a\sum x_i^4 = \sum x_i^2 y_i \end{cases} \quad (1.18)$$

Точность описания корреляционной связи между параметром выхода и переменным фактором нагляднее всего характеризует средняя погрешность аппроксимации ($\bar{\delta}, \%$), которая рассчитывается по следующей формуле:

$$\bar{\delta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left|\frac{y_i - \tilde{y}_i}{y_i}\right| \times 100\%\right\} \quad (1.19)$$

Очевидно, что лучшей зависимостью для описания связи между x и y будет та, которая обеспечивает минимальную среднюю погрешность аппроксимации $\bar{\delta} \to min$.

2. МОДЕЛЬНЫЙ ПРИМЕР

Приведена зависимость напряжённости электростатического поля (y) в процессе гидромониторной промывки ёмкости после нефтепродукта от диэлектрической проницаемости моющей жидкости (x). Результаты эксперимента представлены в таблице 2.1.

Таблица 2.1 – Результаты эксперимента

№ п/п	X	Y	№ п/п	X	Y	№ п/п	X	Y
1	50	32	11	55	54	21	62	72
2	51	35	12	56	50	22	63	71,5
3	52	29	13	57	60	23	63,5	73
4	52,5	31,5	14	58	52	24	64	75
5	53	30	15	59	65	25	64,5	74
6	54	35	16	60	69	26	65	77
7	54,5	38	17	61	68,5	27	66	75

Построим эмпирическую линию регрессии, чтобы определить вид уравнения.

Для этого разобьем рассматриваемый диапазон значений x на n равных интервалов. Для определения n числа интервалов воспользуемся формулой Стерджеса:

$$n = 1 + 3{,}322lgN = 1 + 3{,}322lg21 = 5{,}39 \quad (2.1)$$

Округляем полученный результат до 5.

Ширина интервала h будет равна

$$\frac{X_{max} - X_{min}}{h} = \frac{66 - 50}{5} = 3{,}2 \quad (2.2)$$

Таким образом получим следующие границы и частные средние значения интервалов (таблица 2.2).

Таблица 2.2 – Границы и частные средние значения интервалов

Интервал	Границы интервала	Частные средние $\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{ji}}{n_j}$
1	[50;53,2]	31,5
2	(53,2;56,4]	44,25
3	(56,4;59,6]	59
4	(59,6;62,8]	69,8
5	(62,8;66]	74,25

Соединим полученные точки отрезками прямой (рисунок 2.1).

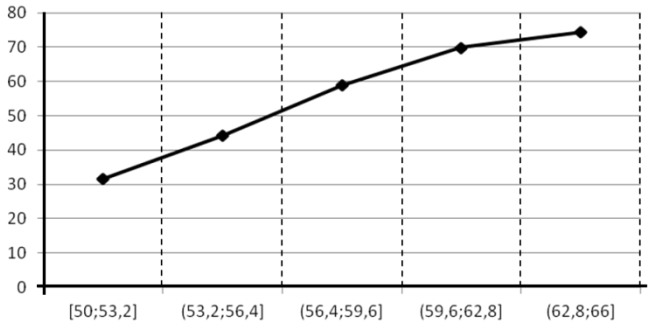


Рисунок 2.1 – Эмпирическая линия регрессии

Исходя из полученного графика, можно предположить, что данная зависимость имеет линейный, либо степенной вид. Соответственно, необходимо найти коэффициенты для двух уравнений регрессии: $y = ax + b$ и $y = ax^b$.

Линейное уравнение $y = ax + b$.

Коэффициенты линейного уравнения определяются по формулам (1.10) и (1.11.)

Промежуточные расчёты представлены в таблице 2.3

Таблица 2.3 – Промежуточные расчёты

Величина	Значение
∑X	1221
∑У	1166,5
$\sum X^2$	71515
∑XУ	69589,25
$(\sum X)^2$	1490841

$$b = \frac{1166{,}5 \times 71515 - 1221 \times 69589{,}25}{21 \times 71515 - 1490841} = -140{,}9$$

$$a = \frac{21 \times 69589{,}25 - 1221 \times 1166{,}5}{21 \times 71515 - 1490841} = 3{,}38$$

Степенное уравнение $y = ax^b$

Коэффициенты данного уравнения будут определяться по формуле (1.14). a=22,296; b=0,3954.

