Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

Санкт-Петербургский горный университет

Кафедра высшей математики

Расчетно-графическая работа

По дисциплине _____Математика_____
(наименование учебной дисциплине согласно учебному плану)

Тема работы: __Основы математической статистики___

Вариант 106

Выполнил: ст. группы КРС-21 _____ Крюков Т.В.
(должность) (подпись) (Ф.И.О)

Дата: _____

Проверил: доцент _____ Мансурова С.Е.
(должность) (подпись) (Ф.И.О)

Санкт-Петербург
2022

Задача 1.

Статистическая обработка массива данных.

Выборочные оценки генеральной совокупности.

Условие задачи:

1) Составить сгруппированный статистический ряд:

а) определить нужное кол-во интервалов по формуле Стерджесса

$k \approx 1 + \log_2 N$, где N-объем выборки;

б) найти наибольшее и наименьшее значение измеряемой величины;

в) определить ширину интервала. Используемая формула:

$$\Delta x = \left[\frac{x_{max} - x_{min}}{k} \right] + 1, \; где [\,] - округление\; вверх$$

г) интервалы берутся открытыми слева и закрытыми справа: $(\alpha_i;\beta_i]$, i=1..k

д) концы интервалов вычисляются по формулам:

$$\alpha_i = x_{min} - \frac{\Delta x}{2}; \alpha_{i+1} = \alpha_i + \Delta x, \beta_i = \alpha_i + \Delta x$$

е) определить середины интервалов: $x_i = \frac{\alpha_i + \beta_i}{2}$

2) Для сгруппированного статистического ряда вычислить выборочное среднее, выборочную дисперсию, выборочное среднее квадратическое отклонение, исправленную дисперсию, исправленное среднее квадратическое отклонение.

3) Считая генеральную совокупность нормально распределенной, найти интервальные оценки математического ожидания и среднего квадратического отклонения с надежностью 0,95

4) Построить гистограмму, полигон частот и эмпирическую функцию распределения.

Исходные данные:

A
Вар.
106
39
2
17
29
17
18
2
6
17
27
1
48
4
26
18
8
42
25
36
47

	A
28	13
29	15
30	25
31	-6
32	27
33	42
34	21
35	27
36	22
37	18
38	-5
39	24
40	16
41	5
42	22
43	15
44	34
45	5
46	13
47	33
48	29
49	22
50	11
51	29
52	6
53	19
54	37
55	9

	A
55	9
56	29
57	16
58	25
59	32
60	28
61	18
62	22
63	12
64	-2
65	3
66	17
67	39
68	16
69	-2
70	2
71	13
72	41
73	21
74	22
75	29
76	16
77	44
78	23
79	37
80	26
81	24

	A
79	37
80	26
81	24
82	31
83	27
84	43
85	-5
86	41
87	31
88	15
89	6
90	3
91	20
92	4
93	4
94	42
95	16
96	13
97	20
98	27
99	25
100	1
101	5
102	25
103	21
104	
105	
106	

Решение:

1) Объем выборки N=100. Определим кол-во интервалов для группировки по формуле Стеджесса:

k¿¿ , где […] – символ округления (вверх) до целого.

Найдем наибольшее и наименьшее значения измеряемой величины

Xmin	Xmax
-6	48

Определим ширину интервала.

$$\Delta x = \left[\frac{x_{max} - x_{min}}{k}\right] + 1 = \left[\frac{48+6}{8}\right] + 1 = [7] + 1 = 8$$

Составим таблицу со сгруппированным статистическим рядом:

Номер интервала	Интервал	Середина интервала	Частота	Относительная частота
i	$(\alpha_i;\beta_i]$	x_i	m_i^*	p_i^*
1	(-10; -2]	-6	5	0,05
2	(-2;6]	2	17	0,17
3	(6; 14]	10	8	0,08
4	(14; 22]	18	27	0,27
5	(22; 30]	26	22	0,22
6	(30; 38]	34	9	0,09
7	(38; 46]	42	10	0,1
8	(46; 54]	50	2	0,02
Контрольные суммы			100	1

Рис.1. Таблица со сгруппированным статистическим рядом

Начало первого интервала вычисляется по формуле: $\alpha_i = x_{min} - \frac{\Delta x}{2} = -6 - \frac{8}{2} = -10$;

