МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра математического обеспечения и применения ЭВМ

КУРСОВАЯ РАБОТА

по дисциплине «Статистические методы обработки экспериментальных данных»

Тема: Программная реализация и компьютерное исследование алгоритмов обработки экспериментальных данных

Студент гр. 8383 _____ Киреев К.А.

Преподаватель _____ Середа А.-В.И.

Санкт-Петербург
2022

МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра математического обеспечения и применения ЭВМ

КУРСОВАЯ РАБОТА

по дисциплине «Статистические методы обработки экспериментальных данных»

Тема: Программная реализация и компьютерное исследование алгоритмов обработки экспериментальных данных

Студент гр. 8383 _____ Муковский Д.В.

Преподаватель _____ Середа А.-В.И.

Санкт-Петербург

2022

ЗАДАНИЕ

НА КУРСОВУЮ РАБОТУ

Студент Киреев К.А.

Группа 8383

Тема работы: Программная реализация и компьютерное исследование алгоритмов обработки экспериментальных данных.

Исходные данные:

Из представленной генеральной совокупности формируется выборка заданного объема. Необходимо провести выравнивание статистических рядов, выполнить корреляционный, регрессионный и кластерный анализы.

Содержание пояснительной записки:

«Аннотация», «Содержание», «Введение», «Заключение», «Список использованных источников».

Предполагаемый объем пояснительной записки:

Не менее 20 страниц.

Дата выдачи задания: 05.04.2022

Дата сдачи реферата: 07.04.2022

Дата защиты реферата: 07.04.2022

Студент _____ Киреев К.А.

Преподаватель _____ Середа А.-В.И.

ЗАДАНИЕ

НА КУРСОВУЮ РАБОТУ

Студент Муковский Д.В.

Группа 8383

Тема работы: Программная реализация и компьютерное исследование алгоритмов обработки экспериментальных данных

Исходные данные:

Из представленной генеральной совокупности формируется выборка заданного объем. Необходимо провести выравнивание статистических рядов, выполнить корреляционный, регрессионный и кластерный анализы.

Содержание пояснительной записки:

«Аннотация», «Содержание», «Введение», «Заключение», «Список использованных источников».

Предполагаемый объем пояснительной записки:

Не менее 20 страниц.

Дата выдачи задания: 05.04.2022

Дата сдачи реферата: 07.04.2022

Дата защиты реферата: 07.04.2022

Студент _____ Муковский Д.В.

Преподаватель _____ Середа А.-В.И.

АННОТАЦИЯ

В данной курсовой работе исследуется двумерная выборка, состоящая из данных наблюдений относительно объемного веса nu ($\frac{\text{г}}{\text{см}^3}$) при влажности 10% и модуля упругости E ($\frac{\text{кг}}{\text{см}^2}$) при сжатии вдоль волокон древесины резонансной ели. Исследование включает в себя выравнивание статистических рядов, нахождение точечных и интервальных статистических оценок, построение регрессионных кривых, проверку статистических гипотез о нормальном распределении выборки и о равенстве коэффициента корреляции нулю. Методы исследования включают в себя корреляционный анализ, регрессионный анализ и кластерный анализ, в частности методы k-means и метод поиска сгущений.

SUMMARY

This course work examines a two-dimensional sample consisting of observational data on bulk density nu $\left(\frac{g}{sm^3}\right)$ at 10% moisture content and modulus of elasticity E ($\frac{kg}{sm^2}$) under compression along the fibers of resonant spruce wood. The study includes the alignment of statistical series, finding point and interval statistical estimates, building regression curves, testing statistical hypotheses about the normal distribution of the sample and about the equality of the correlation coefficient to zero. Research methods include correlation analysis, regression analysis and cluster analysis, in particular, k-means methods and the method of searching for clusters.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В ходе данной работы необходимо ознакомиться с основными правилами формирования выборки и подготовки выборочных данных к статистическому анализу, получить практические навыки нахождения точечных статистических оценок параметров распределения. Получить практические навыки вычисления интервальных статистических оценок параметров распределения выборочных данных и проверки статистических гипотез.

Необходимо освоить основные понятия, связанные с корреляционной зависимостью между случайными величинами, доверительными интервалами, статистическими гипотезами и их проверкой. Ознакомиться с основными положениями метода наименьших квадратов, со статистическими свойствами МНК оценок, с понятием функции регрессии и роли МНК в регрессионном анализе, с корреляционным отношением, как мерой тесноты произвольной корреляционной связи.

Необходимо освоить и реализовать методы кластерного анализа, такие как, метод k-means и метод поиска сгущений.

1. ВЫРАВНИВАНИЕ СТАТИСТИЧЕСКИХ РЯДОВ

1.1. Основные теоретические положения

Ранжированный ряд – это распределение отдельных единиц совокупности в порядке возрастания или убывания исследуемого признака. Ранжирование позволяет легко разделить количественные данные по группам, сразу обнаружить наименьшее и наибольшее значения признака, выделить значения, которые чаще всего повторяются. Вариационный ряд – последовательность значений заданной выборки $x^m = (x_1, \dots, x_m)$, расположенных в порядке неубывания:

$$x^{(1)} \le x^{(2)} \le \dots \le x^{(m)}$$

Интервальный ряд распределения – это таблица, состоящая из двух столбцов (строк) – интервалов варьирующего признака X_i и числа единиц совокупности, попадающих в данный интервал (частот - f_i), или долей этого числа в общей численности совокупностей (частостей - d_i). Полигоном частот называют ломанную, отрезки которой соединяют точки $(x_1, n_1), (x_2, n_2), \dots, (x_k, n_k)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им частоты n_i. Точки (x_i, n_i) соединяют отрезками прямых и получают полигон частот. Гистограммой частот (частостей) называется ступенчатая фигура, состоящая из прямоугольников с основаниями, равными интервалам значений h_i и высотами, равными отношению частот (или частостей) к шагу. Эмпирической функцией распределения, построенной по выборке $x^m = (x_1, \dots, x_m)$ объема m, называется случайная функция $\hat{F}_m(x)$, равная

$$\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} I_{\{x_i<x\}}.$$

Значения эмпирической функции распределения принадлежат отрезку [0,1].

Математическим ожиданием дискретной случайной величины называется сумма произведений ее возможных значений на соответствующие им вероятности:

$$M(X) = \frac{1}{N}\sum_{i=1}^{n} x_i n_i$$

Дисперсией случайной величины называется математическое ожидание квадрата ее отклонения от ее математического ожидания:

$$D(X) = M\big(X - M(X)\big)^2$$

Среднеквадратическим отклонением случайной величины X называется квадратный корень из ее дисперсии:

$$\sigma = \sqrt{D(X)}$$

Выборочная дисперсия определяется по формуле:

$$D_B = \frac{1}{N}\sum_{i=1}^{k} (x_i - \bar{x})^2 \, n_i$$

Исправленная выборочная дисперсия определяется по формуле:

$$s^2 = \frac{N}{N-1} D_B$$

Центральным моментом порядка k случайной величины X называется математическое ожидание величины:

$$M\big(X - M(X)\big)^k = m_k.$$

Асимметрией, или коэффициентом асимметрии, называется числовая характеристика, определяемая выражением:

$$A_s = \frac{m_3}{s^3},$$

где m_3 – центральный эмпирический момент третьего порядка, s – исправленная выборочная дисперсия.

Эксцессом, или коэффициентом эксцесса, называется численная характеристика случайной величины, которая определяется выражением:

$$E = \frac{m_4}{S^4} - 3.$$

Мода дискретной случайной величины – это наиболее вероятное значение этой случайной величины.

$$M_o = x_o + \frac{(m_2 - m_1)}{(m_2 - m_1) + (m_2 - m_3)} h$$

Медиана случайной величины X – это такое ее значение M_e, для которого выполнено равенство

$$M_e = x_o + \frac{0{,}5 * n - n_{m-1}^n}{n_o} h$$

Доверительным называют интервал, который с заданной надежностью γ покрывает заданный параметр. Доверительный интервал для оценки математического ожидания при неизвестном СКО, который покрывает неизвестное значение параметра a с надежностью γ можно построить как:

$$\left(\overline{x_в} - t_\gamma \frac{s}{\sqrt{N}}, \overline{x_в} + t_\gamma \frac{s}{\sqrt{N}}\right)$$

Интервальной оценкой среднеквадратического отклонения σ по исправленной выборочной дисперсии служит доверительный интервал:

$$s(1 - q) \leq \sigma \leq s(1 - q),$$

Критерий Пирсона, или критерий χ^2, применяют для проверки гипотезы о соответствии эмпирического распределения предполагаемому теоретическому распределению. Метод позволяет оценить статистическую значимость различий двух или нескольких относительных показателей.

Теоретические частоты вычисляются по формуле:

$$n_i^{'} = p_i * N,$$

$$p_i = \Phi(z_{i+1}) - \Phi(z_i),$$

где $\Phi(z_i)$ – функция Лапласа

Если $\chi^2_{\text{наб}} \leq \chi^2_{\text{крит}}$ – гипотеза принимается, иначе – гипотеза отвергается.

1.2. Формирование и первичная обработка выборки. Ранжированный и интервальный ряды.

Выборка состоит из данных наблюдений относительно объемного веса nu $(\frac{г}{см^3})$ при влажности 10% и модуля упругости E $(\frac{кг}{см^2})$ при сжатии вдоль волокон древесины резонансной ели.

Формирование репрезентативной выборки заданного объема из имеющейся генеральной совокупности экспериментальных данных представлено в таблице 1. Объём выборки: 104.

Таблица 1

№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***
1	460	124.5	*25*	394	112.1	*49*	411	112.9	*73*	428	131.6	*97*	378	103.8
2	525	148.3	*26*	434	118.6	*50*	451	124.3	*74*	510	140.6	*98*	576	170.1
3	503	146.6	*27*	518	151.3	*51*	466	130.3	*75*	478	126.6	*99*	452	116.1
4	482	148.2	*28*	522	143.8	*52*	433	130.0	*76*	421	115.1	*100*	543	155.4
5	470	124.0	*29*	511	149.5	*53*	492	137.5	*77*	510	153.9	*101*	538	165.0
6	400	114.6	*30*	437	124.3	*54*	503	148.5	*78*	351	102.9	*102*	523	172.8
7	398	109.0	*31*	352	87.7	*55*	451	128.6	*79*	493	149.7	*103*	434	108.7
8	514	174.6	*32*	406	112.4	*56*	415	107.1	*80*	411	115.2	*104*	458	128.0
9	518	154.0	*33*	448	125.9	*57*	459	145.4	*81*	422	108.6			
10	383	109.7	*34*	493	129.7	*58*	442	123.4	*82*	402	120.8			
11	412	117.9	*35*	468	128.9	*59*	424	117.1	*83*	438	126.7			
12	320	64.5	*36*	345	95.9	*60*	397	108.6	*84*	485	138.6			
13	473	137.9	*37*	523	152.6	*61*	414	113.5	*85*	496	155.3			
14	438	134.1	*38*	498	144.3	*62*	437	129.2	*86*	453	126.4			
15	359	71.9	*39*	482	139.9	*63*	512	169.9	*87*	377	96.1			
16	569	157.4	*40*	487	146.0	*64*	525	165.9	*88*	540	156.7			
17	423	115.9	*41*	331	84.6	*65*	546	177.0	*89*	502	137.2			
18	460	140.7	*42*	416	120.5	*66*	422	122.9	*90*	408	110.0			
19	372	81.7	*43*	358	98.3	*67*	495	150.9	*91*	417	124.3			
20	383	107.4	*44*	463	144.9	*68*	452	131.0	*92*	474	132.5			
21	409	116.7	*45*	462	138.8	*69*	465	140.7	*93*	480	153.9			
22	444	130.0	*46*	413	110.8	*70*	391	107.5	*94*	483	130.3			
23	463	136.7	*47*	506	153.5	*71*	426	128.2	*95*	472	135.6			
24	482	150.1	*48*	465	140.9	*72*	482	136.4	*96*	477	146.0			

В таблице 2 представлена выборка только для nu.

Таблица 2

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	460	25	394	49	411	73	428	97	378
2	525	26	434	50	451	74	510	98	576
3	503	27	518	51	466	75	478	99	452
4	482	28	522	52	433	76	421	100	543
5	470	29	511	53	492	77	510	101	538
6	400	30	437	54	503	78	351	102	523
7	398	31	352	55	451	79	493	103	434
8	514	32	406	56	415	80	411	104	458
9	518	33	448	57	459	81	422		
10	383	34	493	58	442	82	402		
11	412	35	468	59	424	83	438		
12	320	36	345	60	397	84	485		
13	473	37	523	61	414	85	496		
14	438	38	498	62	437	86	453		
15	359	39	482	63	512	87	377		
16	569	40	487	64	525	88	540		
17	423	41	331	65	546	89	502		
18	460	42	416	66	422	90	408		
19	372	43	358	67	495	91	417		
20	383	44	463	68	452	92	474		
21	409	45	462	69	465	93	480		
22	444	46	413	70	391	94	483		
23	463	47	506	71	426	95	472		
24	482	48	465	72	482	96	477		

- Ранжированный ряд

В таблице 3 представлено преобразование выборки в ранжированный ряд.

Таблица 3

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	320	25	413	49	452	73	482	97	525
2	331	26	414	50	452	74	483	98	525
3	345	27	415	51	453	75	485	99	538
4	351	28	416	52	458	76	487	100	540
5	352	29	417	53	459	77	492	101	543
6	358	30	421	54	460	78	493	102	546

7	359	*31*	422	*55*	460	*79*	493	*103*	569
8	372	*32*	422	*56*	462	*80*	495	*104*	576
9	377	*33*	423	*57*	463	*81*	496		
10	378	*34*	424	*58*	463	*82*	498		
11	383	*35*	426	*59*	465	*83*	502		
12	383	*36*	428	*60*	465	*84*	503		
13	391	*37*	433	*61*	466	*85*	503		
14	394	*38*	434	*62*	468	*86*	506		
15	397	*39*	434	*63*	470	*87*	510		
16	398	*40*	437	*64*	472	*88*	510		
17	400	*41*	437	*65*	473	*89*	511		
18	402	*42*	438	*66*	474	*90*	512		
19	406	*43*	438	*67*	477	*91*	514		
20	408	*44*	442	*68*	478	*92*	518		
21	409	*45*	444	*69*	480	*93*	518		
22	411	*46*	448	*70*	482	*94*	522		
23	411	*47*	451	*71*	482	*95*	523		
24	412	*48*	451	*72*	482	*96*	523		

В таблице 3 можно заметить, что наименьшее значение в выборке $x_{min} = 320$, а наибольшее значение $x_{max} = 576$.

- Вариационный ряд

В таблицах 4 и 5 представлено преобразование полученной выборки в вариационный ряд с абсолютными и относительными частотами соответственно.

Таблица 4

i	x_i	n_i	i	x_i	n_i	i	x_i	n_i	i	x_i	n_i
1	320	1	*22*	412	1	*43*	453	1	*64*	493	2
2	331	1	*23*	413	1	*44*	458	1	*65*	495	1
3	345	1	*24*	414	1	*45*	459	1	*66*	496	1
4	351	1	*25*	415	1	*46*	460	2	*67*	498	1
5	352	1	*26*	416	1	*47*	462	1	*68*	502	1
6	358	1	*27*	417	1	*48*	463	2	*69*	503	2
7	359	1	*28*	421	1	*49*	465	2	*70*	506	1
8	372	1	*29*	422	2	*50*	466	1	*71*	510	2
9	377	1	*30*	423	1	*51*	468	1	*72*	511	1
10	378	1	*31*	424	1	*52*	470	1	*73*	512	1

11	383	2	*32*	426	1	*53*	472	1	*74*	514	1
12	391	1	*33*	428	1	*54*	473	1	*75*	518	2
13	394	1	*34*	433	1	*55*	474	1	*76*	522	1
14	397	1	*35*	434	2	*56*	477	1	*77*	523	2
15	398	1	*36*	437	2	*57*	478	1	*78*	525	2
16	400	1	*37*	438	2	*58*	480	1	*79*	538	1
17	402	1	*38*	442	1	*59*	482	4	*80*	540	1
18	406	1	*39*	444	1	*60*	483	1	*81*	543	1
19	408	1	*40*	448	1	*61*	485	1	*82*	546	1
20	409	1	*41*	451	2	*62*	487	1	*83*	569	1
21	411	2	*42*	452	2	*63*	492	1	*84*	576	1

Таблица 5

i	x_i	$\overline{n_i}$	i	x_i	$\overline{n_i}$	i	x_i	$\overline{n_i}$	i	x_i	$\overline{n_i}$
1	320	0.0096	*22*	412	0.0096	*43*	453	0.0096	*64*	493	0.0192
2	331	0.0096	*23*	413	0.0096	*44*	458	0.0096	*65*	495	0.0096
3	345	0.0096	*24*	414	0.0096	*45*	459	0.0096	*66*	496	0.0096
4	351	0.0096	*25*	415	0.0096	*46*	460	0.0192	*67*	498	0.0096
5	352	0.0096	*26*	416	0.0096	*47*	462	0.0096	*68*	502	0.0096
6	358	0.0096	*27*	417	0.0096	*48*	463	0.0192	*69*	503	0.0192
7	359	0.0096	*28*	421	0.0096	*49*	465	0.0192	*70*	506	0.0096
8	372	0.0096	*29*	422	0.0192	*50*	466	0.0096	*71*	510	0.0192
9	377	0.0096	*30*	423	0.0096	*51*	468	0.0096	*72*	511	0.0096
10	378	0.0096	*31*	424	0.0096	*52*	470	0.0096	*73*	512	0.0096
11	383	0.0192	*32*	426	0.0096	*53*	472	0.0096	*74*	514	0.0096
12	391	0.0096	*33*	428	0.0096	*54*	473	0.0096	*75*	518	0.0192
13	394	0.0096	*34*	433	0.0096	*55*	474	0.0096	*76*	522	0.0096
14	397	0.0096	*35*	434	0.0192	*56*	477	0.0096	*77*	523	0.0192
15	398	0.0096	*36*	437	0.0192	*57*	478	0.0096	*78*	525	0.0192
16	400	0.0096	*37*	438	0.0192	*58*	480	0.0096	*79*	538	0.0096
17	402	0.0096	*38*	442	0.0096	*59*	482	0.0385	*80*	540	0.0096
18	406	0.0096	*39*	444	0.0096	*60*	483	0.0096	*81*	543	0.0096
19	408	0.0096	*40*	448	0.0096	*61*	485	0.0096	*82*	546	0.0096
20	409	0.0096	*41*	451	0.0192	*62*	487	0.0096	*83*	569	0.0096
21	411	0.0192	*42*	452	0.0192	*63*	492	0.0096	*84*	576	0.0096

- Интервальный ряд

С помощью формулы Стерджесса было вычислено количество интервалов:

$$k = 1 + 3.31 * \lg N = 7$$

Получено нечетное количество интервалов.

Ширина интервала h была вычислена по формуле:

$$h = \frac{x_{max} - x_{min}}{k} = \frac{576 - 320}{7} \approx 37$$

В таблице 6 представлен полученный интервальный ряд.

Таблица 6

Границы интервалов	***Середины интервалов***	***Абсолютная частота***	***Относительная частота***
[320, 357)	338.5	5	0.048
[357, 394)	375.5	8	0.077
[394, 431)	412.5	23	0.221
[431, 468)	449.5	25	0.240
[468, 505)	486.5	24	0.231
[505, 542)	523.5	15	0.144
[542, 576)	559	4	0.038

- Графики для интервального ряда абсолютных частот

Полигон представлен на рис. 1.2.1.

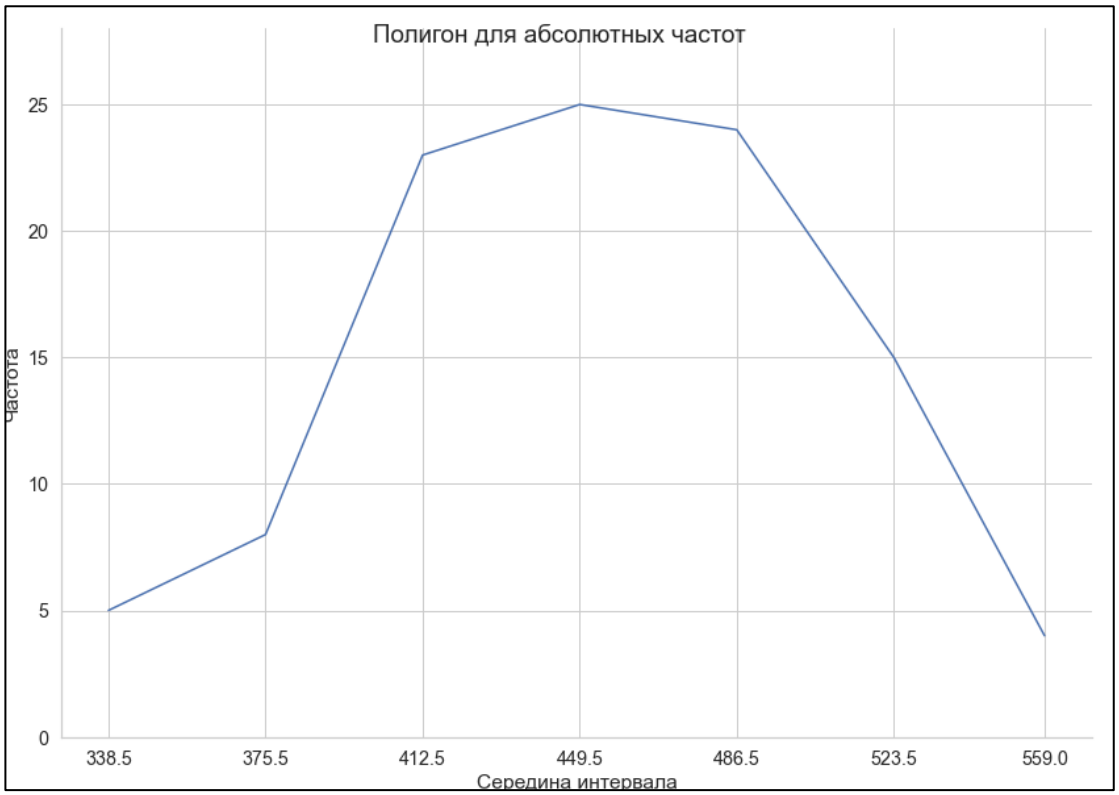

Рисунок 1.2.1 – Полигон для абсолютных частот

Гистограмма, представлена на рис. 1.2.2.