Проверим среднюю погрешность аппроксимации ($\bar{\delta}, \%$) для каждого из построенных уравнений регрессии. Расчёт проводится по формуле (1.19).

$$\bar{\delta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left|\frac{y_i - \tilde{y}_i}{y_i}\right| \times 100\%\right\}$$

$\bar{\delta}_{\text{лин.}} = 9{,}83\%$; $\bar{\delta}_{\text{степен.}} = 126{,}66\%$

Исходя из значений средней погрешности аппроксимации можно сделать вывод, что уравнение регрессии линейного вида точнее описывает эмпирическую зависимость.

3 ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ.

Используя данные из таблицы 3.1 построить уравнения регрессии, описывающие зависимость Y от X.

Таблица 3.1 – Индивидуальные задания.

№п/п	X	1	2	3	4	5
		Y	Y	Y	Y	Y
1	50	4,3	150	25	7,8	1,9
2	51	3,5	155	52	6,5	1,5
3	52	4,7	157	40	8,65	2,1
4	52,5	4	158	44	7,25	1,8
5	53	4,4	160	38	8,15	2
6	54	4,5	163	53	8,3	2
7	54,5	5	165	30	9	2,1
8	55	4	166	60	7,4	1,8
9	56	4,5	170	47	8,2	1,9
10	57	4,9	172	51	9,1	2,1
11	58	4	175	40	7,6	1,8
12	59	4,6	178	62	8,6	2
13	60	4,6	181	72	8,8	2
14	61	5	184	75	9,5	2,2
15	62	5,1	187	79	7,8	1,8
16	63	4,5	190	84	8,3	2
17	63,5	5	192	64	9,6	2,3
18	64	4,2	193	81	8	1,9
19	64,5	4,6	194	69	9	2,1
20	65	4,7	196	79	9,1	2,2
21	66	5,1	199	80	10	2,3
22	67,2	4,2	202	82	8,1	2
23	68	4,65	205	91	9,1	2,4
24	69	5,1	208	95	9,9	2,5
25	69,5	4,2	209	93	8,4	2,4
26	70	4,8	211	97	9,4	2,6

ПОСТРОЕНИЕ РЕГРЕССИОННОЙ МОДЕЛИ

Методические указания к практическим занятиям
по дисциплине "Методы анализа данных"

Ростов-на-Дону
2021

ВВЕДЕНИЕ

К задачам регрессионного анализа относятся:
1) установление формы зависимости; 2) определение уравнения регрессии т.е. определение неизвестных коэффициентов модели; 3) оценка неизвестных значений зависимой переменной.

Различают линейную и нелинейную связи.

Линейная связь имеет место, когда с возрастанием (или убыванием) значений X значения Y увеличиваются (или уменьшаются) более или менее равномерно.

Математически линейная связь может быть выражена уравнением прямой, которое называется линейным *уравнением регрессии:*

$$Y_{теор} = b_0 + b_1 \cdot X , \qquad (1)$$

где X- факторный признак; $Y_{теор}$ – результативный признак; b_0, b_1- коэффициенты уравнения.

Если же она выражается уравнением какой-либо кривой линии (параболы, гиперболы, степенной, показательной, экспоненциальной и т. д.), то такую связь называют **нелинейной.** На практике часто пользуются *уравнением параболы второго порядка:*

$$Y_{теор} = b_0 + b_1 \cdot X + b_2 \cdot X^2 \qquad (2)$$

Уравнение нелинейной связи может быть выражено и в виде *уравнения гиперболы:*

$$Y_{теор} = b_0 + b_1/X$$

или показательной функции:

$$Y_{теор} = b_0 \cdot b_1^X$$

После определения формы связи, т.е. вида уравнения регрессии, по эмпирическим данным определяют коэффициенты искомого уравнения.

1 РАСЧЕТ КОЭФФИЦИЕНТОВ УРАВНЕНИЯ РЕГРЕССИИ

Коэффициенты b_1 и b_0 уравнения (1) определяются по формулам:

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

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

Для экспоненциальной (степенной) зависимости

$$y = b_0 \cdot e^{b_1 \cdot x}$$

коэффициенты b_1 и b_0 определяются по формулам

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

$$b_0 = \exp\left[\frac{1}{n} \cdot \left(\sum_{i=1}^{n} \ln y_i - b_1 \cdot \sum_{i=1}^{n} x_i\right)\right] \quad (14)$$

Для параболическая зависимости

$$y = b_0 + b_1 x + b_2 x^2 .$$

Коэффициенты b_0, b_1, b_2 определяются при решении системы из трех уравнений (например, методом Гаусса):