для остальных : $\alpha_{i+1} = \alpha_i + \Delta x$, $\beta_i = \alpha_i + \Delta x$

2)Числовые характеристики

(точечные оценки параметров генеральной совокупности)

Выборочное среднее:

$$\bar{x}^* = \frac{1}{N}\sum_{i=1}^{4} x_i m_i^* \quad \text{или} \quad \bar{x}^* = \sum_{i=1}^{4} x_i p_i^* .$$

Выборочная дисперсия:

$$D^*(x) = \frac{1}{N}\sum_{i=1}^{4} x_i^2 m_i^* - (\bar{x}^*)^2 \quad \text{или} \quad D^*(x) = \sum_{i=1}^{4} x_i^2 p_i^* - (\bar{x}^*)^2 .$$

Выборочное среднее квадратическое отклонение:

$$\sigma_x^* = \sqrt{D^*(x)} .$$

Несмещенная дисперсия:

$$s_x^{\,2} = \frac{n}{n-1} D^*(x)$$

Исправленное среднее квадратическое отклонение

$$s_x = \sqrt{s_x^2}$$

:

Полученные результаты из Excel:

Выборочное среднее	Выборочная дисперсия	Выборочное среднее квадратичное отклонени	Несмещенная дисперсия	Исправление среднее квадратичного отклонен
x̄*	D*(x)	δx*	S_x^2	S_x
19,68	189,8	13,78	191,7	13,85

3) Интервальные оценки параметров генеральной совокупности

Теория:1. Оценка мат. ожидания при неизвестной дисперсии

Интервальной **называют оценку, которая определяется двумя числами — концами интервала, покрывающего оцениваемый параметр.**

Доверительным **называют интервал, который с заданной надежностью γ покрывает заданный параметр.**

1. ***Интервальной оценкой (с надежностью γ) математического ожидания a нормально распределенного количественного признака X*** **по выборочной средней $\bar{x}_в$ при известном среднем квадратическом отклонении σ генеральной совокупности служит доверительный интервал**

$$\bar{x}_в - t(\sigma/\sqrt{n}) < a < \bar{x}_в + t(\sigma/\sqrt{n}),$$

где $t(\sigma/\sqrt{n}) = \delta$ — точность оценки, n — объем выборки, t — значение аргумента функции Лапласа $\Phi(t)$ (см. приложение 2), при котором $\Phi(t) = \gamma/2$; при неизвестном σ (и объеме выборки $n < 30$)

$$\bar{x}_в - t_\gamma(s/\sqrt{n}) < a < \bar{x}_в + t_\gamma(s/\sqrt{n}),$$

где s — «исправленное» выборочное среднее квадратическое отклонение, t_γ находят по таблице приложения 3 по заданным n и γ.

Для решаемой задачи: объем выборки N=100; надежность γ=0,95; исправленное квадратическое отклонение s=13,85. По таблице значений для t_γ находим: t_γ=1,984 (таблица значений ниже)

Тогда точность оценки (радиус доверительного интервала):

$$= \varepsilon_a = \frac{t_\gamma \cdot s_x}{\sqrt{N}} = \frac{1{,}984 * 13{,}85}{\sqrt{100}} = 2{,}75$$

Таким образом, с вероятностью 0,95 математическое ожидание генеральной совокупности принадлежит интервалу (19,68-2,75; 19,68+2,75), т.е. 16,866<M(ε)< 22,494

Таблица значений $t_\gamma = t(\gamma, n)$

n \ γ	0,95	0,99	0,999	n \ γ	0,95	0,99	0,099
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,720	3,600
9	2,31	3,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,991	2,640	3,418
15	2,15	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,02	120	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Теория: 2. Оценка ср.кв. отклонения.

2. *Интервальной оценкой (с надежностью γ) среднего квадратического отклонения σ нормально распределенного количественного признака X* по «исправленному» выборочному среднему квадратическому отклонению s служит доверительный интервал

$$s(1-q) < \sigma < s(1+q) \text{ (при } q < 1),$$
$$0 < \sigma < s(1+q) \text{ (при } q > 1),$$

где q находят по таблице приложения 4 по заданным n и γ.