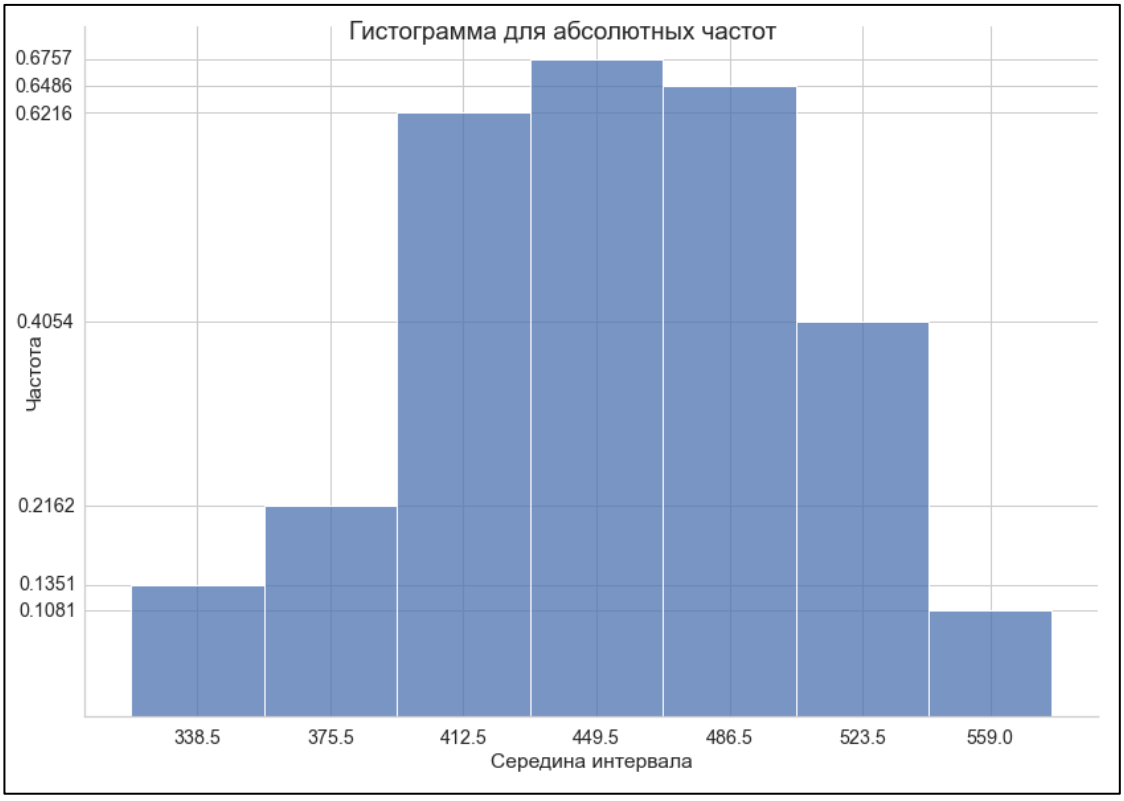

Рисунок 1.2.2 – Гистограмма для абсолютных частот

- Графики для интервального ряда относительных частот

Полигон представлен на рис. 1.2.3.

Рисунок 1.2.3 – Полигон для относительных частот

Гистограмма, представлена на рис. 1.2.4.

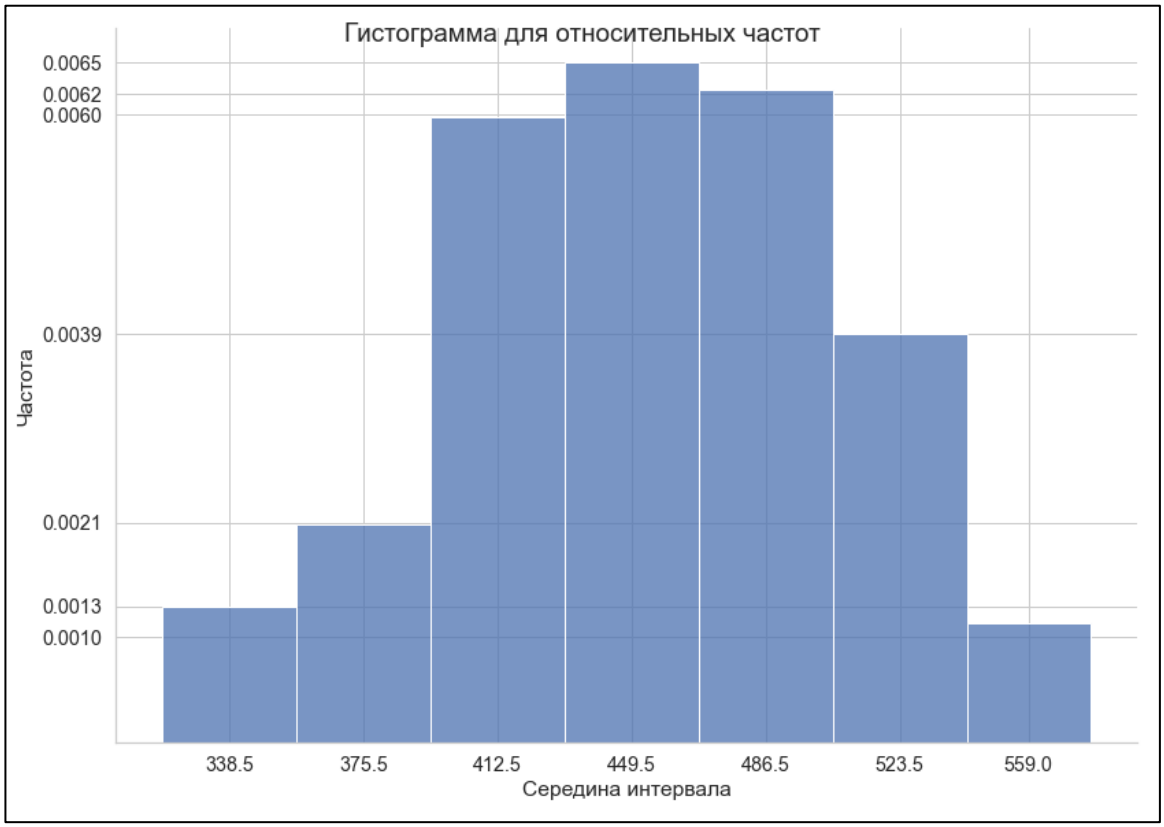

Рисунок 1.2.4 – Гистограмма для относительных частот

Эмпирическая функция распределения, построенная применительно к интервальному ряду для относительных частот представлен на рис. 1.2.5.

Функция распределения:

$$F(x) = \begin{cases} 0, & x = 338.5 \\ 0.048, & x = 375.5 \\ 0.125, & x = 412.5 \\ 0.346, & x = 449.5 \\ 0.587, & x = 486.5 \\ 0.817, & x = 523.5 \\ 0.962, & x = 559 \\ 1, & x > 559 \end{cases}$$

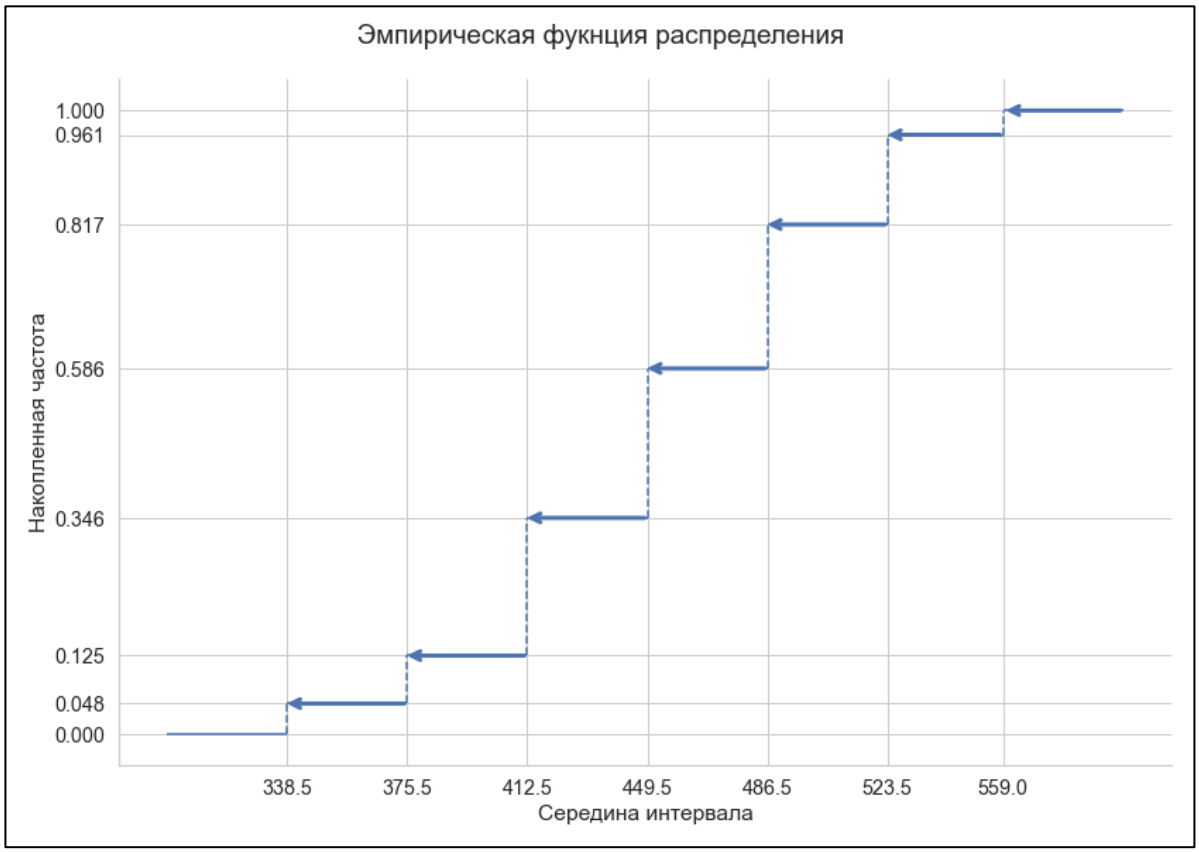

Рисунок 1.2.5 – График эмпирической функции распределения

1.3. Нахождение точечных оценок параметров распределения.

Условные варианты можно найти как $u_j = \frac{x_j - C}{h}$, где C – условный ноль.

Условные моменты k-го порядка:

$$\overline{M_k^*} = \frac{1}{N}\sum n_j \left(\frac{x_j - C}{h}\right)^k = \frac{1}{N}\sum n_j u_j^k$$

Результаты вычислений представлены в табл. 7.

Таблица 7

v	n	u	$n * u$	$n * u^2$	$n * u^3$	$n * u^4$	$n * (u+1)^4$
338.5	0.048	-3	-0.144	0.432	-1.296	3.888	0.768
375.5	0.077	-2	-0.154	0.308	-0.616	1.232	0.077
412.5	0.221	-1	-0.221	0.221	-0.221	0.221	0.0
449.5	0.240	0	0.0	0.0	0.0	0.0	0.24
486.5	0.231	1	0.231	0.231	0.231	0.231	3.696
523.5	0.144	2	0.288	0.576	1.152	2.304	11.664
559	0.039	3	0.117	0.351	1.053	3.159	9.984
∑	1	–	0.117	2.119	0.303	11.035	26.429

Сумма элементов последнего столбца является контрольной суммой, и так как в данном случае во втором столбце записаны относительные частоты, должно быть выполнено равенство:

$$\sum n_j * u_j^4 + 4 * \sum n_j * u_j^3 + 6 * \sum n_j * u_j^2 + 4 * \sum n_j * u_j + 1 = \sum n_j * (u_j + 1)^4$$

$$11.035 + 4 * 0.303 + 6 * 2.119 + 4 * 0.117 + 1 = 26.429$$

Эмпирические начальные и центральные моменты вычислены ниже:

$$\overline{x_в} = \overline{M_1} = \overline{M_1^*}h + C = 453.829$$

$$D_в = \overline{m_2} = \left(\overline{M_2^*} - \left(\overline{M_1^*}\right)^2\right)h^2 = 2882.171$$

$$\overline{m_3} = \left(\overline{M_3^*} - 3\overline{M_2^*}\,\overline{M_1^*} + 2\left(\overline{M_1^*}\right)^3\right)h^3 = -22164.019$$

$$\overline{m_4} = \left(\overline{M_4^*} - 4\overline{M_3^*}\,\overline{M_1^*} + 6\overline{M_2^*}\left(\overline{M_1^*}\right)^2 + 2\left(\overline{M_1^*}\right)^4\right)h^4 = 20740732.146$$

Найдем выборочное среднее и дисперсию с помощью стандартных формул.

Статистическая оценка математического ожидания:

$$\overline{x_в} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = 453.716$$

Статистическая оценка дисперсии:

$$D_{\text{в}} = \frac{1}{N}\sum_{i=1}^{k}(x_i - \overline{x_{\text{в}}})^2\, n_i = 2865.503$$

Данная статистическая оценка является смещенной оценкой, поэтому вычислим исправленную оценку дисперсии:

$$s^2 = \frac{N}{N-1} D_{\text{в}} = \frac{104}{103} * 2865.503 = 2893.324$$

Статистические оценки СКО:

$$\sigma_{\text{в}} = \sqrt{D_{\text{В}}} = \sqrt{2865.503} = 53.53$$

$$s = \sqrt{s^2} = \sqrt{2893.324} = 53.78$$

Статистические оценки математического ожидания и дисперсии, вычисленные по стандартным формулам и с помощью условных вариант совпадают с небольшой погрешностью.

Статистические оценки коэффициентов асимметрии и эксцесса можно вычислить по формулам:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3}$$

$$\bar{E} = \frac{\overline{m_4}}{s^3} - 3$$

Центральные эмпирические моменты третьего и четвертого порядков были найдены выше.

Статистическая оценка коэффициента асимметрии:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3} = \frac{-22164.019}{53.78^3} = -0.000000915$$

Статистическая оценка коэффициента эксцесса:

$$\overline{\text{E}} = \frac{\overline{m_4}}{s^4} - 3 = \frac{20740732.146}{53.78^4} - 3 = -2.99$$

Коэффициент асимметрии отрицателен, следовательно, в данном случае это левосторонняя асимметрия, которая характеризуется удлиненным левым хвостом, а также неравенством $\overline{x_{\text{в}}} < M_o$, но полученное значение незначительно и скос распределения небольшой.

Коэффициент эксцесса также отрицателен, следовательно, эмпирическое распределение является более низким и пологим относительно нормального распределения.

Вычислим моду и медиану заданного распределения для интервального ряда.

Мода заданного распределения:

$$M_o = x_0 + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} h,$$

$$M_o = 431 + 37\frac{25 - 23}{(25 - 23) + (25 - 24)} = 455.67$$

Медиана заданного распределения:

$$M_e = x_o + \frac{0.5n - n_{m-1}^n}{n_m} h,$$

$$M_e = 431 + \frac{0.5 * 104 - 36}{25} 37 = 454.68$$

1.4. Нахождение интервальных оценок параметров распределения. Проверка статистической гипотезы о нормальном законе распределения.

Вычислим точность и доверительный интервал для математического ожидания при неизвестном СКО для доверительной точности γ

Случайная величина t:

$$t = \frac{\overline{x_в} - a}{s/\sqrt{N}}$$

Эта случайная величина распределена по закону Стьюдента с $k = N - 1$ степенями свободы. Справедливо соотношение:

$$P\left(\left|\frac{\overline{x_в} - a}{s/\sqrt{N}}\right| < t_\gamma\right) = 2\int_0^{t_\gamma} S(t, N)dt = \gamma$$

$$P\left(\overline{x_в} - t_\gamma\, s/\sqrt{N} < a < \overline{x_в} + t_\gamma\, s/\sqrt{N}\right) = \gamma$$

Доверительный интервал для оценки математического ожидания:

$$\left(\overline{x_в} - t_\gamma \frac{s}{\sqrt{N}}, \overline{x_в} + t_\gamma \frac{s}{\sqrt{N}}\right), \text{где}$$

$\bar{x_в}$ – выборочное среднее

s – исправленное СКО

$t_\gamma = 1.984$ – определено из соответствующей таблицы

(по заданным значениям $\gamma = 0.95, N = 104$)

$$\bar{x_в} - t_\gamma \frac{s}{\sqrt{N}} = 453.71 - 1.984 * \frac{53.79}{\sqrt{104}} = 443.25$$

$$\bar{x_в} + t_\gamma \frac{s}{\sqrt{N}} = 453.71 + 1.984 * \frac{53.79}{\sqrt{104}} = 464.17$$

Можно сделать вывод, что интервал $(443.25; 464.17)$ с вероятностью (надежностью) $\gamma = 0.95$ содержит в себе истинное значение математического ожидания.

Построим доверительный интервал для среднеквадратического отклонения:

$$P(s - \delta < \sigma < s + \delta) = \gamma$$

$$P(s(1 - \delta/s) \ < \sigma < s(1 + \delta/s)\ ) = \gamma$$

$$q = \delta/s$$

Доверительный интервал для оценки СКО:

$$s(1 - q) < \sigma < s(1 + q), \text{где}$$

s – исправленное СКО

$q = 0.141$ – определено из соответствующей таблицы

(по заданным значениям $\gamma = 0.95, N = 104$)

$$s(1 - q) = 53.79 * 0.859 = \ 46.206$$

$$s(1 + q) = 53.79 * 1.141 = 61.374$$

Можно сделать вывод, что интервал $(46.206; 61.374)$ с вероятностью (надежностью) $\gamma = 0.95$ содержит в себе истинное значение среднеквадратического отклонения.

Проверим гипотезу о нормальности заданного распределения с помощью критерия Пирсона χ^2

Гипотеза H_0 – выборочные данные представляют значения случайной величины, распределённой по нормальному закону распределения. Согласно критерию Пирсона, вычисляется наблюдаемое значение случайной величины χ^2:

$$\chi^2_{\text{набл}} = \sum_{1}^{K} \frac{(n_i - n_i')^2}{n_i'}$$

Распределение хи-квадрат зависит от числа степеней свободы k, которое вычисляется как $k = K - 3$. По числу степеней свободы и уровню значимости вычисляется значение $\chi^2_{\text{крит}} = \chi^2(\alpha, k)$. Область принятия гипотезы H_0 определяется условием:

$$\chi^2_{\text{набл}} \leq \chi^2_{\text{крит}}$$

Найдем теоретические частоты. Вычисления представлены в табл. 8.

Таблица 8

x_i	x_{i+1}	n_i	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	p_i	n_i'
320.0	357.0	5	$-\infty$	-1.8	-0.5	-0.4641	0.0359	3.7336
357.0	394.0	8	-1.8	-1.11	-0.4641	-0.3665	0.0976	10.1504
394.0	431.0	23	-1.11	-0.42	-0.3665	-0.1628	0.2037	21.1848
431.0	468.0	25	-0.42	0.27	-0.1628	0.1064	0.2692	27.9968
468.0	505.0	24	0.27	0.95	0.1064	0.3289	0.2225	23.14
505.0	542.0	15	0.95	1.64	0.3289	0.4495	0.1206	12.5424
542.0	576.0	4	1.64	$+\infty$	0.4495	0.5	0.0505	5.252

Вычислим наблюдаемое значение критерия $\chi^2_{\text{набл}}$. Результаты представлены в табл. 9.

Таблица 9

n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$(n_i - n_i')^2/n_i'$
5	3.7336	1.2664	1.6038	0.4296

8	10.1504	-2.1504	4.6242	0.4556
23	21.1848	1.8152	3.295	0.1555
25	27.9968	-2.9968	8.9808	0.3208
24	23.14	0.86	0.7396	0.032
15	12.5424	2.4576	6.0398	0.4816
4	5.252	-1.252	1.5675	0.2985

$$\chi^2_{\text{набл}} = \sum_{1}^{K} \frac{(n_i - n_i')^2}{n_i'} = 2.1736$$

Найдем $\chi^2_{\text{крит}}$ по заданному уровню значимости $\alpha = 0.05$ и числу степеней свободы $k = K - 3 = 4$: $\chi^2_{\text{крит}} = 9.5$

Сравним $\chi^2_{\text{крит}}$ с наблюдаемым значением:

$$\chi^2_{\text{набл}} = 2.1736;\ \chi^2_{\text{крит}} = 9.5$$

$$\chi^2_{\text{набл}} \leq \chi^2_{\text{крит}}$$

Из полученных результатов можно сделать вывод, что выдвинутая нулевая гипотеза принимается, то есть выборочные данные позволяют предположить, что случайная величина распределена по нормальному закону распределения.