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum_{i=1}^{n} x_i + b_2 \cdot \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} y_i; \\ b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} x_i^{\,2} + b_2 \cdot \sum_{i=1}^{n} x_i^{\,3} = \sum_{i=1}^{n} x_i \cdot y_i; \\ b_0 \cdot \sum_{i=1}^{n} x_i^{\,2} + b_1 \cdot \sum_{i=1}^{n} x_i^{\,3} + b_2 \cdot \sum_{i=1}^{n} x_i^{\,4} = \sum_{i=1}^{n} x_i^{\,2} \cdot y_i \end{cases}$$

$$D = \begin{vmatrix} N & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{vmatrix}, \qquad D_1 = \begin{vmatrix} \sum y & \sum x & \sum x^2 \\ \sum xy & \sum x^2 & \sum x^3 \\ \sum x^2 y & \sum x^3 & \sum x^4 \end{vmatrix}$$

$$D_2 = \begin{vmatrix} N & \sum y & \sum x^2 \\ \sum x & \sum xy & \sum x^3 \\ \sum x^2 & \sum x^2 y & \sum x^4 \end{vmatrix}, \qquad D_3 = \begin{vmatrix} N & \sum x & \sum y \\ \sum x & \sum x^2 & \sum xy \\ \sum x^2 & \sum x^3 & \sum x^2 y \end{vmatrix}$$

$$D_1 = \sum y \sum x^2 \sum x^4 - \sum y \left(\sum x^3\right)^2 - \sum x \sum xy \sum x^4 + \sum x \sum x^2 y \sum x^3 + \sum x^2 \sum xy \sum x^3 - \left(\sum x^2\right)^2 \sum x^2 y;$$

$$D_2 = N \sum xy \sum x^4 - N \sum x^2 y \sum x^3 - \sum x \sum y \sum x^4 + \sum y \sum x^2 \sum x^3 + \sum x \sum x^2 \sum x^2 y - \left(\sum x^2\right)^2 \sum xy;$$

$$D_3 = N \sum x^2 \sum x^2 y - N \sum x^3 \sum xy - \left(\sum x\right)^2 \sum x^2 y + \sum x \sum x^2 \sum x\ y + \sum x \sum y \sum x^3 - \left(\sum x^2\right)^2 \sum y;$$

$$D = N \sum x^2 \sum x^4 - N \left(\sum x^3\right)^2 - \left(\sum x\right)^2 \sum x^4 - \left(\sum x^2\right)^3 + 2 \sum x \sum x^2 \sum x^3 .$$

$$b_0 = \frac{D_1}{D}; \qquad b_1 = \frac{D_2}{D}; \qquad b_2 = \frac{D_3}{D}$$

Для функции y = f(x), имеющей вид:

$$y = b_0 + b_1/X$$

система уравнений для определения коэффициентов уравнения регрессии

имеет вид:

$$\begin{cases} nb_0 + b_1 \cdot \sum \frac{1}{x_i} = \sum y_i \\ b_0 \sum \frac{1}{x_i} + b_1 \sum \frac{1}{x^2} = \sum \frac{x_i}{y_i} \end{cases}$$

Для функции y = f(x), имеющей вид:

$$y=b_0 \cdot b_1^x$$

система уравнений для определения коэффициентов уравнения имеет вид:

$$\begin{cases} \sum \lg y = n \lg b_0 + \lg b_1 \sum x \\ \sum x \lg y = \lg b_0 \sum x + \lg b_1 \sum x^2 \end{cases}.$$

2 ПРИМЕР РАСЧЕТА КОЭФФИЦИЕНТОВ УРАВНЕНИЯ РЕГРЕССИИ

Имеются статистические данные о зависимости рентабельности производства продукции (%) по ряду предприятий, производящих одноименную продукцию, от выработки (в стоимостных показателях) на одного среднесписочного работника производственно-промышленного персонала. Полученные данные представлены в табл. 1:

Таблица 1 – Статистические данные по предприятиям

Номер	X	Y
1	907	11,20
2	926	11,05
3	506	6,84
4	741	9,21
5	789	9,42
6	889	10,08
7	874	9,45
8	510	6,73
9	529	7,24
10	420	6,12
11	679	7,63
12	872	9,43
13	924	9,46
14	607	7,64
15	452	6,92
16	729	8,95
17	794	9,33
18	844	10,23
19	1010	11,77
Итого	14623	176,11

Для прогноза результирующего признака Y применим модель парной регрессии, в которой используется только одна факторная переменная — *X*. Анализ табличных данных показывает наличие прямой линейной зависимости между факторным X (выработки продукции) и результативным

признаком Y (рентабельностью производства). Тесноту и направление связи между факторным и результативным признаками определим с помощью коэффициентом корреляции r.

$$r = \frac{n \cdot \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n \cdot \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

где X_i и Y_i - значения факторного и результативного признаков соответственно; n – объем выборки (число пар исходных данных).