Таблица значений $q = q(\gamma, n)$

n \ γ	0,95	0,99	0,999	n \ γ	0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

По таблице значений для $q(\gamma,N)$ находим: $q(0{,}95;100)=0{,}143$. Тогда концы доверительного интервала:

$$\alpha_\sigma = s(1-q) = 13{,}85(1-0{,}143) = 11{,}867;$$

$$\beta_\sigma = s(1+q) = 13{,}85(1+0{,}143) = 15{,}827.$$

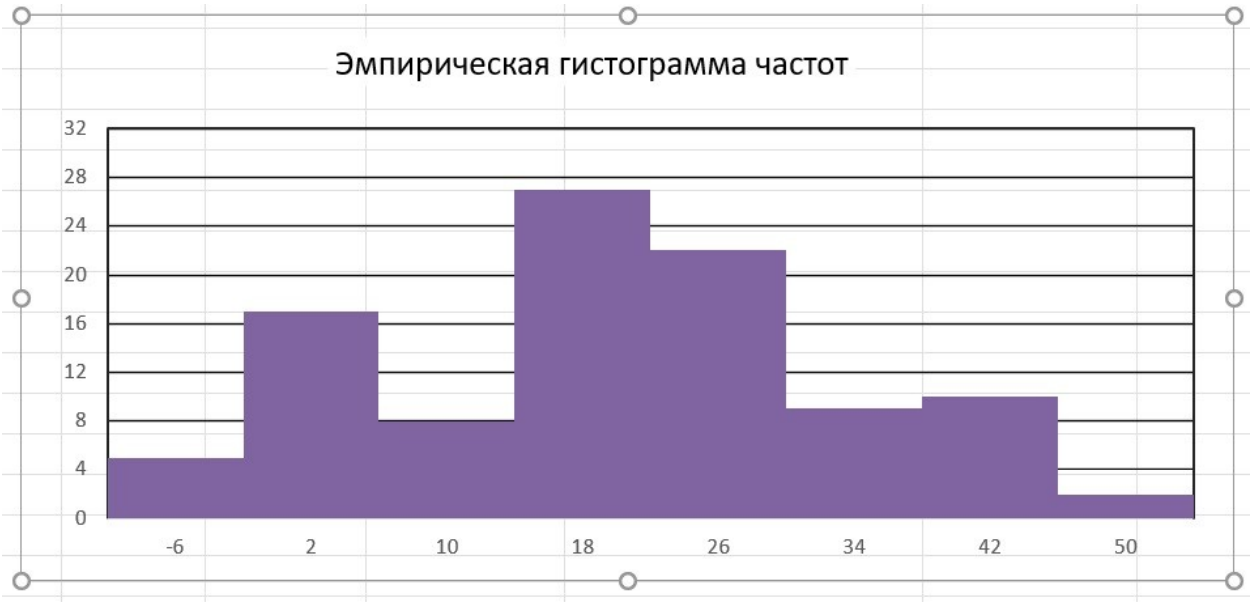
Таким образом, с вероятностью 0,95 среднее квадратическое отклонение генеральной совокупности принадлежит интервалу $(11{,}867;15{,}827)$, *т.е.*

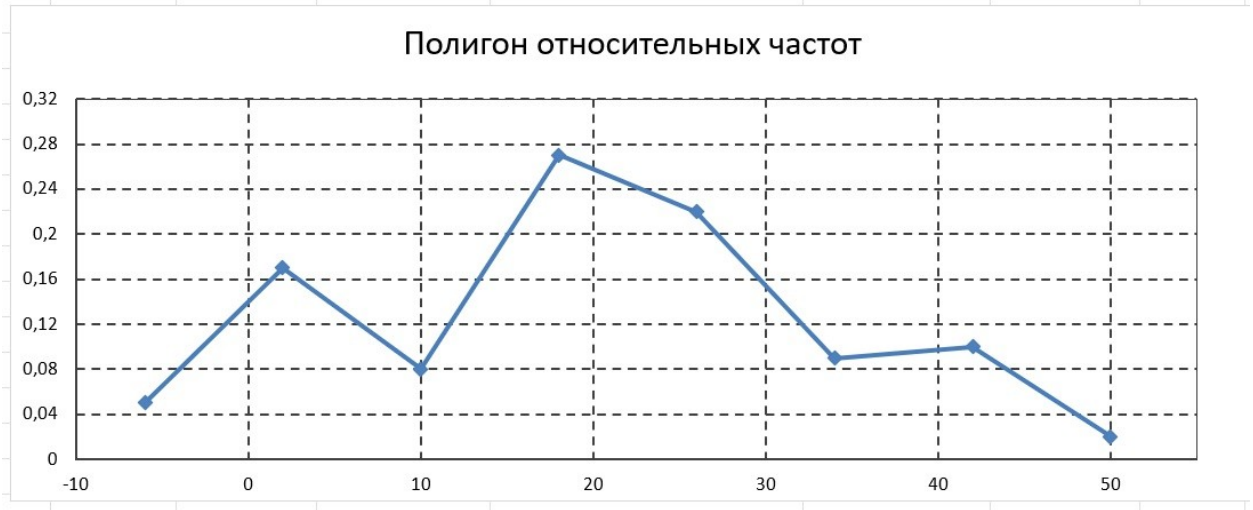
$$11{,}867 < \sigma(\varepsilon) < 15{,}827$$