1.5. Выводы.

Из генеральной совокупности была сформирована репрезентативная выборка. Выборка была преобразована в ранжированный, вариационный и интервальный ряды. Используя полученный интервальный ряд был построен полигон и гистограмма для абсолютных и относительных частот. Для интервального ряда относительных частот был построен график эмпирической функции распределения.

Элементы ранжированного ряда расположены в порядке возрастания их значений, поэтому можно определить минимальный и максимальный элемент выборки. Для данной выборки $x_{min} = 320$, $x_{max} = 576$.

Вариационный ряд получается в результате объединения одинаковых элементов, поэтому можно определить варианту с наибольшей частотой повторения в выборке. Для данной выборки это $x_{59} = 482$ с абсолютной частотой $n_{59} = 4$ и относительной частотой $\overline{n_{59}} = 0.0385$.

Интервальный ряд был построен с помощью деления вариационного ряда на интервалы. По формуле Стерджесса было получено нечетное количество интервалов $k = 7$. По сформированному интервальному ряду можно увидеть, что наибольшая частота попадания значений вариант в интервале $[431, 468)$.

Такой же результат можно увидеть на построенных полигоне и гистограмме. Форма графиков не меняется для абсолютных и относительных частот, меняется ось ординат, которая для полигонов обозначает частоты (абсолютные или относительные), а для гистограмм уже площадь прямоугольника обозначает частоты, что можно проверить путем умножения высоты столбца на ширину $h = 37$. Сумма площадей прямоугольников гистограммы для абсолютных частот равна объему выборки $n = 104$, а для относительных частот равна 1. На графике эмпирической функции распределения можно увидеть отношение накопленных частот до середины интервалов к объему выборки.

По виду полигона и гистограммы можно сделать предположение о том, что анализируемая переменная имеет примерно нормальное распределение.

Для интервального ряда были найдены середины интервалов и накопленные частоты, далее для полученных вариант были вычислены условные варианты. Были вычислены условные эмпирические моменты через условные варианты, и с их помощью вычислены начальные и центральные эмпирические моменты. Корректность вычислений была проверена контрольной суммой, которая дала понять, что вычисления были верны.

Были посчитаны выборочное среднее и дисперсия с помощью стандартных формул и с помощью условных вариант. Статистические оценки, вычисленные по стандартным формулам и с помощью условных вариант совпали.

Была найдена статистическая оценка коэффициентов асимметрии и эксцесса. Коэффициент асимметрии получился отрицательным, то есть — это левосторонняя асимметрия, которая характеризуется удлиненным левым хвостом, а также неравенством $\bar{x_в} < M_o$, но полученное значение незначительно и скос распределения небольшой. Коэффициент эксцесса также отрицателен, следовательно, эмпирическое распределение является более низким и пологим относительно нормального распределения. Данные наблюдения также можно увидеть на рисунке 1.

Для интервального ряда была вычислена мода и медиана. Мода оказалась смещена относительно центра модального интервала в сторону правого интервала с большей частотой. Медиана также смещена правее, так как по правую сторону находится большее количество вариант.

Вычислен доверительный интервал для математического ожидания при неизвестном СКО с доверительной точностью $\gamma = 0.95$. Исходя из полученных результатов можно сделать вывод, что интервал $(443.25; 464.17)$ с вероятностью (надежностью) $\gamma = 0.95$ содержит в себе истинное значение математического ожидания. Были вычислены границы доверительного интервала для среднеквадратического отклонения. Определено, что интервал $(46.206; 61.374)$ с вероятностью (надежностью) $\gamma = 0.95$ содержит в себе истинное значение среднеквадратического отклонения.

Была выполнена проверка гипотезы о нормальности заданного распределения с помощью критерия χ^2 (Пирсона). Было выяснено, что $\chi^2_{\text{набл}} \leq \chi^2_{\text{крит}}$, следовательно, выдвинутая нулевая гипотеза принимается, то есть выборочные данные позволяют предположить, что случайная величина распределена по нормальному закону распределения.

2. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

2.1. Основные теоретические положения.

Корреляционный анализ.

Рассмотрим систему двух случайных величин $\{X; Y\}$. Эти случайные величины могут быть независимыми:

$$f(x, y) = f_1(x) \cdot f_2(y)$$

В противном случае между ними может быть:

- Функциональная зависимость:

$$y = g(x)$$

- Статистическая зависимость:

$$\varphi(x/y) = \frac{f(x, y)}{f_2(y)}$$

$$\phi(y/x) = \frac{f(x, y)}{f_1(x)}$$

Частным случаем статистической зависимости является корреляционная зависимость. Корреляционной называют статистическую зависимость двух случайных величин, при которой изменение значения одной из случайных величин приводит к изменению математического ожидания другой случайной величины:

$$M(X/y) = q_1(y)$$

$$M(Y/x) = q_2(x)$$

Корреляционный момент:

$$\mu_{xy} = M\{[x - M(X)] \cdot [y - M(Y)]\}$$

Коэффициент корреляции:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y}$$

Для коэффициента корреляции справедливо соотношение:

$$|r_{xy}| \leq 1$$

Случайные величины называют коррелированными, если их корреляционный момент или их коэффициент корреляции отличен от нуля. В противном случае эти величины некоррелированные. Если случайные величины X и Y коррелированы, то они зависимы.

Коэффициент корреляции служит мерой тесноты линейной зависимости между случайными величинами X и Y. При $|r_{xy}| = 1$ эта зависимость становится функциональной.

Значение $\bar{r}_{xy}$ – статистической оценки r_{xy} – коэффициента корреляции можно вычислить по формуле:

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij} y_i x_j - N\bar{x}_{\text{в}}\bar{y}_{\text{в}}}{N S_x S_y}$$

При $N > 50$ в случае нормального распределения системы случайных величин $\{X; Y\}$ для оценки значения $\bar{r}_{xy}$ можно использовать соотношение:

$$\bar{r}_{xy} - 3\frac{1-\bar{r}_{xy}^2}{\sqrt{N}} \le r_{xy} \le \bar{r}_{xy} + 3\frac{1+\bar{r}_{xy}^2}{\sqrt{N}}$$

Для построения доверительного интервала с помощью преобразования Фишера перейдём к случайной величине z:

$$\bar{z} = 0.5 \ln \frac{1+\bar{r}_{xy}}{1-\bar{r}_{xy}}$$

Распределение z при неограниченном возрастании объёма выборки асимптотически нормальное со значением СКО:

$$\bar{\sigma}_z = \frac{1}{\sqrt{N-3}}$$

Доверительный интервал для генерального значения:

$$(\bar{z} - \lambda(\gamma)\bar{\sigma}_z; \bar{z} + \lambda(\gamma)\bar{\sigma}_z), \text{где}$$

$$\Phi\big(\lambda(\gamma)\big) = \frac{\gamma}{2}$$

Для пересчёта интервала в доверительный интервал для коэффициента корреляции с тем же значением γ необходимо воспользоваться обратным преобразованием Фишера:

$$r = th(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Гипотеза $H_0: r_{xy} = 0$. Альтернативой будет гипотеза $H_1: r_{xy} \neq 0$. Если основная гипотеза отвергается, то это означает, что выборочный коэффициент корреляции $\bar{r}_{xy}$ значимо отличается от нуля (значим). В качестве критерия проверки статистической гипотезы о значимости выборочного коэффициента корреляции можно принять случайную величину:

$$T = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}}$$

При справедливости нулевой гипотезы случайная величина T распределена по закону Стьюдента с $k = K - 2$ степенями свободы. Критическая область для данного критерия двусторонняя. Если $|T_{\text{набл}}| \leq t_{\text{крит}}(\alpha, k)$ – нет оснований отвергать гипотезу H_0. Если $|T_{\text{набл}}| > t_{\text{крит}}(\alpha, k)$ – основная гипотеза H_0 с выборочными данными должна быть отвергнута.

Метод наименьших квадратов (МНК) — метод, основанный на поиске минимума суммы квадратов отклонений значений некоторых функций от заданного множества значений. МНК является одним из основных методов регрессионного анализа и применяется для оценки параметров регрессионных моделей на основе выборочных данных.

$$M(X/y) = q_1(y)$$
$$M(Y/x) = q_2(x)$$

Пусть имеется двумерная случайная величина $\{X, Y\}$, где X и Y зависимые случайные величины. Функцию $g(x)$ называют линейной функцией среднеквадратической регрессии Y на X.

$$g(x) = m\left(\frac{Y}{x}\right) = m(Y) + r_{xy}\frac{\sigma_y}{\sigma_x}[x - m(X)]$$

В случае, когда известны только выборочные данные – двумерная выборка значений случайных величин X и Y, возможно построение только выборочных прямых среднеквадратической регрессии.

Уравнения выборочных прямых среднеквадратической регрессии:

$$\overline{y_x} = \overline{y_в} + \overline{r_{xy}}\frac{S_y}{S_x}(x - \overline{x_в})$$

$$\overline{x_y} = \overline{x_в} + \overline{r_{xy}}\frac{S_x}{S_y}(y - \overline{y_в})$$

Для оценки корреляционной зависимости между случайными величинами в общем, а не только линейной, может быть использовано так называемое корреляционной отношение.

Оценку общей дисперсии X можно представить, как сумму внутригрупповой и межгрупповой дисперсии:

$$D_{общ} = D_{межгр} + D_{внгр}$$

Внутригрупповая дисперсия вычисляется, как взвешенная по объемам групп средняя арифметическая групповых дисперсий.

Межгрупповая дисперсия вычисляется, как дисперсия условных (групповых) средних $\overline{x_{y_i}}$ относительно выборочной средней $\overline{x_в}$.

Выборочное корреляционное отношение Y к X определяется в соответствии с выражением:

$$\overline{\eta_{yx}} = \frac{\overline{\sigma_{\overline{y_x}}}}{\overline{\sigma_y}} = \sqrt{\frac{D_{межгр}}{D_{общ}}},$$

где $\overline{\sigma_{\overline{y_x}}} = \sqrt{D_{межгр}}$, $\overline{\sigma_y} = \sqrt{D_{общ}}$ – выборочные значения СКВО $\overline{y_x}$ и Y соответственно. Аналогично определяется выборочное корреляционное отношение X к Y.

Выборочное уравнение регрессии Y на X параболического вида:

$$\overline{y_x} = ax^2 + bx + c$$

Значения коэффициентов a, b и c определим с помощью МНК, что приводит к необходимости решать систему линейных уравнений третьего порядка.

2.2. Элементы корреляционного анализа. Проверка статистической гипотезы о равенстве коэффициента корреляции нулю.

Формирование репрезентативной выборки заданного объема из имеющейся генеральной совокупности экспериментальных данных представлено в таблице 10. Объём выборки: 104.

Таблица 10

№	nu	E	№	nu	E	№	nu	E	№	nu	E	№	nu	E
1	460	124.5	*25*	394	112.1	*49*	411	112.9	*73*	428	131.6	*97*	378	103.8
2	525	148.3	*26*	434	118.6	*50*	451	124.3	*74*	510	140.6	*98*	576	170.1
3	503	146.6	*27*	518	151.3	*51*	466	130.3	*75*	478	126.6	*99*	452	116.1
4	482	148.2	*28*	522	143.8	*52*	433	130.0	*76*	421	115.1	*100*	543	155.4
5	470	124.0	*29*	511	149.5	*53*	492	137.5	*77*	510	153.9	*101*	538	165.0
6	400	114.6	*30*	437	124.3	*54*	503	148.5	*78*	351	102.9	*102*	523	172.8
7	398	109.0	*31*	352	87.7	*55*	451	128.6	*79*	493	149.7	*103*	434	108.7
8	514	174.6	*32*	406	112.4	*56*	415	107.1	*80*	411	115.2	*104*	458	128.0
9	518	154.0	*33*	448	125.9	*57*	459	145.4	*81*	422	108.6			
10	383	109.7	*34*	493	129.7	*58*	442	123.4	*82*	402	120.8			
11	412	117.9	*35*	468	128.9	*59*	424	117.1	*83*	438	126.7			
12	320	64.5	*36*	345	95.9	*60*	397	108.6	*84*	485	138.6			
13	473	137.9	*37*	523	152.6	*61*	414	113.5	*85*	496	155.3			
14	438	134.1	*38*	498	144.3	*62*	437	129.2	*86*	453	126.4			
15	359	71.9	*39*	482	139.9	*63*	512	169.9	*87*	377	96.1			
16	569	157.4	*40*	487	146.0	*64*	525	165.9	*88*	540	156.7			
17	423	115.9	*41*	331	84.6	*65*	546	177.0	*89*	502	137.2			
18	460	140.7	*42*	416	120.5	*66*	422	122.9	*90*	408	110.0			
19	372	81.7	*43*	358	98.3	*67*	495	150.9	*91*	417	124.3			
20	383	107.4	*44*	463	144.9	*68*	452	131.0	*92*	474	132.5			
21	409	116.7	*45*	462	138.8	*69*	465	140.7	*93*	480	153.9			
22	444	130.0	*46*	413	110.8	*70*	391	107.5	*94*	483	130.3			
23	463	136.7	*47*	506	153.5	*71*	426	128.2	*95*	472	135.6			
24	482	150.1	*48*	465	140.9	*72*	482	136.4	*96*	477	146.0			

В таблице 11 представлена выборка только для E.

Таблица 11

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	124.5	25	112.1	49	112.9	73	131.6	97	103.8
2	148.3	26	118.6	50	124.3	74	140.6	98	170.1
3	146.6	27	151.3	51	130.3	75	126.6	99	116.1
4	148.2	28	143.8	52	130.0	76	115.1	100	155.4
5	124.0	29	149.5	53	137.5	77	153.9	101	165.0
6	114.6	30	124.3	54	148.5	78	102.9	102	172.8
7	109.0	31	87.7	55	128.6	79	149.7	103	108.7
8	174.6	32	112.4	56	107.1	80	115.2	104	128.0
9	154.0	33	125.9	57	145.4	81	108.6		
10	109.7	34	129.7	58	123.4	82	120.8		
11	117.9	35	128.9	59	117.1	83	126.7		
12	64.5	36	95.9	60	108.6	84	138.6		
13	137.9	37	152.6	61	113.5	85	155.3		
14	134.1	38	144.3	62	129.2	86	126.4		
15	71.9	39	139.9	63	169.9	87	96.1		
16	157.4	40	146.0	64	165.9	88	156.7		
17	115.9	41	84.6	65	177.0	89	137.2		
18	140.7	42	120.5	66	122.9	90	110.0		
19	81.7	43	98.3	67	150.9	91	124.3		
20	107.4	44	144.9	68	131.0	92	132.5		
21	116.7	45	138.8	69	140.7	93	153.9		
22	130.0	46	110.8	70	107.5	94	130.3		
23	136.7	47	153.5	71	128.2	95	135.6		
24	150.1	48	140.9	72	136.4	96	146.0		

В таблице 12 представлено преобразование выборки в ранжированный ряд.

Таблица 12

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	64.5	25	114.6	49	128.6	73	140.9	97	157.4
2	71.9	26	115.1	50	128.9	74	143.8	98	165.0
3	81.7	27	115.2	51	129.2	75	144.3	99	165.9
4	84.6	28	115.9	52	129.7	76	144.9	100	169.9
5	87.7	29	116.1	53	130.0	77	145.4	101	170.1
6	95.9	30	116.7	54	130.0	78	146.0	102	172.8
7	96.1	31	117.1	55	130.3	79	146.0	103	174.6
8	98.3	32	117.9	56	130.3	80	146.6	104	177.0
9	102.9	33	118.6	57	131.0	81	148.2		
10	103.8	34	120.5	58	131.6	82	148.3		

11	107.1	*35*	120.8	*59*	132.5	*83*	148.5		
12	107.4	*36*	122.9	*60*	134.1	*84*	149.5		
13	107.5	*37*	123.4	*61*	135.6	*85*	149.7		
14	108.6	*38*	124.0	*62*	136.4	*86*	150.1		
15	108.6	*39*	124.3	*63*	136.7	*87*	150.9		
16	108.7	*40*	124.3	*64*	137.2	*88*	151.3		
17	109.0	*41*	124.3	*65*	137.5	*89*	152.6		
18	109.7	*42*	124.5	*66*	137.9	*90*	153.5		
19	110.0	*43*	125.9	*67*	138.6	*91*	153.9		
20	110.8	*44*	126.4	*68*	138.8	*92*	153.9		
21	112.1	*45*	126.6	*69*	139.9	*93*	154.0		
22	112.4	*46*	126.7	*70*	140.6	*94*	155.3		
23	112.9	*47*	128.0	*71*	140.7	*95*	155.4		
24	113.5	*48*	128.2	*72*	140.7	*96*	156.7		

В таблице 12 можно заметить, что наименьшее значение в выборке $x_{min} = 64.5$, а наибольшее значение $x_{max} = 177$.

В таблицах 13 и 14 представлено преобразование полученной выборки в вариационный ряд с абсолютными и относительными частотами соответственно.

Таблица 13

i	x_i	n_i	i	x_i	n_i	i	x_i	n_i	i	x_i	n_i
1	64.5	1	*25*	115.1	1	*49*	129.7	1	*73*	146.6	1
2	71.9	1	*26*	115.2	1	*50*	130.0	2	*74*	148.2	1
3	81.7	1	*27*	115.9	1	*51*	130.3	2	*75*	148.3	1
4	84.6	1	*28*	116.1	1	*52*	131.0	1	*76*	148.5	1
5	87.7	1	*29*	116.7	1	*53*	131.6	1	*77*	149.5	1
6	95.9	1	*30*	117.1	1	*54*	132.5	1	*78*	149.7	1
7	96.1	1	*31*	117.9	1	*55*	134.1	1	*79*	150.1	1
8	98.3	1	*32*	118.6	1	*56*	135.6	1	*80*	150.9	1
9	102.9	1	*33*	120.5	1	*57*	136.4	1	*81*	151.3	1
10	103.8	1	*34*	120.8	1	*58*	136.7	1	*82*	152.6	1
11	107.1	1	*35*	122.9	1	*59*	137.2	1	*83*	153.5	1
12	107.4	1	*36*	123.4	1	*60*	137.5	1	*84*	153.9	2
13	107.5	1	*37*	124.0	1	*61*	137.9	1	*85*	154.0	1
14	108.6	2	*38*	124.3	3	*62*	138.6	1	*86*	155.3	1
15	108.7	1	*39*	124.5	1	*63*	138.8	1	*87*	155.4	1

16	109.0	1	*40*	125.9	1	*64*	139.9	1	*88*	156.7	1
17	109.7	1	*41*	126.4	1	*65*	140.6	1	*89*	157.4	1
18	110.0	1	*42*	126.6	1	*66*	140.7	2	*90*	165.0	1
19	110.8	1	*43*	126.7	1	*67*	140.9	1	*91*	165.9	1
20	112.1	1	*44*	128.0	1	*68*	143.8	1	*92*	169.9	1
21	112.4	1	*45*	128.2	1	*69*	144.3	1	*93*	170.1	1
22	112.9	1	*46*	128.6	1	*70*	144.9	1	*94*	172.8	1
23	113.5	1	*47*	128.9	1	*71*	145.4	1	*95*	174.6	1
24	114.6	1	*48*	129.2	1	*72*	146.0	2	*96*	177.0	1

Таблица 14

i	x_i	$\overline{n_i}$	i	x_i	$\overline{n_i}$	i	x_i	$\overline{n_i}$	i	x_i	$\overline{n_i}$
1	64.5	0.0096	*25*	115.1	0.0096	*49*	129.7	0.0096	*73*	146.6	0.0096
2	71.9	0.0096	*26*	115.2	0.0096	*50*	130.0	0.0192	*74*	148.2	0.0096
3	81.7	0.0096	*27*	115.9	0.0096	*51*	130.3	0.0192	*75*	148.3	0.0096
4	84.6	0.0096	*28*	116.1	0.0096	*52*	131.0	0.0096	*76*	148.5	0.0096
5	87.7	0.0096	*29*	116.7	0.0096	*53*	131.6	0.0096	*77*	149.5	0.0096
6	95.9	0.0096	*30*	117.1	0.0096	*54*	132.5	0.0096	*78*	149.7	0.0096
7	96.1	0.0096	*31*	117.9	0.0096	*55*	134.1	0.0096	*79*	150.1	0.0096
8	98.3	0.0096	*32*	118.6	0.0096	*56*	135.6	0.0096	*80*	150.9	0.0096
9	102.9	0.0096	*33*	120.5	0.0096	*57*	136.4	0.0096	*81*	151.3	0.0096
10	103.8	0.0096	*34*	120.8	0.0096	*58*	136.7	0.0096	*82*	152.6	0.0096
11	107.1	0.0096	*35*	122.9	0.0096	*59*	137.2	0.0096	*83*	153.5	0.0096
12	107.4	0.0096	*36*	123.4	0.0096	*60*	137.5	0.0096	*84*	153.9	0.0192
13	107.5	0.0096	*37*	124.0	0.0096	*61*	137.9	0.0096	*85*	154.0	0.0096
14	108.6	0.0192	*38*	124.3	0.0288	*62*	138.6	0.0096	*86*	155.3	0.0096
15	108.7	0.0096	*39*	124.5	0.0096	*63*	138.8	0.0096	*87*	155.4	0.0096
16	109.0	0.0096	*40*	125.9	0.0096	*64*	139.9	0.0096	*88*	156.7	0.0096
17	109.7	0.0096	*41*	126.4	0.0096	*65*	140.6	0.0096	*89*	157.4	0.0096
18	110.0	0.0096	*42*	126.6	0.0096	*66*	140.7	0.0192	*90*	165.0	0.0096
19	110.8	0.0096	*43*	126.7	0.0096	*67*	140.9	0.0096	*91*	165.9	0.0096
20	112.1	0.0096	*44*	128.0	0.0096	*68*	143.8	0.0096	*92*	169.9	0.0096
21	112.4	0.0096	*45*	128.2	0.0096	*69*	144.3	0.0096	*93*	170.1	0.0096
22	112.9	0.0096	*46*	128.6	0.0096	*70*	144.9	0.0096	*94*	172.8	0.0096
23	113.5	0.0096	*47*	128.9	0.0096	*71*	145.4	0.0096	*95*	174.6	0.0096
24	114.6	0.0096	*48*	129.2	0.0096	*72*	146.0	0.0192	*96*	177.0	0.0096

С помощью формулы Стерджесса было вычислено количество интервалов:

$$k = 1 + 3.31 * \lg N = 7$$

Получено нечетное количество интервалов.