Для рассматриваемого примера значение коэффициента корреляции составляет:

$$r = \frac{20 \times 134127{,}9 - 14623 \times 176{,}11}{\sqrt{20 \times 11306109 - 14623^2} \times \sqrt{20 \times 1602{,}0971 - 176{,}11^2}} = 0{,}955$$

Значение коэффициента корреляции показывает на наличие довольно сильной связи.

Рассчитаем параметры уравнения регрессии: $Y_{теор} = b_0 + b_1 \cdot X$.

Подставив данные из табл. 1 в расчетные формулы получим значения коэффициентов: $b_0 = 2{,}423$; $b_1 = 0{,}00873$.

Таким образом, линейное уравнение регрессии имеет вид:

$$Y_{теор} = 2{,}423 + 0{,}00873 \cdot X .$$

Коэффициент b_1 характеризует наклон линии регрессии. Значение b_1 = 0,00873 и это означает, что при увеличении *X* на единицу ожидаемое значение Y возрастет на 0,00873. Отсюда b_1 может быть интерпретирован как прирост нормы рентабельности, который варьируется в зависимости от средней выручки. Свободный член уравнения $b_0 = 2{,}423$ у.е.. Это значение Y при *X*, равном нулю. Поскольку маловероятно значение выработки, равное нулю, то можно интерпретировать b_0 как меру влияния на величину рентабельности других факторов, не включенных в данное уравнение регрессии.

Регрессионная модель может быть использована для прогноза уровня рентабельности, (который будет на предприятии, например, где средняя выработка на одного работника составит 600 руб.)

Для того, чтобы определить прогнозируемое значение, следует X = 600 подставить в уравнение:

$$Y = 2{,}423 + 0{,}00873 \bullet 600 = 7{,}661.$$

Расчеты показали, что прогнозируемый уровень рентабельности для предприятия со средней выработкой 600 рублей на одного рабочего ППП составит 7,661 %.

ДГТУ
Кафедра "Управление качеством"

Методические указания для выполнения практической работы по дисциплине

«Методы анализа данных»

Ростов-на-Дону, 2021

Методические указания содержат варианты заданий для выполнения практических работ по теме ПФЭ

Задачи. Выработать: умения использования понятийного аппарата теории планирования эксперимента; проведения необходимых этапов подготовки эксперимента, обработки экспериментальных данных.

При выполнении заданий студенту необходимо выбрать свой вариант исходных данных, который соответствует порядковому номер студента в списке группы.

Задание № 1

Составление плана ПФЭ типа 2^n

Задача 1. При оптимизации работы молотильного аппарата при уборке зерновых культур с целью снижения потерь зерна в качестве управляемых факторов выбраны: частота вращения молотильного барабана – Z_1, мин$^{-1}$; зазор на входе – Z_2, мм; зазор на выходе – Z_3, мм.

Пределы изменения факторов следующие:

$Z_{1min} = 800$; $Z_{1max} = 1200$; $Z_{2min} = 16$; $Z_{2max} = 24$; $Z_{3min} = 6$; $Z_{3max} = 10$.

Задание.

- Произвести кодирование факторов. Составить план эксперимента. Графически изобразить область исследования при безразмерном выражении факторов.
- Произвести кодирование факторов. Составить план эксперимента с учетом парных взаимодействий.

Задача 2. При исследовании энергоемкости процесса измельчения листостебельной массы в дробилке агрегата травяной муки выявлены четыре наиболее значимых фактора: Z_1 – зазор между концами молотков и стенкой камеры дробилки, мм; Z_2 – окружная скорость молотков, м/с; Z_3 – количество молотков на роторе; Z_4 – подача материала в дробилку, кг/ч. Центром эксперимента могут быть значения $Z_{10} = 25$; $Z_{20} = 100$; $Z_{30} = 32$; $Z_{40} = 550$. Интервал варьирования факторов принять 20 – 25% от нулевого уровня факторов.

Задание.

Произвести кодирование факторов. Составить план эксперимента с учетом межфакторных взаимодействий.