Результаты из Excel:

εa	aσ	βσ
2,75	11,867	15,827

4)Гистограмма, полигон частот и эмпирическая функция распределения.





Для построения эмпирической функции распределения, с учетом группировки, лучше использовать не кусочно-постоянную, а непрерывную

кусочно-линейную функцию, где до начала первого интервала и на левом конце первого интервала $F^* = 0$, на правом конце первого интервала $F^{\prime}=p_i$ на отрезке $[\alpha_1;\beta_1]$ строим функцию как отрезок, соединяющий конечные точки. На втором отрезке будет $F^{\prime}(\alpha_2)=F^{\prime}(\beta_1)=p_1;$

$F^{\prime}(\beta_2)=p_1+p_2$и т.д. Таким образом, для последнего интервала получим: $F^{\prime}(\beta_k)=1$ и это же значение $F^{\prime}$будет иметь для всех значений x, больших, чем 54. При подобном рассмотрении эмпирической функции распределения вероятностей мы предполагаем, что изменения ее значений происходят не скачком в средних точках интервалов, а накапливаются "плавно" на каждом интервале.

Составим таблицу для построения описанной функции распределения:

	"Запас"	α_1	$\alpha_2=\beta_1$	$\alpha_3=\beta_2$	$\alpha_4=\beta_3$	$\alpha_5=\beta_4$	$\alpha_6=\beta_5$	$\alpha_7=\beta_6$	$\alpha_8=\beta_7$	β_8	"Запас"
x	-20	-10	-2	6	14	22	30	38	46	54	70
F*(x)	0	0	0,05	0,22	0,3	0,57	0,79	0,88	0,98	1	1

Задача 2. Проверка статистических гипотез.

Условие задачи:

По данным из задачи 1:

1) Оценить с помощью критерия Пирсона хи-квадрат согласие данных с нормальным распределением при уровне значимости α=0,05

2) Построить сравнительный график эмпирических (ломаная линия с маркерами) и теоретических (сглаженная линия без маркеров) частот. График рисуется на белом фоне линиями разных цветов, добавляются линии сетки координат и легенда.

Решение:

Оценим с помощью критерия Пирсона хи-квадрат согласие эмпирических данных с нормальным распределением при уровне значимости α=0,05 .

Основная гипотеза H_0: рассматриваемая генеральная совокупность ξ распределена по нормальному закону с параметрами $a=\overline{x}=19,68$ и δ=δ*=13,78 ; альтернативная гипотеза H_1: генеральная совокупность распределена по какому-то другому закону.

Для этого дополним таблицу со сгруппированным статистическим рядом значениями теоретических вероятностей теоретических частот.

Теоретические вероятности попадания значения генеральной совокупности в интервал вычисляются как $p_i=\varphi(x_i)\cdot\Delta x$, где $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ плотность нормального распределения, $i=1..k$. Теоретические частоты вычисляются по формуле $m_i=[p_i\cdot N]$, где [...] — округление до целого.

При этом контрольные суммы для теоретических вероятностей и частот должны ***приближенно*** равняться 1 и объему выборки.