Ширина интервала h была вычислена по формуле:

$$h = \frac{x_{max} - x_{min}}{k} = \frac{177 - 64.5}{7} \approx 16.1$$

В таблице 15 представлен полученный интервальный ряд.

Таблица 15

Границы интервалов	***Середины интервалов***	***Абсолютная частота***	***Относительная частота***
[64.5, 80.6)	72.55	2	0.019
[80.6, 96.7)	88.65	5	0.048
[96.7, 112.8)	104.75	15	0.144
[112.8, 128.9)	120.85	27	0.26
[128.9, 145.0)	136.95	27	0.26
[145.0, 161.1)	153.05	21	0.202
[161.1, 177.0)	169.05	7	0.067

- Графики для интервального ряда абсолютных частот

  Полигон представлен на рис. 2.2.1.

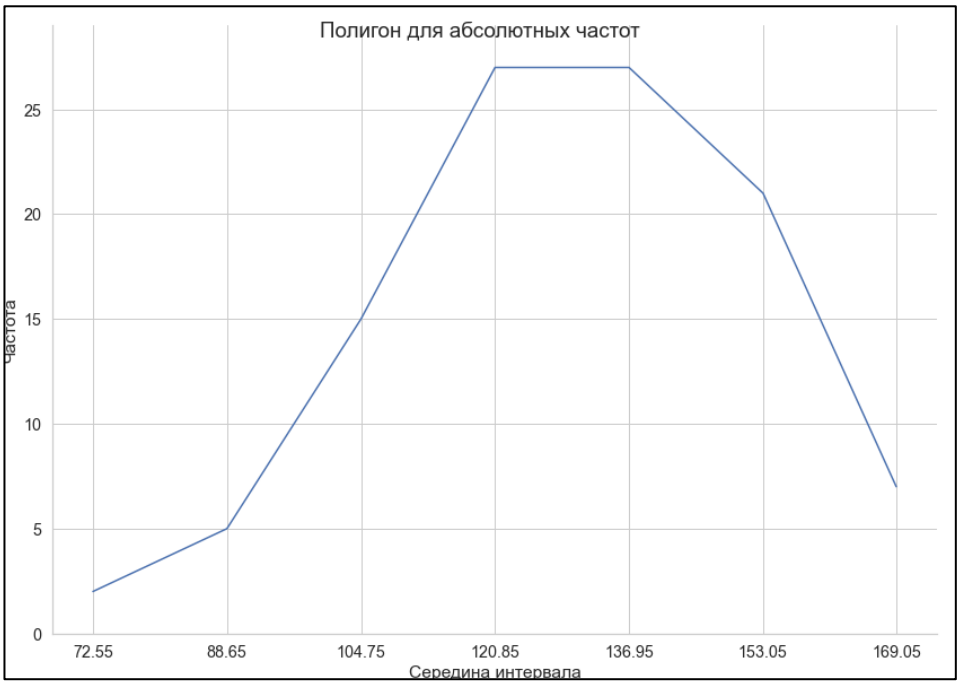

Рисунок 2.2.1 – Полигон для абсолютных частот

Гистограмма, представлена на рис. 2.2.2.

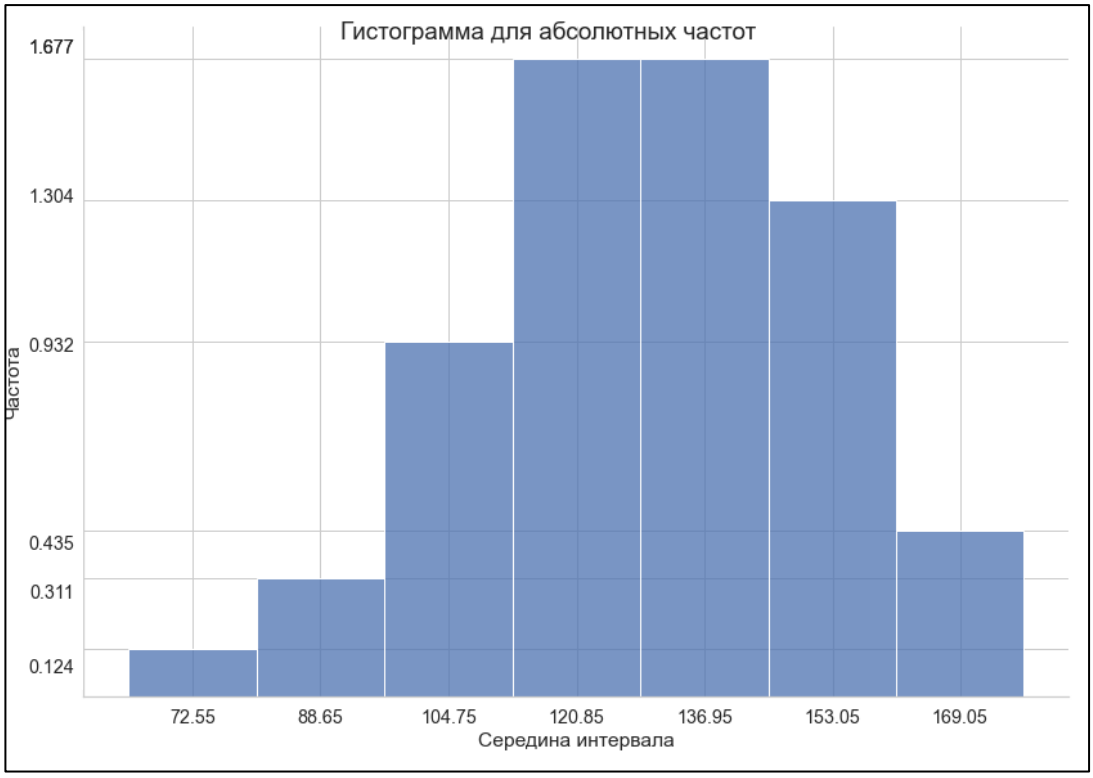

Рисунок 2.2.2 – Гистограмма для абсолютных частот

- Графики для интервального ряда относительных частот

Полигон представлен на рис. 2.2.3.

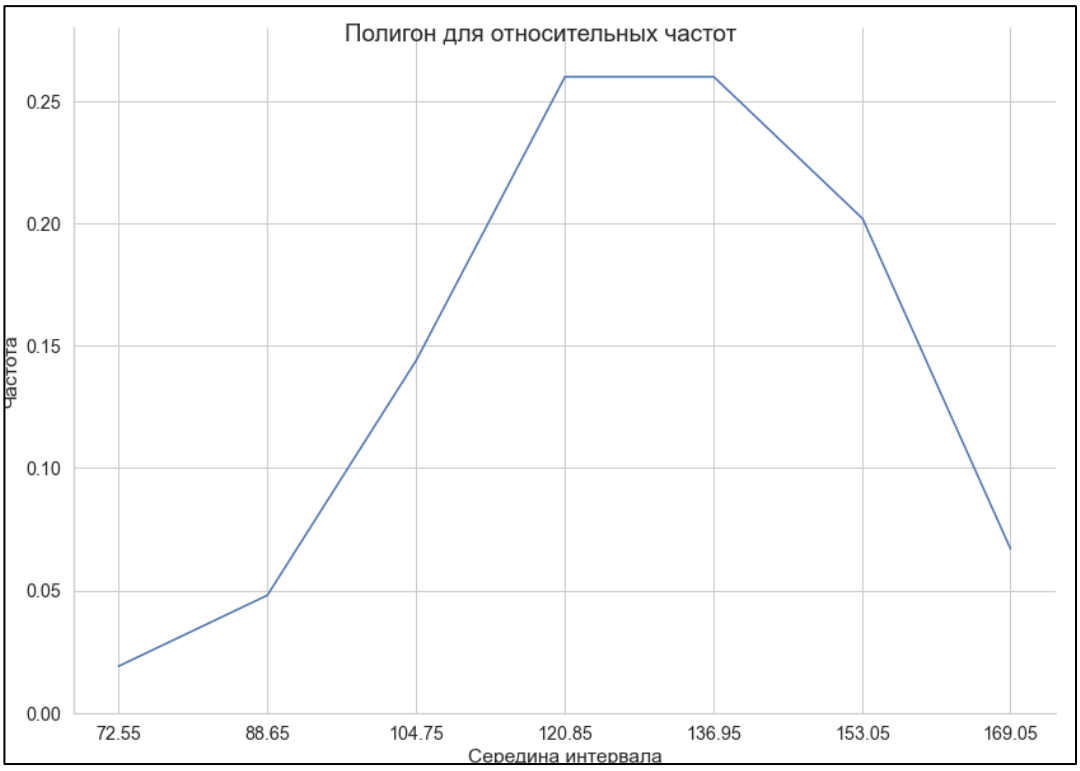

Рисунок 2.2.3 – Полигон для относительных частот

Гистограмма, представлена на рис. 2.2.4.

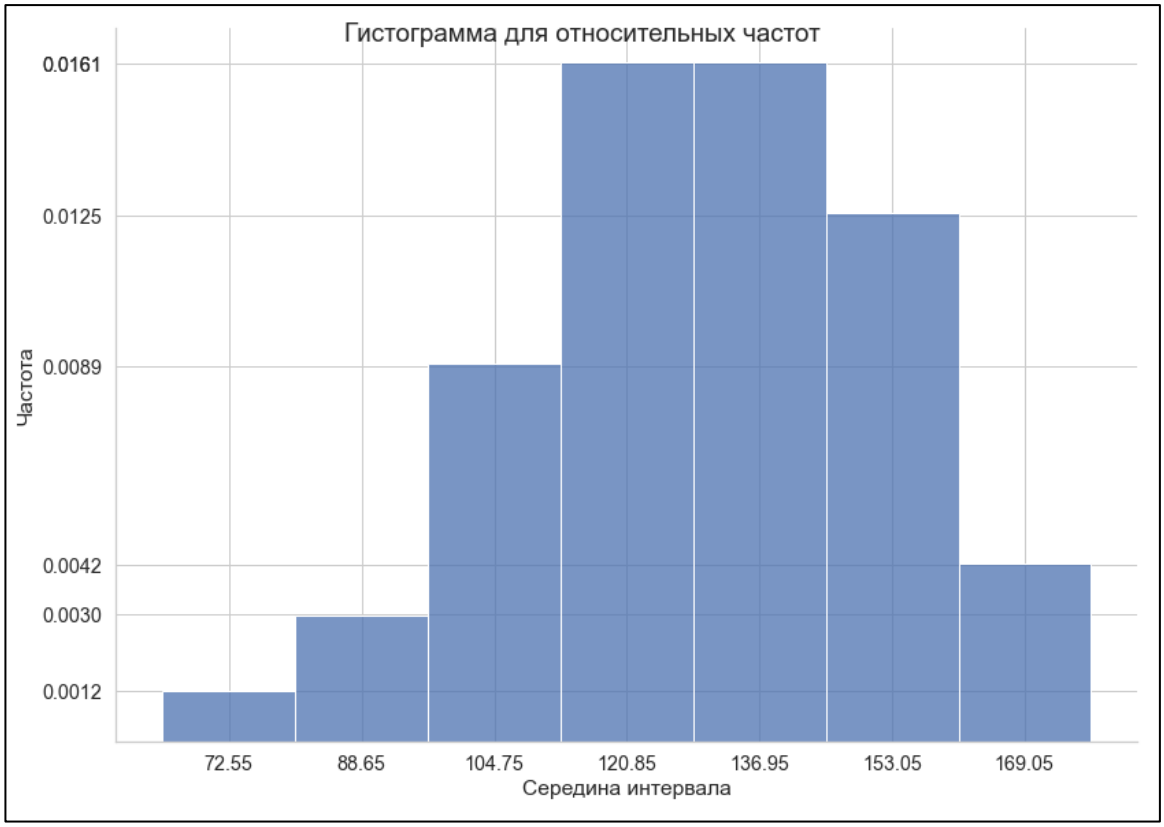

Рисунок 2.2.4 – Гистограмма для относительных частот

Эмпирическая функция распределения, построенная применительно к интервальному ряду для относительных частот представлен на рис. 2.2.5.

Функция распределения:

$$F(x) = \begin{cases} 0, & x = 72.55 \\ 0.019, & x = 88.650 \\ 0.067, & x = 104.75 \\ 0.211, & x = 120.85 \\ 0.471, & x = 136.95 \\ 0.731, & x = 153.05 \\ 0.933, & x = 169.05 \\ 1, & x > 169.05 \end{cases}$$

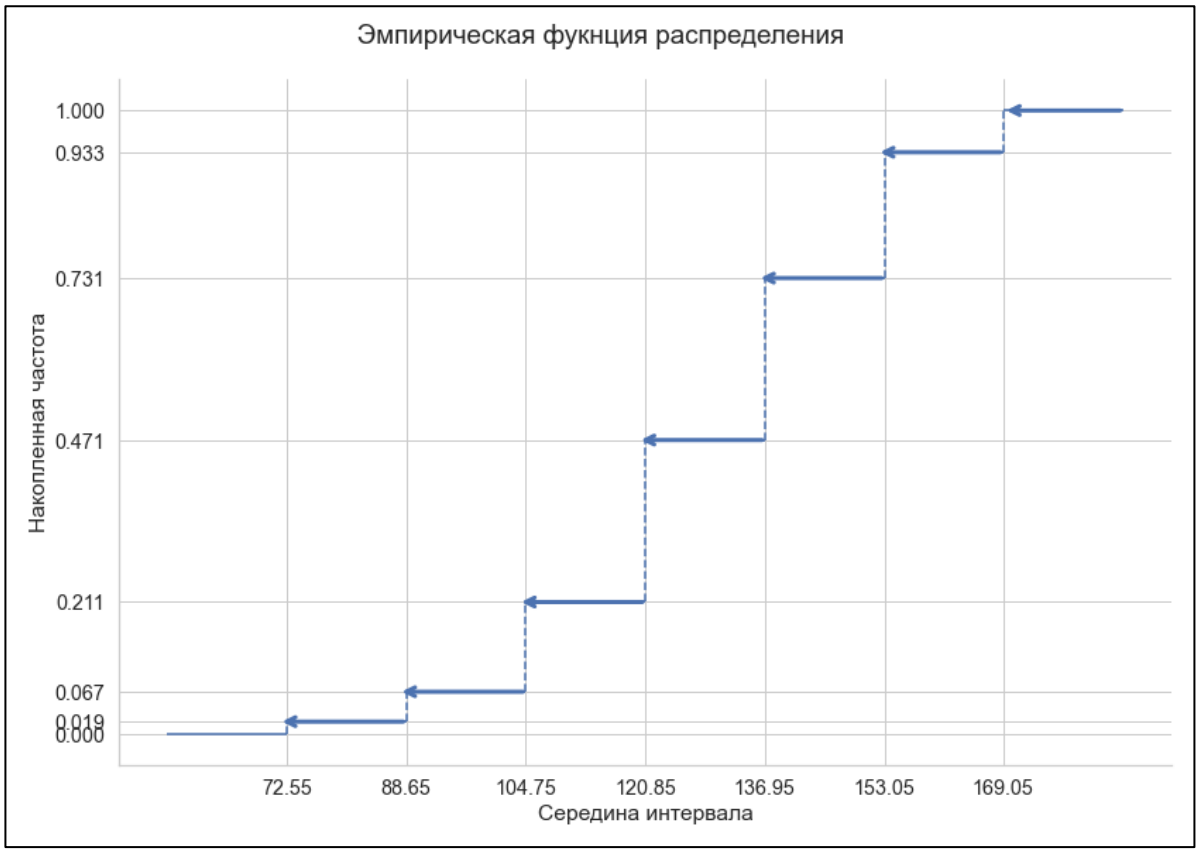

Рисунок 2.2.5 – График эмпирической функции распределения

Для интервального ряда были найдены середины интервалов и накопленные частоты. Интервальный ряд представлен в таблице 16.

Таблица 16

Границы интервалов	***Середины интервалов***	***Абсолютная частота***	***Относительная частота***	***Накопленная частота***
[64.5, 80.6)	72.55	2	0.019	0.019
[80.6, 96.7)	88.65	5	0.048	0.067
[96.7, 112.8)	104.75	15	0.144	0.211
[112.8, 128.9)	120.85	27	0.26	0.471
[128.9, 145.0)	136.95	27	0.26	0.731
[145.0, 161.1)	153.05	21	0.202	0.933
[161.1, 177.0)	169.05	7	0.067	1

Условные варианты можно найти как $u_j = \frac{x_j - C}{h}$, где C – условный ноль.

Условные моменты k-го порядка:

$$\overline{M_k^*} = \frac{1}{N}\sum n_j \left(\frac{x_j - C}{h}\right)^k = \frac{1}{N}\sum n_j u_j^k$$

Результаты вычислений представлены в табл. 17.

Таблица 17

v	n	u	$n*u$	$n*u^2$	$n*u^3$	$n*u^4$	$n*(u+1)^4$
72.55	0.019	-3	-0.057	0.171	-0.513	1.539	0.304
88.65	0.048	-2	-0.096	0.192	-0.384	0.768	0.048
104.75	0.144	-1	-0.144	0.144	-0.144	0.144	0
120.85	0.26	0	0	0	0	0	0.26
136.95	0.26	1	0.26	0.26	0.26	0.26	4.16
153.05	0.202	2	0.404	0.808	1.616	3.232	16.362
169.05	0.067	3	0.201	0.603	1.809	5.427	17.152
$\sum$	1	–	0.568	2.178	2.644	11.37	38.286

Сумма элементов последнего столбца является контрольной суммой, и так как в данном случае во втором столбце записаны относительные частоты, должно быть выполнено равенство:

$$\sum n_j * u_j^4 + 4 * \sum n_j * u_j^3 + 6 * \sum n_j * u_j^2 + 4 * \sum n_j * u_j + 1 = \sum n_j * (u_j + 1)^4$$

$$11.37 + 4 * 2.644 + 6 * 2.178 + 4 * 0.568 + 1 = 38.286$$

Эмпирические начальные и центральные моменты вычислены ниже:

$$\overline{x_в} = \overline{M_1} = \overline{M_1^*}h + C = 129.9948$$

$$D_в = \overline{m_2} = \left(\overline{M_2^*} - \left(\overline{M_1^*}\right)^2\right)h^2 = 480.932$$

$$\overline{m_3} = \left(\overline{M_3^*} - 3\overline{M_2^*}\,\overline{M_1^*} + 2\left(\overline{M_1^*}\right)^3\right)h^3 = -2924.6818$$

$$\overline{m_4} = \left(\overline{M_4^*} - 4\overline{M_3^*}\,\overline{M_1^*} + 6\overline{M_2^*}\left(\overline{M_1^*}\right)^2 + 2\left(\overline{M_1^*}\right)^4\right)h^4 = 622622.816$$

Найдем выборочное среднее и дисперсию с помощью стандартных формул.

Статистическая оценка математического ожидания:

$$\overline{x_в} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = 129.98$$

Статистическая оценка дисперсии:

$$D_{\text{в}} = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x_{\text{в}}})^2\, n_i = 481.8$$

Данная статистическая оценка является смещенной оценкой, поэтому вычислим исправленную оценку дисперсии:

$$s^2 = \frac{N}{N-1} D_{\text{в}} = \frac{104}{103} * 481.8 = 486.5$$

Статистические оценки СКО:

$$\sigma_{\text{в}} = \sqrt{D_{\text{В}}} = \sqrt{481.8} = 21.95$$

$$s = \sqrt{s^2} = \sqrt{486.5} = 22.06$$

Статистические оценки математического ожидания и дисперсии, вычисленные по стандартным формулам и с помощью условных вариант совпадают с небольшой погрешностью.

Статистические оценки коэффициентов асимметрии и эксцесса можно вычислить по формулам:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3}$$

$$\bar{E} = \frac{\overline{m_4}}{s^3} - 3$$

Центральные эмпирические моменты третьего и четвертого порядков были найдены выше.

Статистическая оценка коэффициента асимметрии:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3} = -0.0000254$$

Статистическая оценка коэффициента эксцесса:

$$\overline{\text{E}} = \frac{\overline{m_4}}{s^4} - 3 = -2.99$$

Коэффициент асимметрии отрицателен, следовательно, в данном случае это левосторонняя асимметрия, которая характеризуется удлиненным левым хвостом, а также неравенством $\bar{x}_{\text{в}} < M_o$, но полученное значение незначительно и скос распределения небольшой. Коэффициент эксцесса также отрицателен,

следовательно, эмпирическое распределение является более низким и пологим относительно нормального распределения.

Был построен двумерный интервальный вариационный ряд. Двумерный интервальный ряд представлен в таблице 18.