Задание № 2

Рандомизация опытов

Рандомизировать опыты:

2.1. для ПФЭ типа 2^3 с тремя повторными опытами;

2.2 для ПФЭ типа 2^2 с пятью повторными опытами;

2.3 для ПФЭ типа 2^4 с тремя повторными опытами;

2.4 для ПФЭ типа 2^3 с пятью повторными опытами;

2.5 для ПФЭ типа 2^4 с двумя повторными опытами.

Задание № 3

Проверка воспроизводимости эксперимента

Задача 1. Проверить воспроизводимость эксперимента, данные которого приведены в табл. 1.

Таблица 1 – Экспериментальные данные к задаче 1

u	y_{u1}	y_{u1}	y_{u1}	y_{u1}	$\bar{y}_u$	S_u^2
1	2,4	2,5	2,4	2,6		
2	3,3	3,3	3,3	3,5		
3	2,9	2,8	2,6	2,8		
4	1,8	2,0	1,9	1,8		

Таблица 2 – Экспериментальные данные к задаче 1

u	y_{u1}	y_{u1}	y_{u1}	y_{u1}	$\bar{y}_u$	S_u^2
1	4,4	4,5	4,4	4,6		
2	5,3	5,3	5,3	5,5		
3	4,9	4,8	4,6	4,8		
4	3,8	4,0	3,9	1,8		

Таблица 3 – Экспериментальные данные к задаче 1

u	y_{u1}	y_{u1}	y_{u1}	y_{u1}	$\bar{y}_u$	S_u^2
1	7,1	7,2	7,8	7,3		
2	8,5	8,5	7,3	8,2		
3	7,5	7,9	7,6	7,7		
4	6,7	7,2	7,6	4,9		

Задание № 4

Обработка результатов эксперимента

Задача 1. Проверить однородность дисперсий замеров массовой концентрации аммиака (мг/м^3) в животноводческом помещении с целью выявления систематических погрешностей универсального газоанализатора УГ-2.

Эксперимент проводился на трех уровнях.

Таблица 4 – Экспериментальные данные к задаче 1

1-й уровень	7,5	9,0	8,0	8,0	9,5
2-й уровень	28,5	28,0	25,0	30,0	27,0
3-й уровень	54,0	60,0	55,5	58,0	60,5

Таблица 5 – Экспериментальные данные к задаче 1

1-й уровень	9,5	11,0	10,0	10,0	11,5
2-й уровень	25,5	25,0	22,0	27,0	24,0
3-й уровень	57,0	63,0	58,5	61,0	63,5

Таблица 6 – Экспериментальные данные к задаче 1

1-й уровень	11,2	12,6	12,4	12,8	11,2
2-й уровень	30,5	32,0	28,0	33,0	31,0
3-й уровень	60,0	57,0	58,5	59,0	62,5

Задача 2. Построить эмпирическую зависимость степени измельчения соломы от скорости молотков (измельчитель ИРТ-80) и влажности соломы. Экспериментальные данные приведены в табл. 7. Скорость молотков меняется от 40 до 70 м/с, влажность соломы от 8,9 до 36,6%. Степень измельчения соломы характеризуется процентным содержанием частиц длиной до 50 мм.

Таблица 7 - Экспериментальные данные к задаче 2

u	x_{1u}	x_{2u}	$x_{1u}\,x_{2u}$	Y_u1	Y_u2	Y_u3	$\bar{Y}_u$	S_u^2	$\hat{Y}$
1	-	-	+	50	54	48			
2	+	-	-	64	63	65			
3	-	+	-	22	18	21			
4	+	+	+	86	85	86			

Задание:

- проверить гипотезы о значимости коэффициентов уравнения и адекватности модели.

Задача 3. Работоспособность ременно-планчатых транспортеров валковых жаток (ЖВН-6; ЖРБ-4,2) определяется тяговой способностью, долговечностью и ползучестью (вытяжкой) отдельных лент транспортера. Анализ исследований по тяговой способности плоских транспортных лент и приводных ремней показал, что основными факторами, определяющими тяговую способность лент, являются:

• конструктивное оформление ленты: "+ 1" — лента с вулканизированными прокладками;

"- 1" — серийная лента;

• жесткостные свойства ленты:

"+ Г — лента предварительно обкатана;

"- 1" — лента в состоянии поставки (не обкатана);

• способ сшивки лент:

"+ Г — сшивка "гребешком" с прорезиненной прокладкой; "- 1" — сшивка с помощью металлических накладок;

• удельное начальное натяжение:

"+ 1" - 1,2 МПа;

"- 1" - 0,8 МПа.