Вычислив необходимые величины, получим:

Номер интервала	Середина интервала	Эмпирические значения:		Теоритические значения		
		Относительная частота	Частота	Плотность нормального	Вероятность попадания в	Частоты
i	xi	pi*	mi*	φ(xi)	pi	mi
1	-6	0,05	5	0,00510	0,0408	4
2	2	0,17	17	0,01271	0,1017	10
3	10	0,08	8	0,02262	0,1810	18
4	18	0,27	27	0,02874	0,2299	23
5	26	0,22	22	0,02606	0,2085	21
6	34	0,09	9	0,01687	0,1350	13
7	42	0,1	10	0,00780	0,0624	6
8	50	0,02	2	0,00257	0,0206	2
Σ=	100	1	100		0,9798	97

Т.к. критерий Пирсона не "работает" при малых частотах, объединим 1-й интервал со 2-ми 7й с 8-м. Таким образом, количество интервалов уменьшится до $k=6$.

Добавим в таблицу столбец с наблюдаемыми значениями распределения χ^2, где

$\chi_i^2=\frac{(m_i^2-m_i)^2}{m_i}$, $\chi^2_{набл}=\sum_{i=1}^{k}\chi_i^2$ Получим:

Номер интервала	Интервал	Эмпирическая частота	Теоритическая частота	Наблюдаемые значения χ^2	χ_i^2
i	$(\alpha_i;\beta_i]$	m_i^*	m_i	χ_i^2	7,81473
1	(-10; 6]	22	14	4,571428571	
2	(6; 14]	8	18	5,555555556	
3	(14; 22]	27	23	0,695652174	
4	(22; 30]	22	21	0,0476	
5	(30; 38]	9	13	1,230769231	
6	(38; 54]	12	8	2	
	Σ=	100	97	14,101	

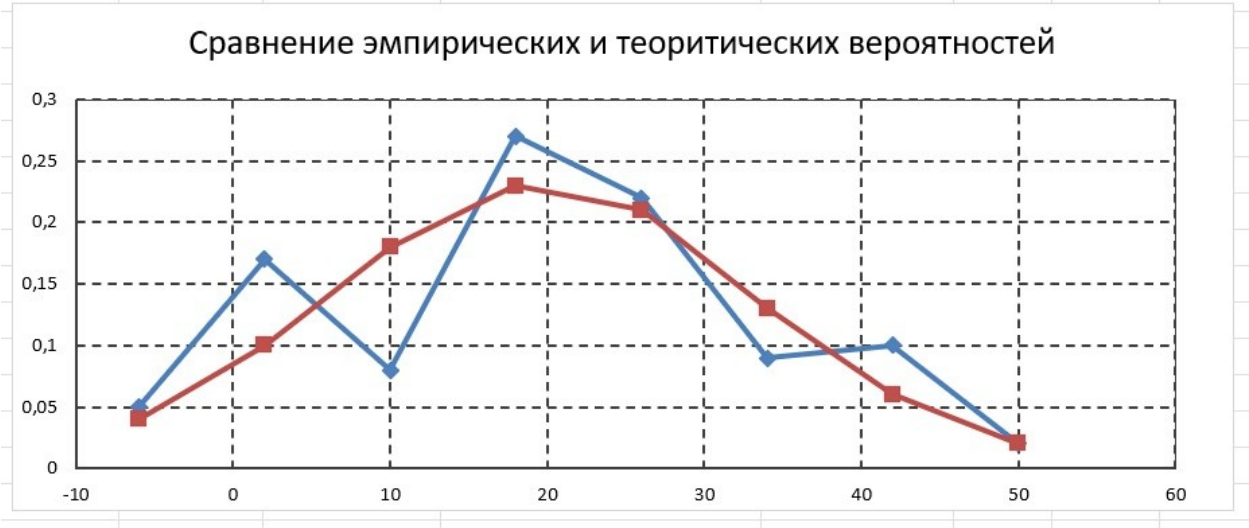
Таким образом, $\chi^2_{набл}=14,101$. По таблице критических точек распределения χ^2 найдем $\chi_{критич}$ при числе степеней свободы $n = k - 3 = 6 - 3 = 3$ и уровне значимости $\alpha = 0,05$:

$$\chi^2_{критич}=7,815$$

Сравнивая найденное значение $\chi^2_{набл}=14,101$ с критическим (14,101> 7,815), определяем, что $\chi^2_{набл}$ лежит в критической области

Таким образом, при уровне значимости α=0,05 можно утверждать, что распределение рассматриваемой генеральной совокупности ξ **сильно отличается от нормальной**

Построим сравнительный график эмпирических (ломаная линия с маркерами) и теоретических (сглаженная линия без маркеров) частот (используя точечную диаграмму).