Таблица 18

Y	***X***							
	338.5	*375.5*	*412.5*	*449.5*	*486.5*	*523.5*	*559*	n_y
72.55	1	1	-	-	-	-	-	2
88.65	3	2	-	-	-	-	-	5
104.75	1	5	8	1	-	-	-	15
120.85	-	-	14	11	2	-	-	27
136.95	-	-	1	12	12	2	-	27
153.05	-	-	-	1	10	8	2	21
169.05	-	-	-	-	-	5	2	7
n_x	5	8	23	25	24	15	4	$n = 104$

При вычислении выборочного коэффициента корреляции необходимо будет посчитать двойную сумму:

$$\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij} y_i x_j = \sum_{i=1}^{K_y} y_i \sum_{j=1}^{K_x} n_{ij} x_j = \sum_{j=1}^{K_x} x_j \sum_{i=1}^{K_y} n_{ij} y_i$$

Данные вычисления были произведены в корреляционной таблице, представленной в таблице 19.

Таблица 19

Y	**X**														X_i	$y_i X_i$
	338.5			*375.5*			*412.5*			*449.5*			*486.5*			*523.5*			*559*				
72.55		338.5			375.5																	714	51800.7
		1			1			-			-			-			-			-			
	72.55			72.55																			
88.65		1015.5			751																	1428	15660 0.225
		3			2			-			-			-			-			-			
	265.95			177.3																			
104.75		338.5			1877.5			3300			449.5											5965.5	62488 6.125
		1			5			8			1			-			-			-			
	104.75			523.75			838			104.75													
120.85							5775			4944.5			973									11692.5	14130 38.625
		-			-			14			11			2			-			-			
							1691.9			1329.35			241.7										
136.95							412.5			5394			5838			1047						12691.5	17381 00.925
		-			-			1			12			12			2			-			
							136.95			1643.4			1643.4			273.9							
153.05										449.5			4865			4188			1118			10620.5	16254 67.525
		-			-			-			1			10			8			2			
										153.05			1530.5			1224.4			306.1				
169.05																2617.5			1118			3735.5	63148 6.275
		-			-			-			-			-			5			2			
																845.25			338.1				
Y_j	443.25			773.6			2666.85			3230.55			3415.6			2343.55			644.2				*6241380.4*
$x_j Y_j$	15040.125			290486.8			1100075.625			1452132.225			1661689.4			1226848.425			360107.8			*6241380.4*	

$$\sum_{i=1}^{K_y} \sum_{j=1}^{K_x} n_{ij} y_i x_j = 6241380.4$$

Исходя из результатов корреляционной таблицы был вычислен выборочный коэффициент корреляции.

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y} \sum_{j=1}^{K_x} n_{ij} y_i x_j - N \bar{x}_в \bar{y}_в}{N S_x S_y} = \frac{6241380.4 - 104 * 453.71 * 129.98}{104 * 53.79 * 22.06} = 0.8765$$

Выборочный коэффициент корреляции отличен от нуля, следовательно X и Y коррелированы. Если случайные величины X и Y коррелированы, то они

зависимы. Также $\bar{r}_{xy}$ – положительный, следовательно можно сказать о положительной корреляционной зависимости, то есть, если X возрастает, то и Y возрастает.

Также выборочный коэффициент корреляции был посчитан с помощью условных вариант. Вычисления представлены в таблице 20.

Таблица 20

Y	**X**								
	-3	*-2*	*-1*	*0*	*1*	*2*	*3*	X_i	y_iX_i
-3	-3 1 -3	-2 1 -3	- 	- 	- 	- 	- 	-5	15
-2	-9 3 -6	-4 2 -4	- 	- 	- 	- 	- 	-13	26
-1	-3 1 -1	-10 5 -5	-8 8 -8	0 1 -1	- 	- 	- 	-21	21
0	- - 	- - 	-14 14 0	0 11 0	2 2 0	- 	- 	-12	0
1	- - 	- - 	1 1 1	0 12 12	12 12 12	4 2 2	- 	15	15
2	- - 	- - 	- 	0 1 2	10 10 20	16 8 16	6 2 4	32	64
3	- - 	- - 	- 	- 	- 	10 5 15	6 2 6	16	48
Y_j	-10	-12	-7	13	32	33	10		*189*
x_jY_j	30	24	7	0	32	66	30	*189*	

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}u_iv_j - N\overline{u_в}\overline{v_в}}{NS_uS_v} = \frac{189 - 104 * 0.115 * 0.567}{104 * 1.45 * 1.364} = 0.8758, \text{где}$$

$u_в = 0.115, v_в = 0.567$ – условные средние для условных вариант,

$S_u = 1.45, S_v = 1.364$ – несмещенные СКО условных вариант

Коэффициенты корреляции, рассчитанные двумя способами, совпадают с точностью до сотых.

В случае нормального распределения системы случайных величин $\{X;Y\}$ для оценки значения r_{xy}, если $\bar{r}_{xy}$ – значим, можно использовать соотношение:

$$\bar{r}_{xy} - 3\frac{1-\bar{r}_{xy}^2}{\sqrt{N}} \le r_{xy} \le \bar{r}_{xy} + 3\frac{1+\bar{r}_{xy}^2}{\sqrt{N}}$$

$$0.8765 - 3\frac{1-0.8765^2}{\sqrt{104}} \le r_{xy} \le 0.8765 + 3\frac{1+0.8765^2}{\sqrt{104}}$$

$$0.8083 \le r_{xy} \le 1$$

С помощью преобразования Фишера перейдём к случайной величине z:

$$\bar{z} = 0.5\ln\frac{1+\bar{r}_{xy}}{1-\bar{r}_{xy}} = 0.5\ln\frac{1+0.8765}{1-0.8765} = 1.36$$

СКО распределения z:

$$\bar{\sigma}_z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{104-3}} = 0.0995$$

Доверительный интервал для генерального значения:

$$(\bar{z} - \lambda(\gamma)\bar{\sigma}_z; \bar{z} + \lambda(\gamma)\bar{\sigma}_z), \text{где}$$

$$\Phi\big(\lambda(\gamma)\big) = \frac{\gamma}{2}$$

Тогда при уровне значимости $\gamma = 0.95$:

$$\Phi\big(\lambda(\gamma)\big) = 0.475$$

$$\lambda(\gamma) = 1.96$$

$$z \in (1.36 - 1.96 * 0.0995; 1.36 + 1.96 * 0.0995)$$

$$z \in (1.165; 1.555)$$

Для пересчёта интервала в доверительный интервал для коэффициента корреляции с тем же значением γ необходимо воспользоваться обратным преобразованием Фишера:

$$r = th(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z} - 1}{e^{2z} + 1}$$

$$\frac{e^{2z_1}-1}{e^{2z_1}+1}=0.8227$$
$$\frac{e^{2z_2}-1}{e^{2z_2}+1}=0.9146$$

Можно сделать вывод, что интервал $(0.8227; 0.9146)$ с вероятностью (надежностью) $\gamma = 0.95$ содержит в себе истинное значение коэффициента корреляции.

Поскольку $\bar{r}_{xy}$ является случайной величиной, то это еще не значит, что r_{xy} тоже отличен от нуля. Проверим гипотезу $H_0: r_{xy} = 0$. Альтернативой будет гипотеза $H_1: r_{xy} \neq 0$.

В качестве критерия проверки гипотезы можно принять случайную величину:

$$T = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}}$$

При справедливости нулевой гипотезы случайная величина T распределена по закону Стьюдента с $k = K - 2 = 5$ степенями свободы.

Найдено $T_{\text{набл}}$ по формуле выше:

$$T_{\text{набл}} = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}} = \frac{0.8765 * \sqrt{102}}{\sqrt{1-0.8765^2}} = 18.388$$

По заданному уровню значимости $\alpha = 0.05$ и значению $k = N - 2 = 102$ из таблицы было определено значение $t_{\text{крит}} = 1.985$

$$T_{\text{набл}} = 18.388$$
$$t_{\text{крит}} = 1.985$$

$|T_{\text{набл}}| > t_{\text{крит}}(\alpha, k)$ – основная гипотеза H_0 должна быть отвергнута, это означает, что выборочный коэффициент корреляции $\bar{r}_{xy}$ значимо отличается от нуля (значим).

2.3. Элементы регрессионного анализа. Выборочные прямые. среднеквадратической регрессии. Корреляционные отношения.

Для заданной выборки построим уравнения средней квадратичной регрессии X на Y и Y на X и отобразим полученные прямые на множестве выборки.

Выборочная прямая средней квадратичной регрессии X на Y:

$$\overline{x_y} = \bar{x}_в + \overline{r_{xy}} \frac{S_x}{S_y}(y - \bar{y}_в)$$

$$x(y) = 453.71 + 0.8765\frac{53.79}{22.06}(y - 129.98) = 2.1372 * y + 175.915$$

Выборочная прямая средней квадратичной регрессии Y на X:

$$\overline{y_x} = \bar{y}_в + \overline{r_{xy}} \frac{S_y}{S_x}(x - \bar{x}_в)$$

$$y(x) = 129.98 + 0.8765\frac{22.06}{53.79}(x - 453.71) = 0.3595 * x - 33.1126$$

Двумерная выборка и выборочные прямые средней квадратичной регрессии представлены на рис. 2.3.1.

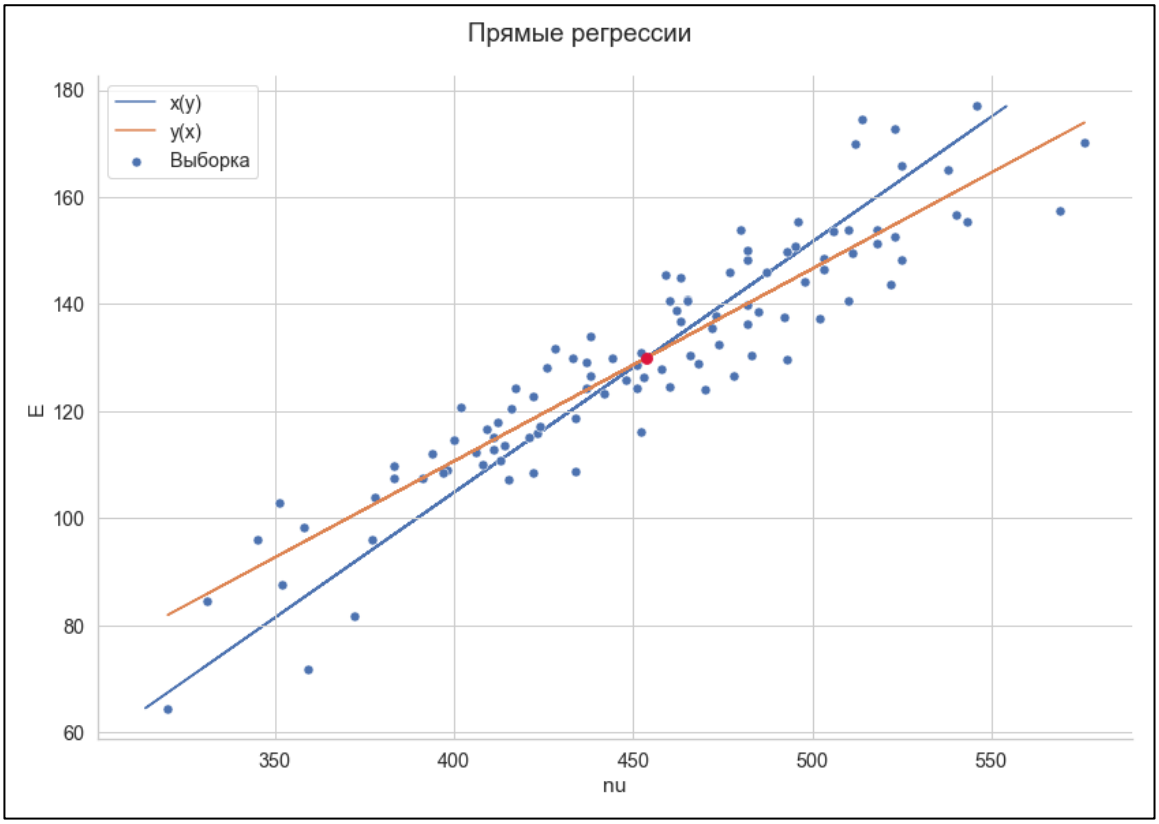

Рисунок 2.3.1 - Выборочные прямые средней квадратичной регрессии

Можно заметить, что пересечение выборочных прямых средней квадратичной регрессии находится в точке с координатами выборочного среднего для каждого из признаков.

Статистические оценки остаточной дисперсии для полученных выборочных прямых регрессии:

$$D_{\text{ост}_x} = S_x^2\left(1 - \overline{r_{xy}}^2\right) = 670.53$$

$$D_{\text{ост}_y} = S_y^2\left(1 - \overline{r_{xy}}^2\right) = 112.78$$

Корреляционная таблица для нахождения выборочного корреляционного отношения представлена в таблице 21. В данной таблице рассчитаны групповые выборочные средние и групповые выборочные дисперсии.

Таблица 21 - Корреляционная таблица

Y	***X***									
	338.5	*375.5*	*412.5*	*449.5*	*486.5*	*523.5*	*559*	n_{y_i}	$\bar{x}_{y_i}$	$D_{x_{y_i}}$
72.55	1	1	-	-	-	-	-	2	357	342.25
88.65	3	2	-	-	-	-	-	5	353.3	328.56
104.75	1	5	8	1	-	-	-	15	397.7	693.63
120.85	-	-	14	11	2	-	-	27	433.06	536.99
136.95	-	-	1	12	12	2	-	27	479.06	638.07
153.05	-	-	-	1	10	8	2	21	505.74	715.28
169.05	-	-	-	-	-	5	2	7	533.64	260.24
n_{x_j}	5	8	23	25	24	15	4	104	-	-
$\bar{y}_{x_j}$	88.65	96.7	115.95	129.22	142.32	156.24	161.05	-	-	-
$D_{y_{x_j}}$	103.68	129.6	77.42	106.69	99.86	108.7	64	-	-	-

Выборочное корреляционное отношение X к Y рассчитывается как отношение выборочных значений СКО $\overline{x_y}$ и X соответсвенно. Для этого были вычислены внутригрупповая, межгрупповая и общая дисперсии.

$$D_{\text{внгр}_{xy}} = \frac{1}{n}\sum_{1}^{K} D_{x_{y_i}} * n_{y_i} = 589.432$$

$$D_{\text{межгр}_{xy}} = \frac{1}{n}\sum_{1}^{K} \left(\bar{x}_{y_i} - \bar{x}_{\text{в}}\right)^2 * n_{y_i} = 2273.804$$

$$D_{\text{общ}_{xy}} = D_{\text{внгр}_{xy}} + D_{\text{межгр}_{xy}} = 2863.236$$

$$\overline{\eta_{xy}} = \sqrt{\frac{D_{\text{межгр}_{xy}}}{D_{\text{общ}_{xy}}}} = 0.8911$$

$$\overline{r_{xy}} = 0.8765$$

Неравенство $\overline{\eta_{xy}} \geq \left|\overline{r_{xy}}\right|$ выполняется.

Выборочное корреляционное отношение Y к X:

$$D_{\text{внгр}_{yx}} = \frac{1}{n}\sum_{1}^{K} D_{y_{x_j}} * n_{x_j} = 98.91$$

$$D_{\text{межгр}_{yx}} = \frac{1}{n}\sum_{1}^{K} \left(\bar{y}_{x_j} - \bar{y}_{\text{в}}\right)^2 * n_{x_j} = 382.72$$

$$D_{\text{общ}_{yx}} = D_{\text{внгр}_{yx}} + D_{\text{межгр}_{yx}} = 481.63$$

$$\overline{\eta_{yx}} = \sqrt{\frac{D_{\text{межгр}_{yx}}}{D_{\text{общ}_{yx}}}} = 0.8914$$

$$\overline{r_{xy}} = 0.8765$$

Неравенство $\overline{\eta_{yx}} \geq \left|\overline{r_{xy}}\right|$ выполняется.

Для заданной выборки построим корреляционную кривую параболического вида $y = \beta_2 x^2 + \beta_1 x^2 + \beta_0$. Выборочное уравнение регрессии Y на X:

$$\overline{y_x} = ax^2 + bx + c$$

Значения коэффициентов определим с помощью МНК, решив систему уравнений:

$$\begin{cases} \left(\sum_{i=1}^{m} n_{x_i} x_i^4\right) a + \left(\sum_{i=1}^{m} n_{x_i} x_i^3\right) b + \left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) c = \sum_{i=1}^{m} n_{x_i} \overline{y_{x_i}} x_i^2 \\ \left(\sum_{i=1}^{m} n_{x_i} x_i^3\right) a + \left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) b + \left(\sum_{i=1}^{m} n_{x_i} x_i\right) c = \sum_{i=1}^{m} n_{x_i} \overline{y_{x_i}} x_i \\ \left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) a + \left(\sum_{i=1}^{m} n_{x_i} x_i\right) b + Nc = \sum_{i=1}^{m} n_{x_i} \overline{y_{x_i}} \end{cases}$$

Для вычисления сумм была построена таблица 22.

Таблица 22 – Таблица сумм МНК

x	n_x	$\overline{y_x}$	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \overline{y_x}$	$n_x \overline{y_x} x$	$n_x \overline{y_x} x^2$
338.5	5.0	88.65	1692.5	572911.25	193930458.125	65645460075.3125	443.25	150040.125	50788582.3125
375.5	8.0	96.7	3004.0	1128002.0	423564751.0	159048564000.5	773.6	290486.8	109077793.4
412.5	23.0	115.95	9487.5	3913593.75	1614357421.875	665922436523.4375	2666.85	1100075.625	453781195.3125
449.5	25.0	129.22	11237.5	5051256.25	2270539684.375	1020607588126.5625	3230.5	1452109.75	652723332.625
486.5	24.0	142.32	11676.0	5680374.0	2763501951.0	1344443699161.5	3415.68	1661728.3199999998	808430827.68
523.5	15.0	156.24	7852.5	4110783.75	2151995293.125	1126569535950.9375	2343.6	1226874.6	642268853.1
559.0	4.0	161.05	2236.0	1249924.0	698707516.0	390577501444.0	644.2	360107.80000000005	201300260.20000002
∑	104.0	-	47186.0	21706845.0	10116597075.5	4772814785282.25	13517.68	6241423.019999999	2918370844.629996

В результате решения системы были получены следующие значения коэффициентов:

$$a = -0.00033$$

$$b = 0.6563$$

$$c = -99.7989$$

Выборочное уравнение регрессии Y на X:

$$y(x) = -0.00033 * x^2 + 0.6563 * x - 99.7989$$

Корреляционная кривая параболического вида представлена на рис. 2.3.2.

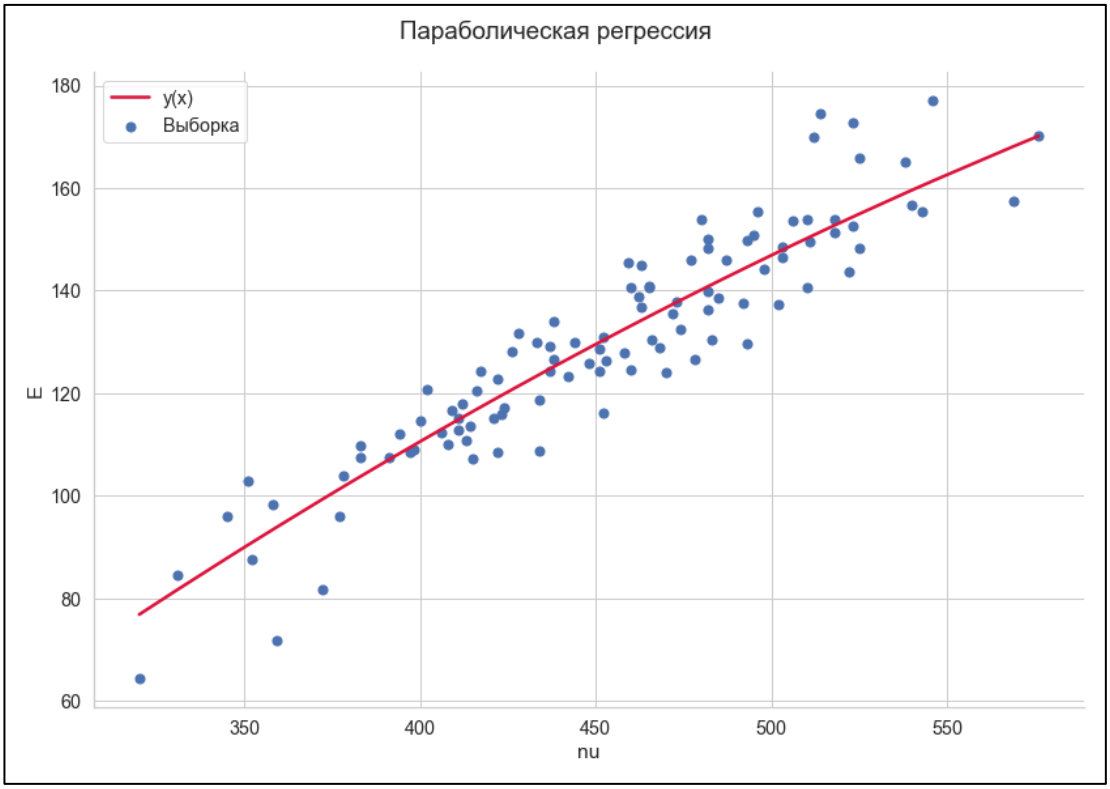

Рисунок 2.3.2 – Корреляционная кривая параболического вида

Для заданной выборки построим корреляционную кривую степенной функции $y = \beta_0 x^{\beta_1}$. Выборочное уравнение регрессии Y на X:

$$\overline{y_x} = a * x^b$$

Значения коэффициентов определим с помощью МНК, решив систему уравнений:

$$\begin{cases} \ln a + \left(\sum_{i=1}^{m} \ln x_i\right) b = \sum_{i=1}^{m} \ln \overline{y_{x_i}} \\ \left(\sum_{i=1}^{m} \ln x_i\right) \ln a + \left(\sum_{i=1}^{m} (\ln x_i)^2\right) b = \sum_{i=1}^{m} \ln x_i * \ln \overline{y_{x_i}} \end{cases}$$

Для вычисления сумм была построена таблица 23.