В качестве функции отклика было выбрано значение критического крутящего момента Мк. Матрица планирования четырехфакторного эксперимента и результаты опытов приведены в табл. 8.

Таблица 8 - Исходные данные к данные к задаче 3

№ опыта	x_1	x_2	x_3	x_4	y_{u1}	y_{u2}	$\bar{y}_u$	S_u^2	$\hat{y}$
1	+	+	+	+	2,58	2,62			
2	-	+	+	+	2,53	2,57			
з	+	-	+	+	2,24	2,26			
4	-	-	+	+	2,14	2,16			
5	+	+	-	+	2,50	2,50			
6	—	+	-	+	2,16	2,14			
7	+	-	-	+	1,96	1,94			
К	—	-	-	+	2,02	2,08			
9	+	+	+	-	4,84	4,86			
ю	—	+	+	-	4,77	4,73			
II	+	—	+	-	3,80	3,70			
12	_	—	+	-	3,90	3,50			
13	+	+	—	-	4,00	3,60			
14		+	—	—	3,54	3,56			
15	+	_	_	—	2,70	2,50			
U»	-	-	-	-	2,96	3,05			

ВАРИАНТЫ ИНДИВИДУАЛЬНЫХ ЗАДАНИЙ ПРИ ВЫПОЛНЕНИИ КОНТРОЛЬНОЙ РАБОТЫ

Номер варианта заданий соответствует номеру в списке группы

Номер варианта	Задание 1	Задание 2	Задание 3	Задание 4
1	Задача 1	2.1	Табл. 1	Табл. 4. Задача 2. Задача 3.
2	Задача 2	2.2	Табл. 2	Табл. 5. Задача 2. Задача 3.
3	Задача 1	2.3	Табл. 3	Табл. 6. Задача 2. Задача 3.
4	Задача 2	2.4	Табл. 1	Табл. 4. Задача 2. Задача 3.
5	Задача 1	2.5	Табл. 2	Табл. 5. Задача 2. Задача 3.
6	Задача 2	2.1	Табл. 3	Табл. 6. Задача 2. Задача 3.
7	Задача 1	2.2	Табл. 1	Табл. 4. Задача 2. Задача 3.
8	Задача 2	2.3	Табл. 2	Табл. 5. Задача 2. Задача 3.
9	Задача 1	2.4	Табл. 3	Табл. 6. Задача 2. Задача 3.
10	Задача 2	2.5	Табл. 1	Табл. 4. Задача 2. Задача 3.
11	Задача 1	2.1	Табл. 2	Табл. 5. Задача 2. Задача 3.
12	Задача 2	2.2	Табл. 3	Табл. 6. Задача 2. Задача 3.
13	Задача 1	2.3	Табл. 1	Табл. 4. Задача 2. Задача 3.
14	Задача 2	2.4	Табл. 2	Табл. 5. Задача 2. Задача 3.
15	Задача 1	2.5	Табл. 3	Табл. 6. Задача 2. Задача 3.
16	Задача 2	2.1	Табл. 1	Табл. 4. Задача 2. Задача 3.
17	Задача 1	2.2	Табл. 2	Табл. 5. Задача 2. Задача 3.
18	Задача 2	2.3	Табл. 3	Табл. 6. Задача 2. Задача 3.
19	Задача 1	2.4	Табл. 1	Табл. 4. Задача 2. Задача 3.
20	Задача 2	2.5	Табл. 2	Табл. 5. Задача 2. Задача 3.
21	Задача 1	2.1	Табл. 3	Табл. 6. Задача 2. Задача 3.
22	Задача 2	2.2	Табл. 1	Табл. 4. Задача 2. Задача 3.
23	Задача 1	2.3	Табл. 2	Табл. 5. Задача 2. Задача 3.
24	Задача 2	2.4	Табл. 3	Табл. 6. Задача 2. Задача 3.
25	Задача 1	2.5	Табл. 1	Табл. 4. Задача 2. Задача 3.
26	Задача 2	2.1	Табл. 2	Табл. 5. Задача 2. Задача 3.
27	Задача 1	2.2	Табл. 3	Табл. 6. Задача 2. Задача 3.
28	Задача 2	2.3	Табл. 1	Табл. 4. Задача 2. Задача 3.
29	Задача 1	2.4	Табл. 2	Табл. 5. Задача 2. Задача 3.
30	Задача 2	2.5	Табл. 3	Табл. 6. Задача 2. Задача 3.