Задача 3. Корреляционный и регрессионный анализ данных.

Условие задачи:

1)Установить наличие или отсутствие связи между случайными величинами X и У, вычислив выборочный коэффициент корреляции.

2)Найти выборочные регрессии У на Х и Х на У, предполагая, что они линейные

3)Построить линии регрессий и точки условных средних на одном чертеже (точки условных средних Х и регрессия Х и У изображаются одним цветом, а точки условных средних У и регрессия У на Х – другим цветом).

Исходные данные:

Вар. 106		X						
		11	**19**	**27**	**35**	**43**	**51**	**59**
Y	**17**	109	79	63	47	31	3	1
	27	78	108	80	64	48	32	16
	37	62	81	107	82	65	49	33
	47	46	66	82	114	79	66	65
	57	16	51	66	71	111	115	82
	67	2	17	50	54	84	84	131

Решение:

Вар. 106		X							mj	$\sum_{i=1}^{7} x_i n_{ji}$	$\bar{x}_{y_j}=\frac{1}{m_j}\sum_{i=1}^{7} x_i n_{ji}$
		11	**19**	**27**	**35**	**43**	**51**	**59**			
Y	**17**	109	79	63	47	31	3	1	333	7591	22,80
	27	78	108	80	64	48	32	16	426	11950	28,05
	37	62	81	107	82	65	49	33	479	15221	31,78
	47	46	66	82	114	79	66	65	518	18562	35,83
	57	16	51	66	71	111	115	82	512	20888	40,80
	67	2	17	50	54	84	84	131	422	19210	45,52
ni		313	402	448	432	418	349	328	2690	93422	34,73
$\sum_{j=1}^{6} y_j n_{ji}$		9461	14404	18156	18584	19896	18013	18176	116690		
$\bar{y}_{x_i}=\frac{1}{n_i}\sum_{j=1}^{6} y_j n_{ji}$		30,23	35,83	40,53	43,02	47,60	51,61	55,41	43,38		

Математические характеристики Х:

1)Математическое ожидание (выборочное среднее)

$$M(X)=\bar{x}=\frac{1}{N}\sum_{i=1}^{7} x_i n_i = 34,73$$

2)Дисперсия (выборочная дисперсия)

$$D(X)=D(x)=\frac{1}{N}\sum_{i=1}^{7} x_i^2 n_i-(\bar{x})^2=229$$

3)Выборочное среднее квадратическое отклонение

$$\sigma(X)=\sigma_x^{\text{¿}}=\sqrt{D^{\text{¿}}(x)}=15,1$$

4)Условные средние, вычисленные в последнем столбце по формуле

$$\overline{x}_{y_j}=\frac{1}{m_j}\sum_{i=1}^{4}x_i n_{ji}$$

Математические характеристики У:

1)Математическое ожидание (выборочное среднее)

$$M(Y)=\overline{y}^{\text{¿}}=\frac{1}{N}\sum_{i=1}^{6}y_i m_i=43{,}3792$$

2)Дисперсия (выборочная дисперсия)

$$D(Y)=D^{\text{¿}}(y)=\frac{1}{N}\sum_{i=1}^{6}y_i^{\,2}m_i-(\overline{y}^{\text{¿}})^2=261,24$$

3)Выборочное среднее квадратическое отклонение

$$\sigma(Y)=\sigma_y^{\text{¿}}=\sqrt{D^{\text{¿}}(y)}=16{,}163$$

4)Условные средние, вычисленные в последнем столбце по формуле

$$\overline{y}_{x_j}=\frac{1}{n_i}\sum_{i=1}^{6}y_j n_{ji}$$

Математическое ожидание совместного появления (произведения) значений X и У:

$$M(XY)=\frac{1}{N}\sum_{i=1}^{6}\sum_{j=1}^{7}x_i y_j n_{ji}=\frac{1}{N}\sum_{i=1}^{6}\left(x_i\sum_{j=1}^{7}y_j n_{ji}\right)=1622,67$$

Выборочный коэффициент корреляции:

$$r_{xy}^{\text{¿}}=\frac{M(XY)-\overline{x}^{\text{¿}}\overline{y}^{\text{¿}}}{\sigma_x^{\text{¿}}\sigma_y^{\text{¿}}}=0{,}4746$$