Таблица 23 – Таблица сумм МНК

x	n_x	$\overline{y_x}$	$\ln x$	$(\ln x)^2$	$\ln y$	$\ln x * \ln y$
338.5	5.0	88.65	5.825	33.925	4.485	26.121
375.5	8.0	96.7	5.928	35.144	4.572	27.102
412.5	23.0	115.95	6.022	36.267	4.753	28.625
449.5	25.0	129.22	6.108	37.309	4.862	29.695
486.5	24.0	142.32	6.187	38.282	4.958	30.677
523.5	15.0	156.24	6.261	39.194	5.051	31.624
559.0	4.0	161.05	6.326	40.02	5.082	32.148
∑	104.0	-	42.657	260.142	33.762	205.991

В результате решения системы были получены следующие значения коэффициентов:

$$\ln a = 0.002575 \Rightarrow a = 1.00258$$

$$b = 0.7914$$

Выборочное уравнение регрессии Y на X:

$$y(x) = 1.00258 * x^{0.7914}$$

Корреляционная кривая степенной функции представлена на рис. 2.3.3.

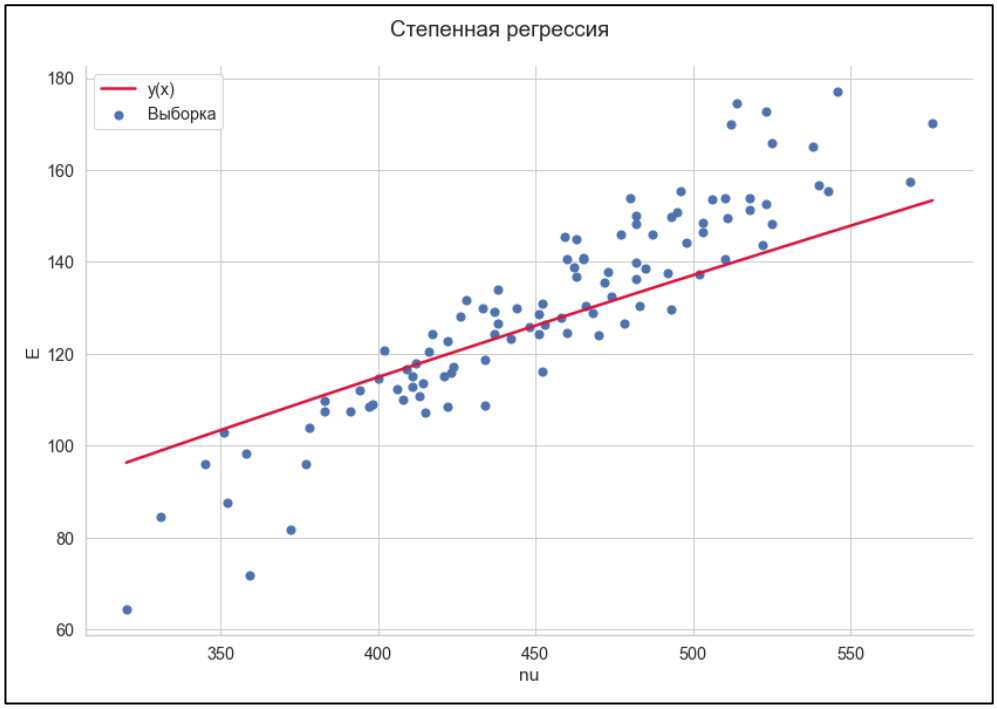

Рисунок 2.3.3 – Корреляционная кривая степенной функции

2.4. Выводы.

Построен двумерный интервальный вариационный ряд. С его помощью была построена корреляционная таблица, где была вычислена двойная сумма для

выборочного коэффициента корреляции. Исходя из результатов корреляционной таблицы был вычислен выборочный коэффициент корреляции $\overline{r_{xy}} = 0.8765$. Выборочный коэффициент корреляции отличен от нуля, следовательно X и Y коррелированы и зависимы. Также $\bar{r}_{xy}$ – положительный, следовательно можно сказать о положительной корреляционной зависимости.

Построен доверительный интервал для коэффициента корреляции при уровне значимости $\gamma = 0.95$. Можно сделать вывод, что интервал $(0.8227; 0.9146)$ с вероятностью (надежностью) $\gamma = 0.95$ содержит в себе истинное значение коэффициента корреляции.

Проведена проверка статистической гипотезы о равенстве коэффициента корреляции нулю при уровне значимости $\alpha = 0.05$. Было выяснено, что $|T_{\text{набл}}| > t_{\text{крит}}(\alpha, k)$, то есть основная гипотеза H_0 должна быть отвергнута, это означает, что выборочный коэффициент корреляции $\bar{r}_{xy}$ значимо отличается от нуля (значим).

Для заданной выборки были получены выборочные прямые средней квадратичной регрессии X на Y и Y на X.

$$x(y) = 2.1372 * y + 175.915$$

$$y(x) = 0.3595 * x - 33.1126$$

Прямые были построены на множестве выборки.

Найдены выборочные корреляционные отношения $\overline{\eta_{xy}} = 0.8911$ и $\overline{\eta_{yx}} = 0.8194$. Определено, что выполняются неравенства $\overline{\eta_{xy}} \geq \left|\overline{r_{xy}}\right|$ и $\overline{\eta_{yx}} \geq \left|\overline{r_{xy}}\right|$. На основе полученных значений выборочного корреляционного отношения можно предположить, что X и Y связаны корреляционной зависимостью, но не линейной корреляционной зависимостью и не функциональной зависимостью. Характер корреляционной зависимости не определен.

Были построены корреляционные кривые параболического и степенного вида. Визуально можно сделать вывод о том, что корреляционная зависимость признаков может быть выражена параболической функцией, но в меньшей мере степенной.

3. КЛАСТЕРНЫЙ АНАЛИЗ

3.1. Основные теоретические положения

Задача кластерного анализа заключается в том, чтобы на основании данных, характеризующих исследуемые объекты, разбить множество объектов G на m кластеров (подмножеств G) $G_1, G_2, \dots, G_m$ таких, что:

$$G_1 \subset G; G_2 \subset G; \dots; G_m \subset G$$

$$G_1 \cup G_2 \cup \dots \cup G_m = G$$

$$G_i \cap G_j = \emptyset \; \forall i \neq j$$

К характеристикам кластера относятся:

- Центр кластера – это среднее геометрическое место точек, принадлежащих кластеру, в пространстве данных.
- Радиус кластера – максимальное расстояние точек, принадлежащих кластеру, от центра кластера.
- Кластеры могут быть перекрывающимися. В этом случае невозможно при помощи используемых процедур однозначно отнести объект к одному из двух или более кластеров. Такие объекты называют спорными.
- Спорный объект – это объект, который по мере сходства может быть отнесен к более, чем одному кластеру.
- Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

Большое значение в кластерном анализе имеет выбор масштаба. Обычно требуется нормировка переменных. Существуют различные способы нормировки данных:

$$z = \frac{(x - \bar{x})}{\sigma}; z = \frac{x}{\bar{x}}; z = \frac{x}{x_{max}}; z = \frac{(x - \bar{x})}{x_{max} - x_{min}}$$

Расстоянием (метрикой) между объектами a и b пространстве параметров называется такая величина d_{ab}, которая удовлетворяет аксиомам:

1. $d_{ab} > 0$, *если* $a \neq b$, 2. $d_{ab} = 0$, *если* $a = b$;
3. $d_{ab} = d_{ba}$; 4. $d_{ab} + d_{bc} \geq d_{ac}$.

Мерой близости (сходства) называется величина μ_{ab}, имеющая предел и возрастающая с возрастанием близости объектов и удовлетворяющая условиям:

$$\mu_{ab} \text{ непрерывна; } \mu_{ab} = \mu_{ba}; 0 \leq \mu_{ab} \leq 1.$$

Существует возможность простого перехода от расстояния к мерам близости:

$$\mu = \frac{1}{1+d}.$$

Суть метода k-средних заключается в том, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \bar{x}\right)^2$$

Центроиды выбираются в тех местах, где визуально скопление точек выше. Алгоритм разбивает множество элементов векторного пространства на заранее известное число кластеров k. Основная идея заключается в том, что на каждой итерации пересчитывается центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения центра масс кластеров.

Возможны две разновидности метода k -средних.

Первая предполагает пересчет центра кластера после каждого изменения его состава, а вторая – лишь после завершения цикла.

Перед началом работы метода целесообразно нормировать характеристики объектов:

$$\widehat{X} = \frac{x - \bar{x}_в}{S_x}; \widehat{Y} = \frac{y - \bar{y}_в}{S_y}$$

Задание количества кластеров является сложным вопросом. Если нет разумных соображений на этот счет, рекомендуется первоначально создать 2 кластера, затем 3, 4, 5 и так далее, сравнивая полученные результаты.

После завершения многомерной классификации необходимо оценить полученные результаты. Для этой цели используются специальные характеристики – функционалы качества. Наилучшим разбиением считается такое, при котором достигается экстремальное (минимальное или максимальное) значение выбранного функционала качества.

В качестве таких функционалов могут быть использованы:

1. Сумма квадратов расстояний до центров кластеров

$$F_1 = \sum_{k=1}^{K} \sum_{i=1}^{N_k} d^2(X_i^{(k)}, X^{(k)}) \Rightarrow \min$$

2. Сумма внутрикластерных расстояний между объектами

$$F_2 = \sum_{k=1}^{K} \sum_{X_i, X_j \in S_k} d^2(X_i, X_j) \Rightarrow \min$$

3. Сумма внутрикластерных дисперсий

$$F_3 = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sigma_{ij}^2 \Rightarrow \min$$

Здесь σ - дисперсия *j*-й переменной в *k*-м кластере.

Оптимальным следует считать разбиение, при котором сумма внутрикластерных (внутригрупповых) дисперсий будет минимальной.

Судить о качестве разбиения позволяют и некоторые простейшие приемы. Например, можно сравнивать средние значения признаков в отдельных кластерах (группах) со средними значениями в целом по всей совокупности объектов. Если групповые средние существенно отличаются от общего среднего значения, то это может являться признаком хорошего разбиения.

Метод поиска сгущений является еще одним итеративным методом кластерного анализа.

Основная идея метода заключается в построении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков в поисках локальных сгущений объектов.

Метод поиска сгущений требует, прежде всего, вычисления матрицы расстояний (или матрицы мер сходства) между объектами и выбора первоначального центра сферы.

На первом шаге центром сферы служит объект, в ближайшей окрестности которого расположено наибольшее число соседей. На основе заданного радиуса сферы (R) определяется совокупность точек внутри этой сферы, и для них вычисляются координаты центра.

Когда очередной пересчет координат центра сферы приводит к такому же результату, как и на предыдущем шаге, перемещение сферы прекращается, а точки, попавшие в нее, образуют кластер, и из дальнейшего процесса кластеризации исключаются.

Перечисленные процедуры повторяются для всех оставшихся точек. Работа алгоритма завершается за конечное число шагов, и все точки оказываются распределенными по кластерам. Число образовавшихся кластеров заранее неизвестно и сильно зависит от заданного радиуса сферы.

Для оценки устойчивости полученного разбиения целесообразно повторить процесс кластеризации несколько раз для различных значений радиуса сферы, изменяя каждый раз радиус на небольшую величину.

Существуют различные способы выбора начального радиуса сферы. В частности, если обозначить через d_{ij} расстояние между i-м и j-м объектами, то в качестве нижней границы значения радиуса сферы можно выбрать минимальное из таких расстояний, а в качестве верхней границы - максимальное:

$$R_{min} = \min_{i,j} d_{ij}$$

$$R_{max} = \max_{i,j} d_{ij}$$

Тогда, если начинать работу алгоритма с

$$R = R_{min} + \delta;\ \delta > 0$$

и при каждом его повторении увеличивать значение δ на некоторую величину, то в конечном итоге можно найти значения радиусов, которые приводят к устойчивому разбиению на кластеры.

После завершения многомерной классификации необходимо оценить полученные результаты. Для этой цели используются специальные характеристики – функционалы качества. Наилучшим разбиением считается такое, при котором достигается экстремальное (минимальное или максимальное) значение выбранного функционала качества.

3.2. Метод k-средних.

Первоначальная выборка представлена на рис. 3.2.1.

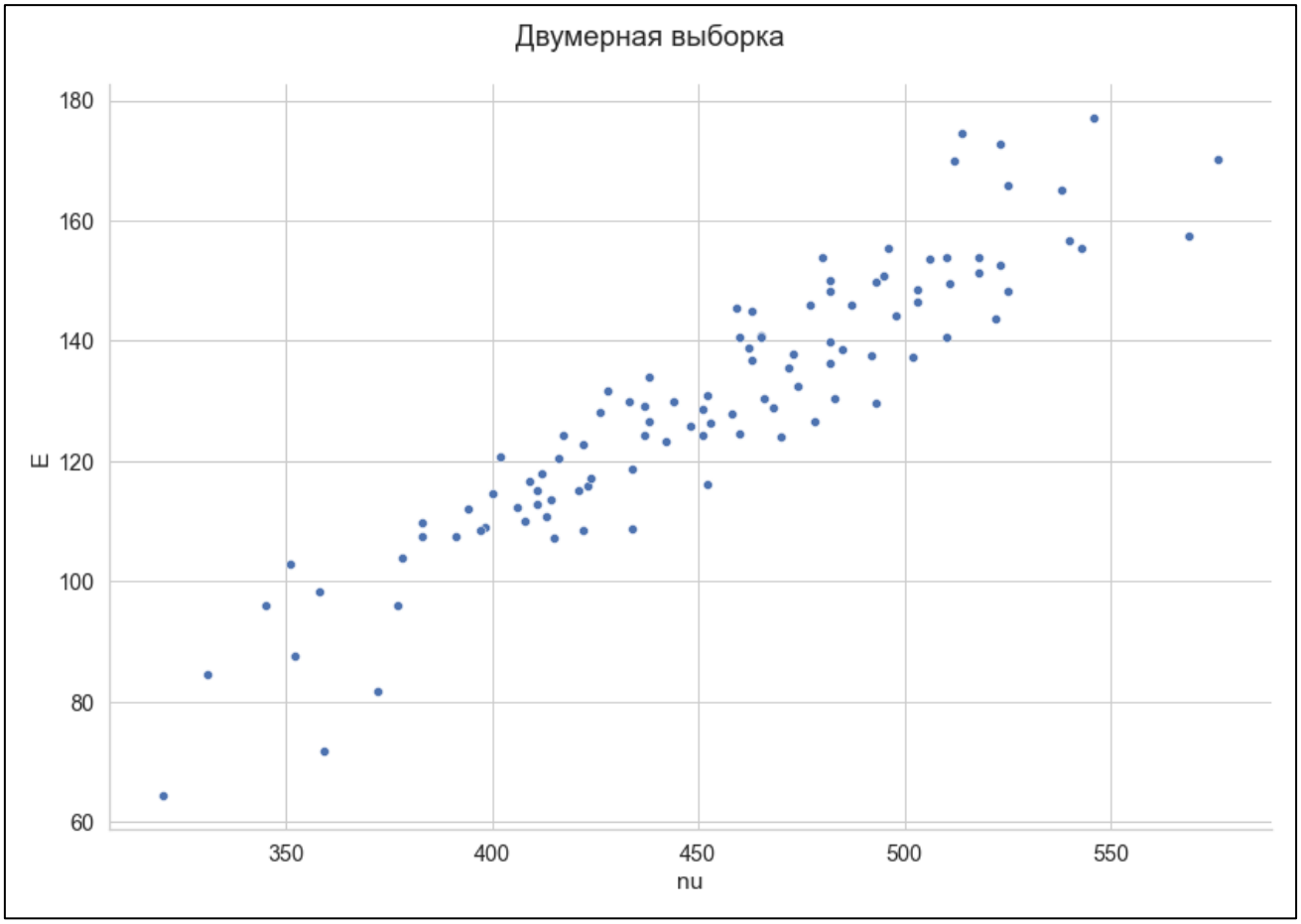

Рисунок 3.2.1 – Первоначальная выборка

Нормализуем множество точек как:

$$z_x = \frac{x - \bar{x_в}}{s_x}; z_y = \frac{y - \bar{y_в}}{s_y}$$

Тогда среднее значение будет равно нулю, а стандартное отклонение единице. Нормализованная выборка представлена на рис. 3.2.2.

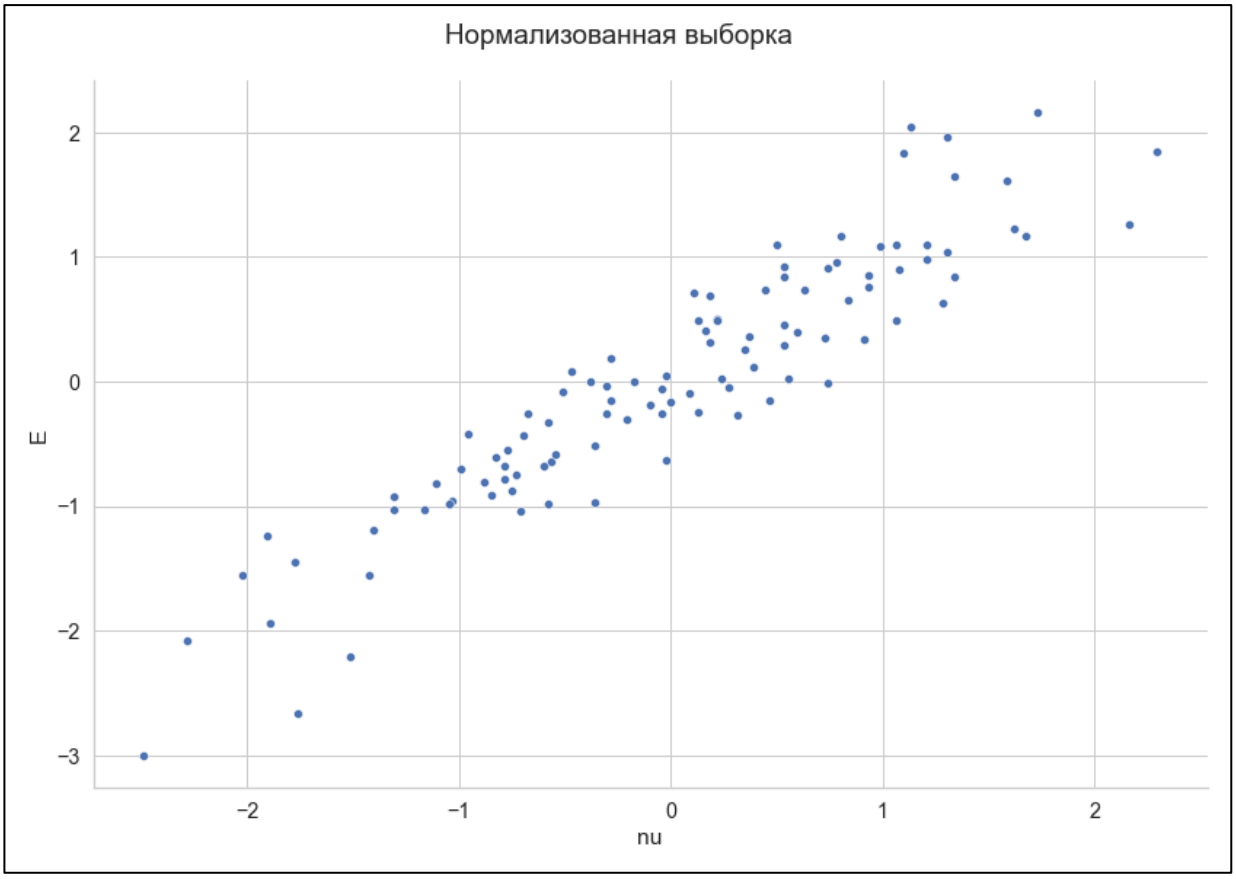

Рисунок 3.2.2 – Нормализованная выборка

Грубая верхняя оценка количества кластеров:

$$\bar{k} = \lfloor \sqrt{N/2} \rfloor = \lfloor \sqrt{104/2} \rfloor = 7$$

- Пересчет центра после шага процедуры

Реализован алгоритм k-means, где пересчет центра кластера осуществляется после шага процедуры (после просмотра всех данных). Количество кластеров от 2 до 7. Полученные кластеры были отображены, выделены свои цветом, были отмечены центроиды. На каждом шаге процедуры вычислены функционалы качества полученного разбиения:

- F_1 – сумма по всем кластерам квадратов расстояний элементов кластеров до центров соответствующих кластеров
- F_2 – сумма по всем кластерам внутрикластерных расстояний между элементами кластера