Регрессия Y на X — функция, показывающая зависимость условных средних $\overline{y}_{x_i}$ от значений x_i $\quad y=\overline{y}^{\text{¿}}+r_{xy}^{\text{¿}}\frac{\sigma_y^{\text{¿}}}{\sigma_x^{\text{¿}}}(x-\overline{x}^{\text{¿}})\Rightarrow$

$$y=43{,}3792+0{,}4746\cdot\frac{16{,}163}{15{,}1}(x-34{,}73)\Rightarrow y=43{,}3792+0{,}51x-17{,}64\Rightarrow$$

$$y_{рег}=25{,}739+0{,}51x$$

Составим таблицу значений регрессии Y на X (для известных X)

Xi	11	19	27	35	43	51	59
Yрег	31,36	35,41	39,46	43,52	47,57	51,62	55,67

Уравнение линейной регрессии X на У: $x=\bar{x}^{¿}+r_{xy}^{¿}\frac{\sigma_x^{¿}}{\sigma_y^{¿}}(y-\bar{y}^{¿})\Rightarrow$

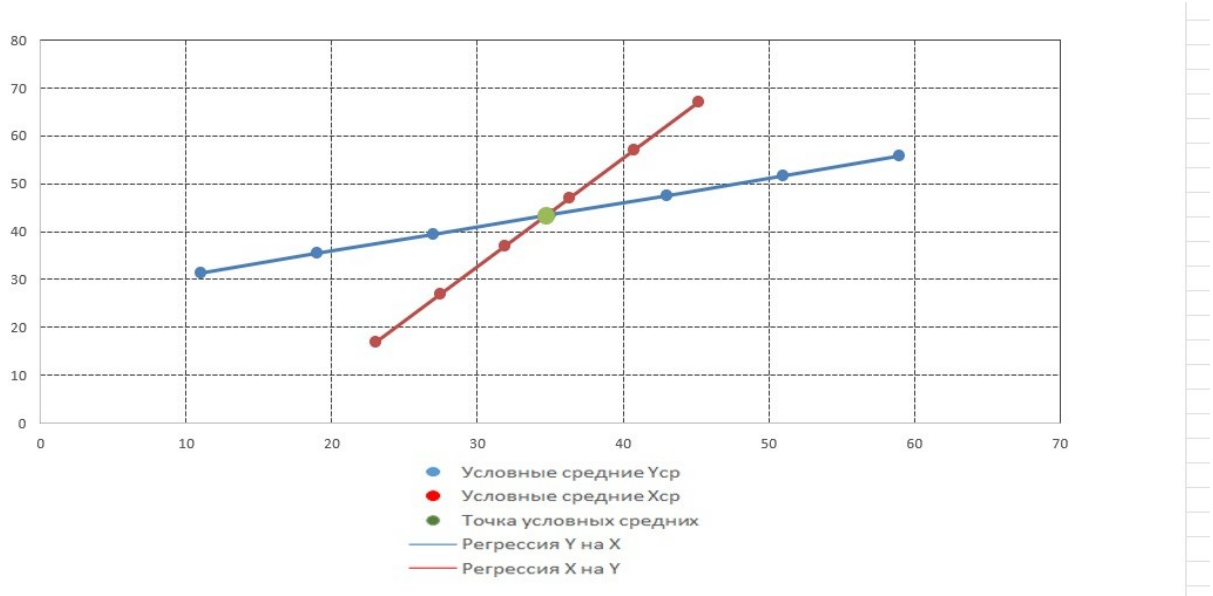
$$x=34,73+0,4746\cdot\frac{15,1}{16,163}(y-43,3792)\Rightarrow x_{рег}=34,73+0,443\,y-19,234\Rightarrow$$

$$x_{рег}=15,496+0,443\,y$$

Составим таблицу значений регрессии X на У (для известных У)

yj	17	27	37	47	57	67
Xрег	23,00	27,45	31,89	36,34	40,78	45,2

Сравнение регрессий:



Анализ корреляции: $r_{xy}^{¿}\approx 0,4746$ — коэфф-т корреляции $\neq 0$(существено больше нуля) ⇒ связь между X и У есть, и существенная (величины X и У коррелированы), при этом значение r_{xy}существенно меньше 1 ⇒ связь нелинейная (существенно нелинейная).