- F_3 – сумма по всем кластерам внутрикластерных дисперсий

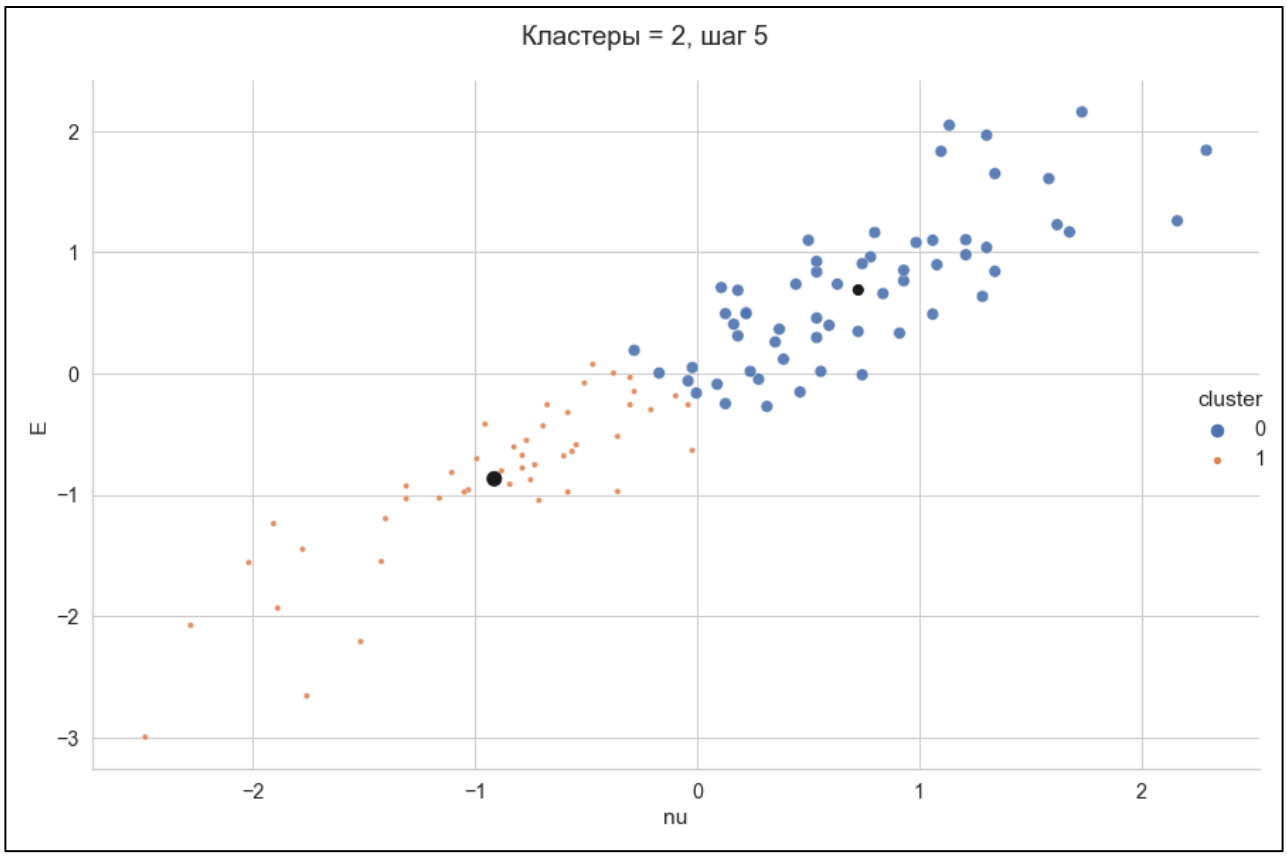

Рисунок 3.2.3 – Кластеризация методом k-средних (2 кластера)

Таблица 24

Центр кластера	Количество элементов
(0.7249; 0.6889)	58
(-0.914; -0.8687)	46

Таблица 25

F_1	F_2	F_3
79.619	4555.727	1.534
78.349	4307.859	1.531
77.567	4134.173	1.53
76.855	4016.464	1.52
76.855	4016.464	1.52

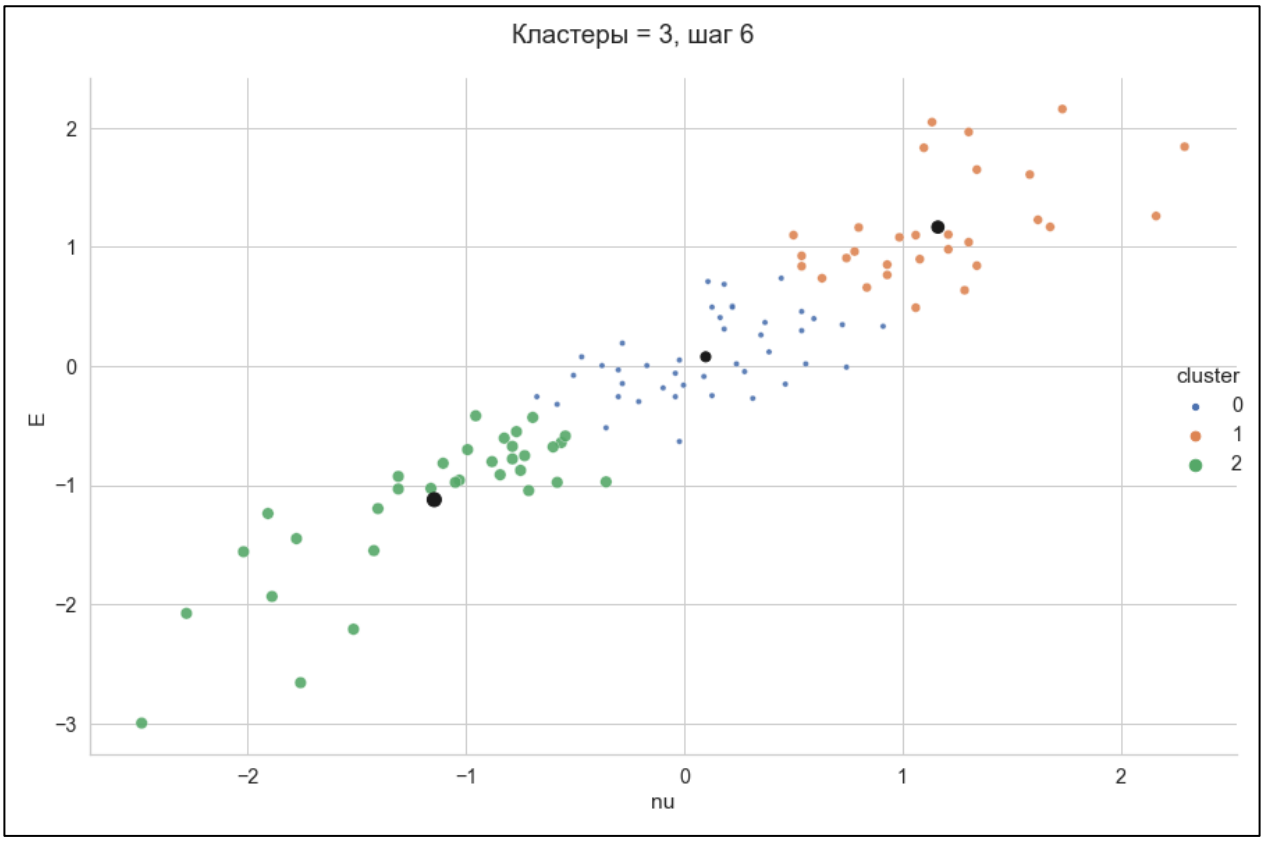

Рисунок 3.2.4 – Кластеризация методом k-средних (3 кластера)

Таблица 26

Центр кластера	Количество элементов
(0.098; 0.0769)	42
(1.162; 1.166)	29
(-1.1459; -1.1225)	33

Таблица 27

F_1	F_2	F_3
51.075	2185.045	1.466
46.297	1719.227	1.385
45.085	1599.229	1.356
44.544	1550.091	1.351
43.857	1497.05	1.349
43.857	1497.05	1.349

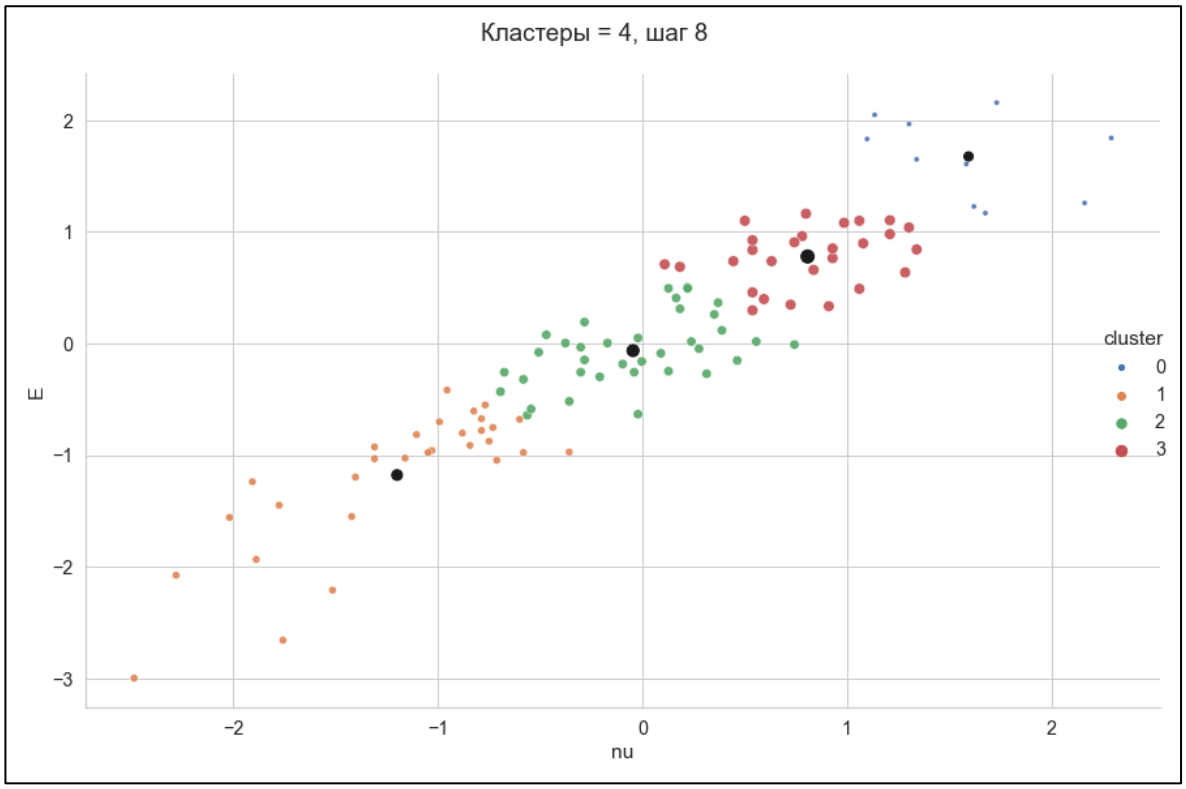

Рисунок 3.2.5 – Кластеризация методом k-средних (4 кластера)

Таблица 28

Центр кластера	Количество элементов
(1.5941; 1.6751)	10
(-1.2003; -1.1793)	30
(-0.0467; -0.0648)	37
(0.8072; 0.7787)	27

Таблица 29

F_1	F_2	F_3
61.594	2888.433	1.784
54.657	2272.014	1.688
46.384	1750.805	1.48
37.374	1186.478	1.354
35.815	1085.379	1.357
35.551	1058.432	1.379
35.379	1053.881	1.381
35.379	1053.881	1.381

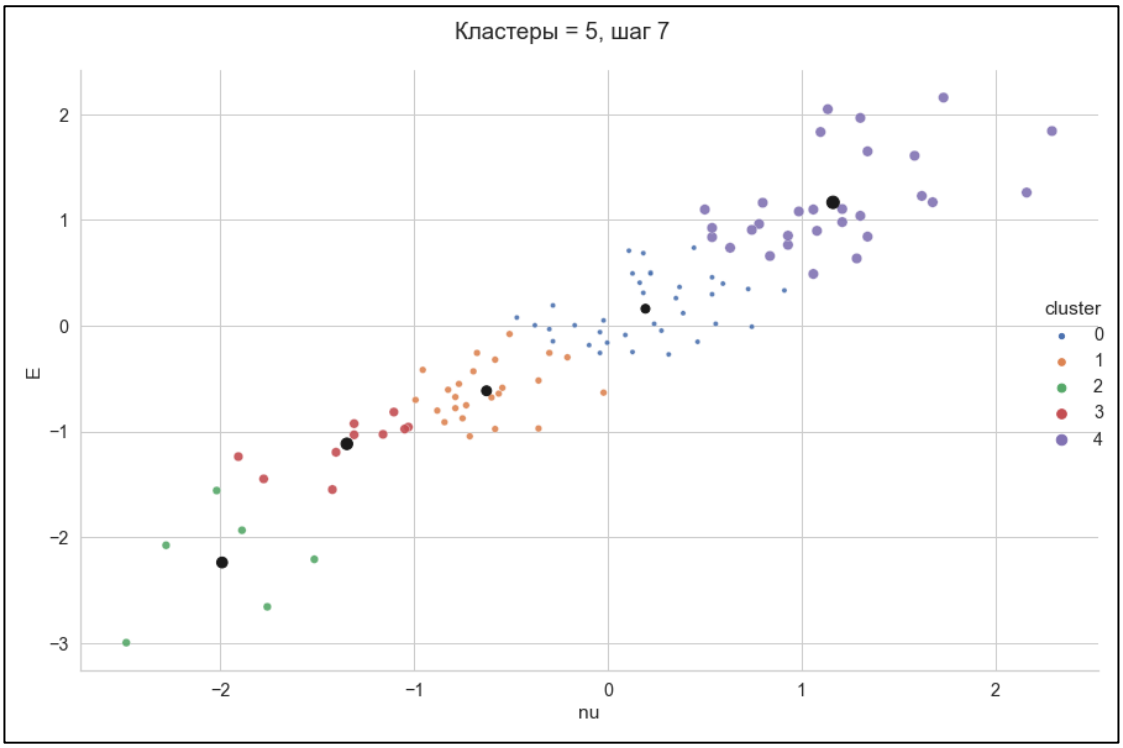

Рисунок 3.2.6 – Кластеризация методом k-средних (5 кластеров)

Таблица 30

Центр кластера	Количество элементов
(0.1936; 0.16)	35
(-0.6268; -0.6164)	24
(-1.9918; -2.2403)	6
(-1.3478; -1.1179)	10
(1.162; 1.166)	29

Таблица 31

F_1	F_2	F_3
32.94	1326.678	1.284
27.323	859.938	1.269
25.084	733.119	1.244
24.407	681.996	1.254
24.16	664.971	1.257
24.084	654.669	1.259
24.084	654.669	1.259

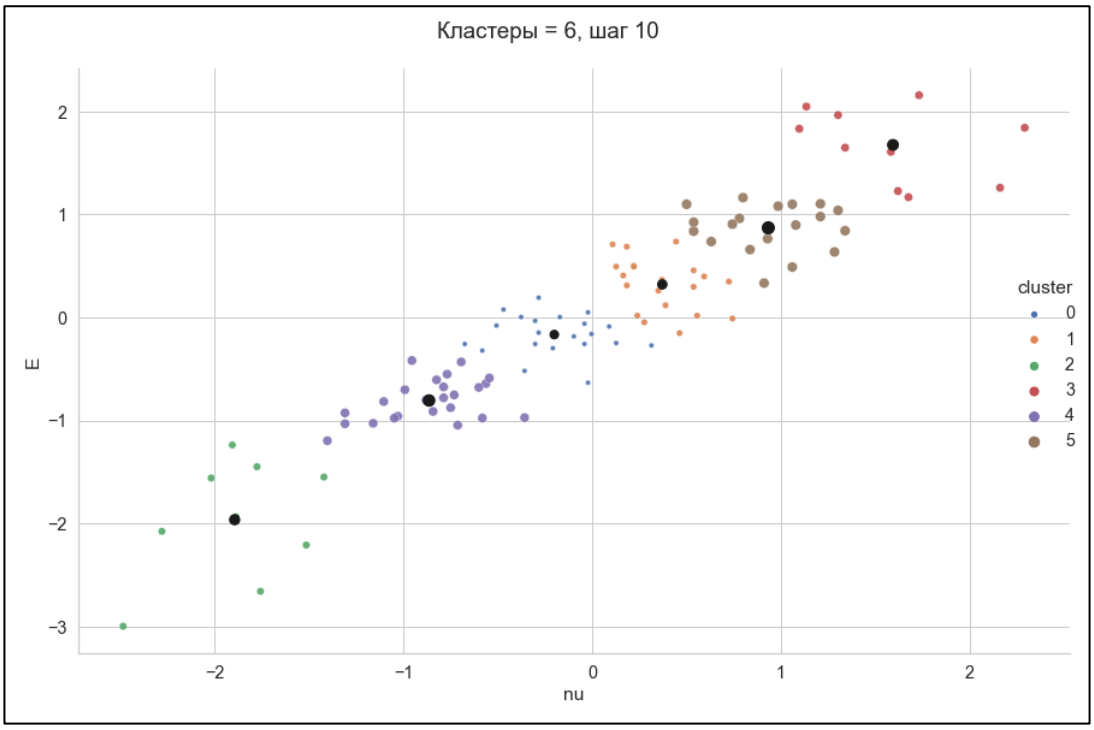

Рисунок 3.2.7 – Кластеризация методом k-средних (6 кластеров)

Таблица 32

Центр кластера	Количество элементов
(-0.2011; -0.1669)	21
(0.3714; 0.3201)	20
(-1.8954; -1.9646)	9
(1.5941; 1.6751)	10
(-0.8648; -0.8068)	24
(0.9333; 0.8698)	20

Таблица 33

F_1	F_2	F_3
44.763	1673.801	1.425
32.729	920.885	1.354
29.655	692.789	1.415
25.679	497.495	1.433
20.715	342.549	1.346
17.233	269.68	1.276
15.541	253.261	1.201
15.174	251.847	1.183
15.096	248.091	1.182
15.096	248.091	1.182

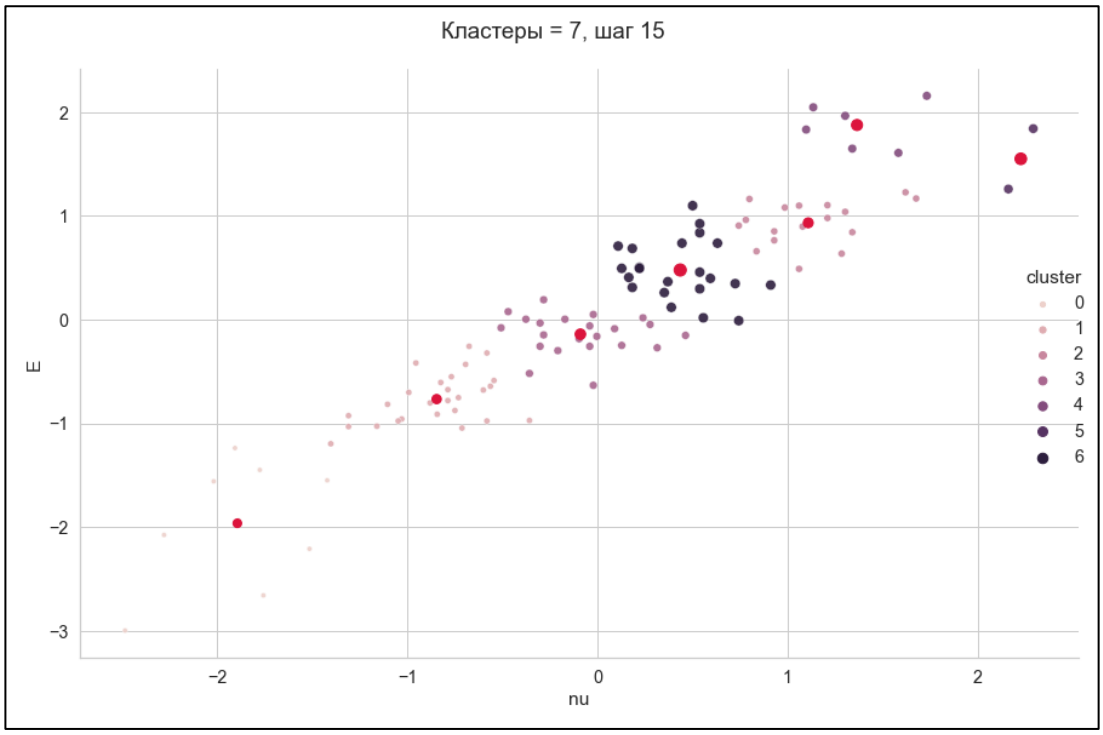

Рисунок 3.2.8 – Кластеризация методом k-средних (7 кластеров)

Таблица 32

Центр кластера	Количество элементов
(-1.8954; -1.9646)	9
(-0.8467; -0.767)	26
(1.1086; 0.933)	17
(-0.0903; -0.1412)	22
(1.3652; 1.8761)	6
(2.227; 1.5499)	2
(0.4349; 0.4779)	22

Таблица 33

F_1	F_2	F_3
29.274	768.748	1.488
25.75	607.845	1.437
23.891	518.596	1.451
21.365	453.482	1.428
19.513	416.434	1.391
...		
14.481	261.669	1.231

В таблице 34 представлены значения функционалов качества на последней итерации и количество итераций для различных значений количества кластеров.

Таблица 34

Количество кластеров	***Количество итераций***	F_1	F_2	F_3
2	5	76.855	4016.464	1.52
3	6	43.857	1497.05	1.349
4	8	35.379	1053.881	1.381
5	7	24.084	654.669	1.259
6	10	15.096	248.091	1.182
7	15	14.481	261.669	1.231

Можно заметить, что при увеличении количества кластеров увеличивается число итераций и минимизируются значения функционалов качества.

- Пересчет центра после каждого изменения состава кластера

При реализации алгоритма в случае изменения центра кластера после каждого изменения его состава были получены следующие значения количества итераций, представленные в таблице 35.

Таблица 35

Количество кластеров	***Количество итераций (центр меняется в конце итерации)***	***Количество итераций (центр меняется после каждого объекта)***
2	5	2
3	6	2
4	8	2
5	7	3
6	10	4
7	15	4

Из таблицы можно увидеть, что количество итераций второй версии алгоритма меньше, чем первой, так как центр корректируется больше раз, что лучше минимизирует функционал качества.

Найдем оптимальное количество кластеров:

1. С помощью метода локтя

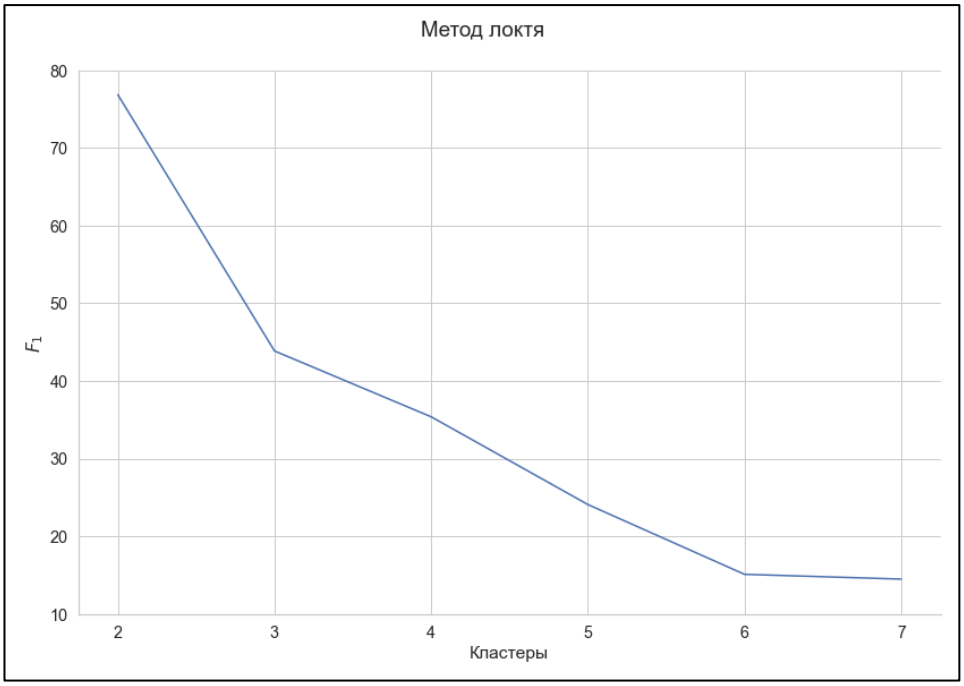

Рисунок 3.2.9 – Метод локтя

Значение числа кластеров можно получить в точке сгиба, после которой значение функционала качества изменяется медленно. Исходя из рис. 3.2.9 можно предположить, что оптимальное число кластеров – 6.

2. С помощью метода силуэтов

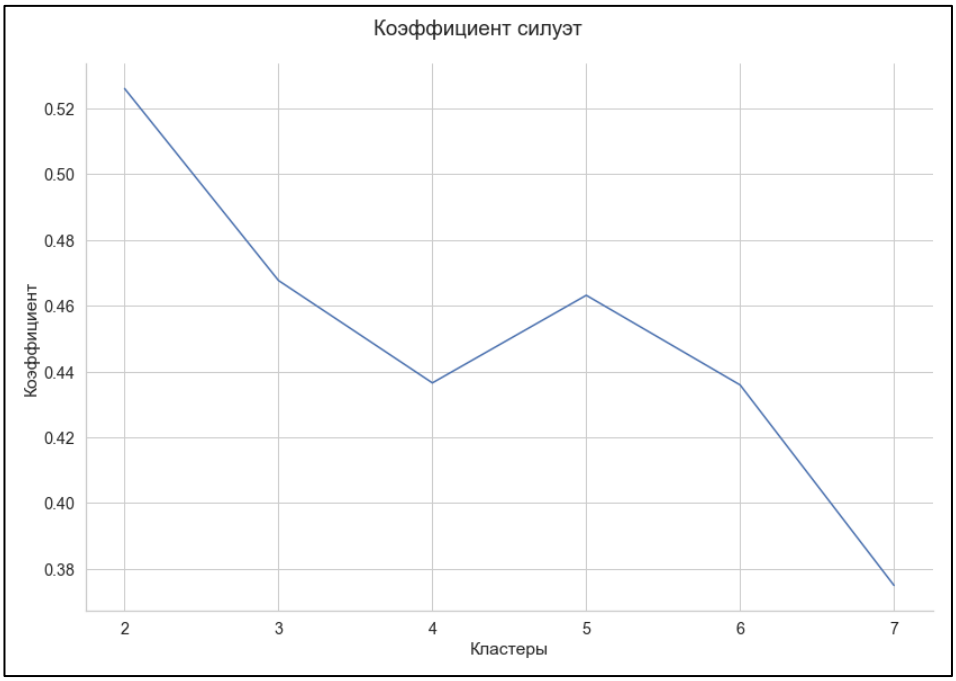

Рисунок 3.2.10 – Метод силуэтов

Коэффициент «силуэт» вычисляется с помощью среднего внутрикластерного расстояния и среднего расстояния до ближайшего кластера по каждому образцу. Можно вычислить среднее значение силуэта по всем образцам и использовать его как метрику для оценки количества кластеров.

График силуэта имеет пиковый характер, в отличие от мягко изогнутого графика при использовании метода локтя. Исходя из рис. 3.2.10 можно предположить, что оптимальное количество кластеров – 2 или 5.

3.3. Метод поиска сгущений

Был реализован алгоритм поиска сгущений. Полученные кластеры были отображены разными цветами.

Вычислены нижняя и верхняя границы радиуса сферы:

$$R_{min} = mind_{ij} = 0.0092$$

$$R_{max} = maxd_{ij} = 6.8015$$

Начиная с $R_{min} = 0.1$ и изменяя радиус на $\delta = 0.05$ было найдено значение радиуса, которое приводит к устойчивому разбиению.

$$R = 1.15$$

В качестве центра сферы на первом шаге был выбран объект, в ближайшей окрестности которого было расположено максимальное число соседей – $(x_{58}; y_{58}) = (-0.2091; -0.2994)$.

В заголовках рисунков можно увидеть изменение центров кластеров и количества элементов.

Формирование кластеров представлено на рис. 3.3.3 - 3.3.16:

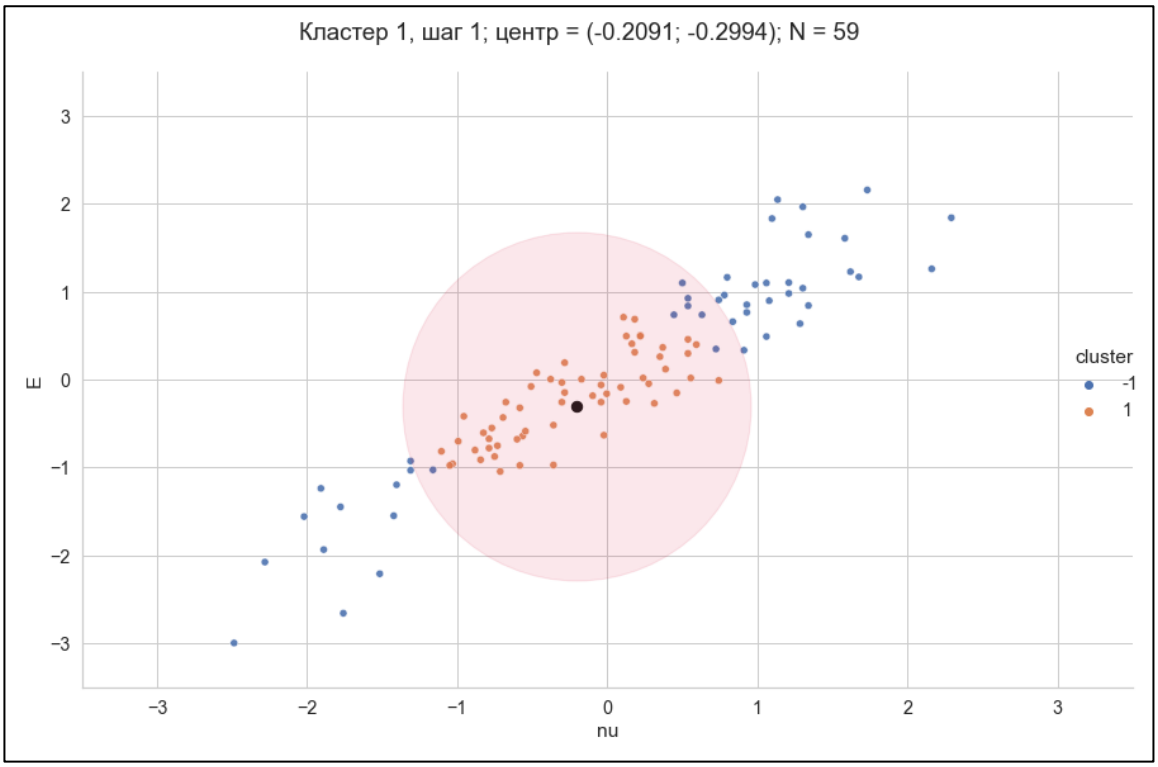

Рисунок 3.3.3

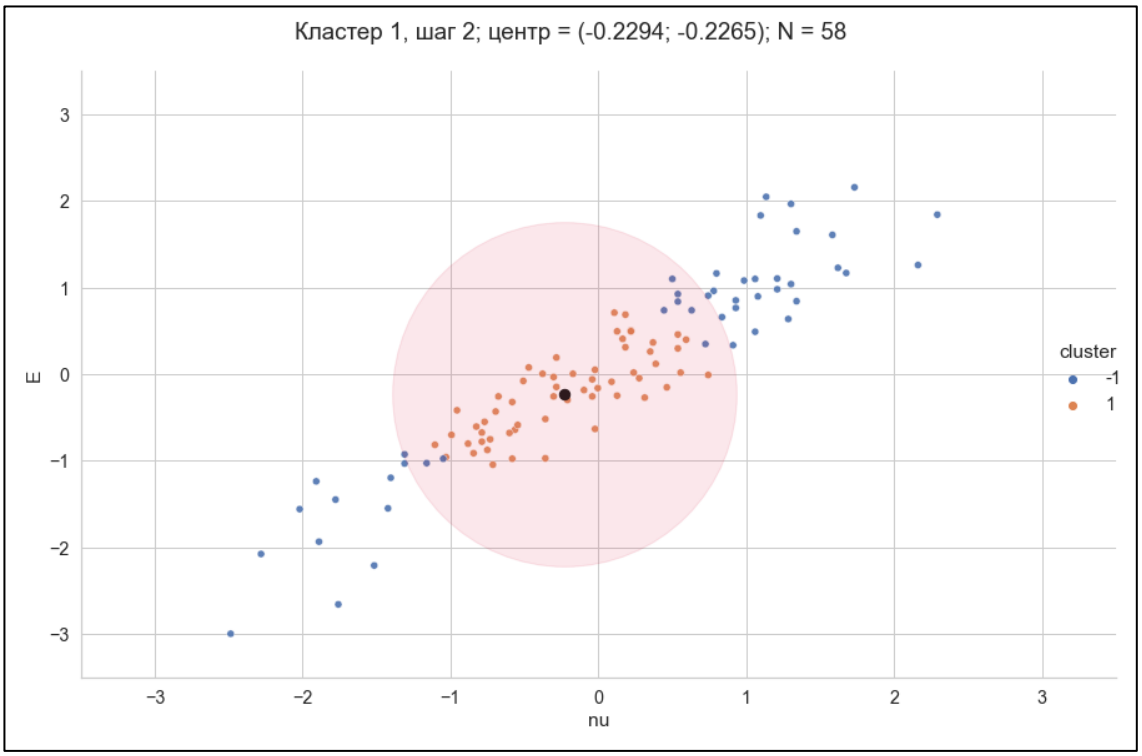

Рисунок 3.3.4

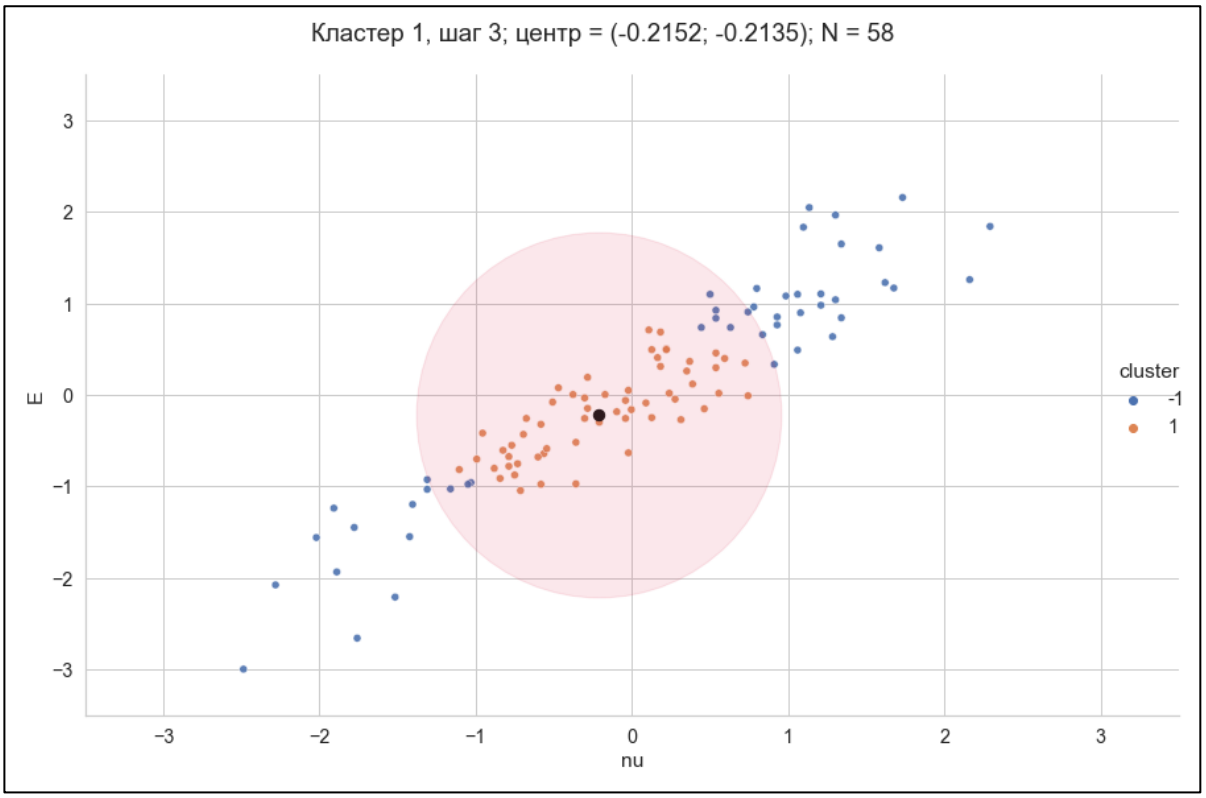

Рисунок 3.3.5

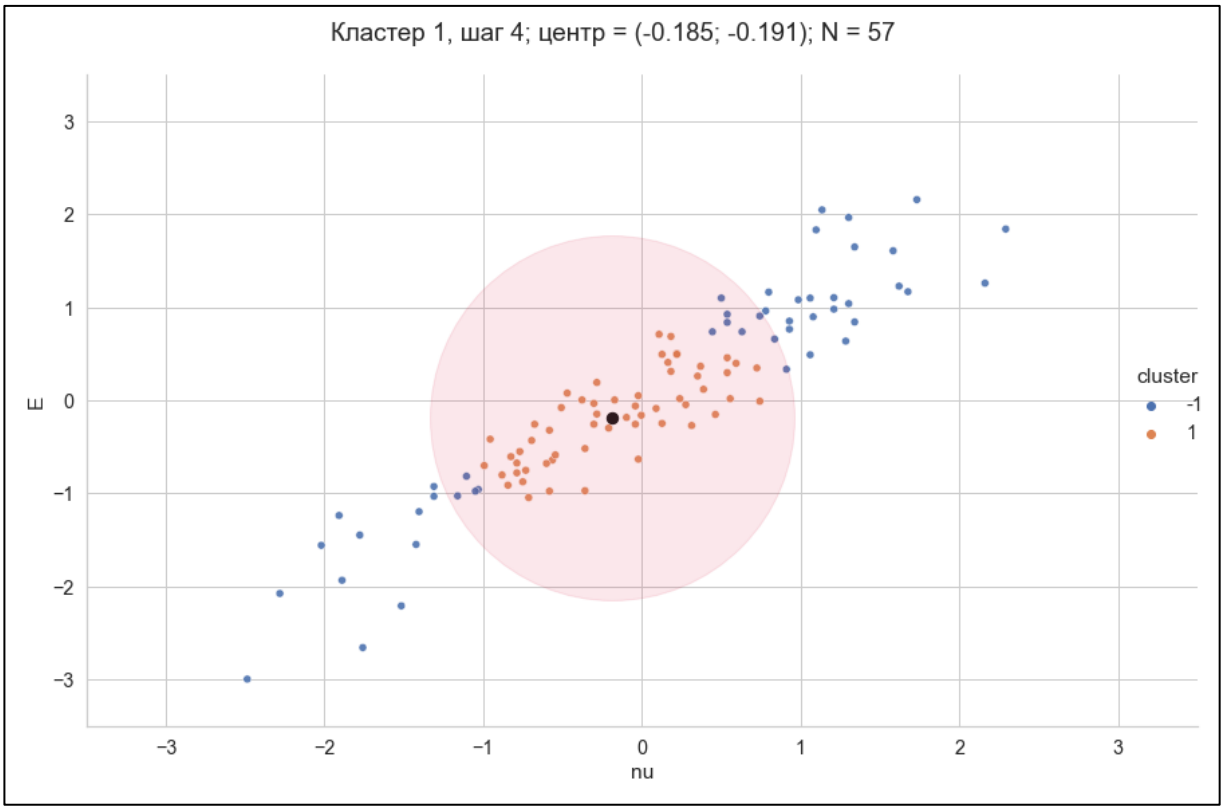

Рисунок 3.3.6

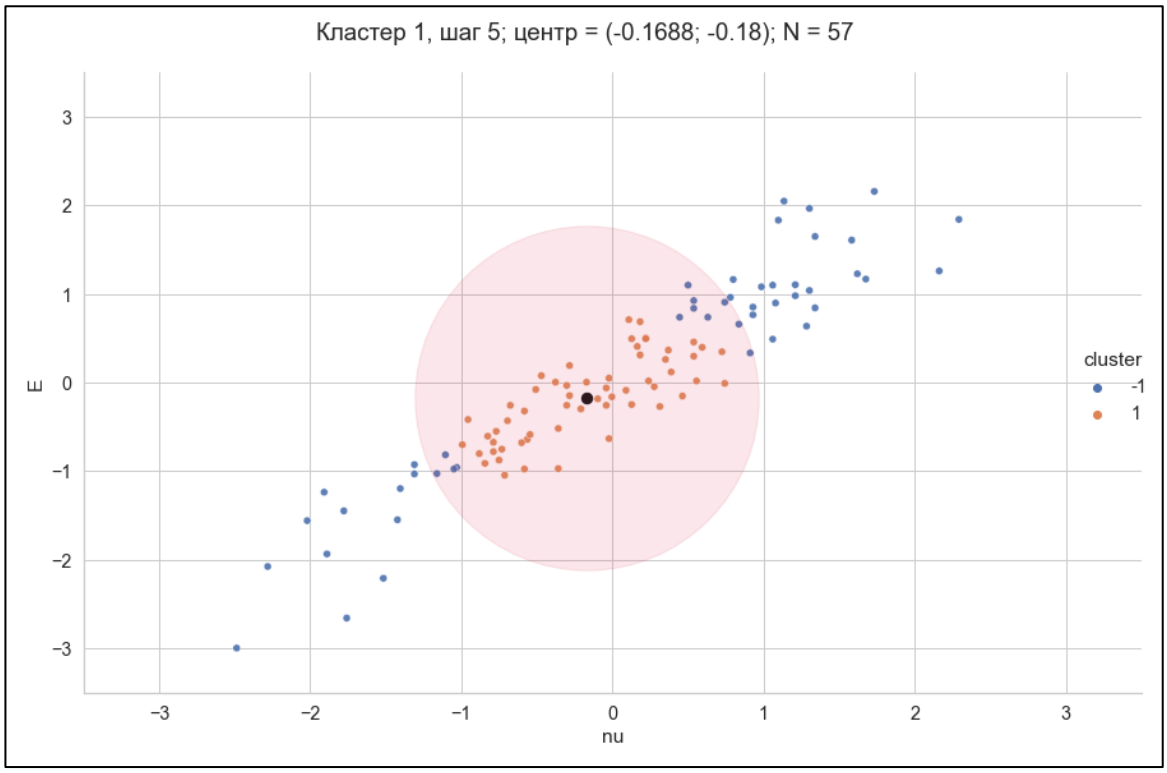

Рисунок 3.3.7

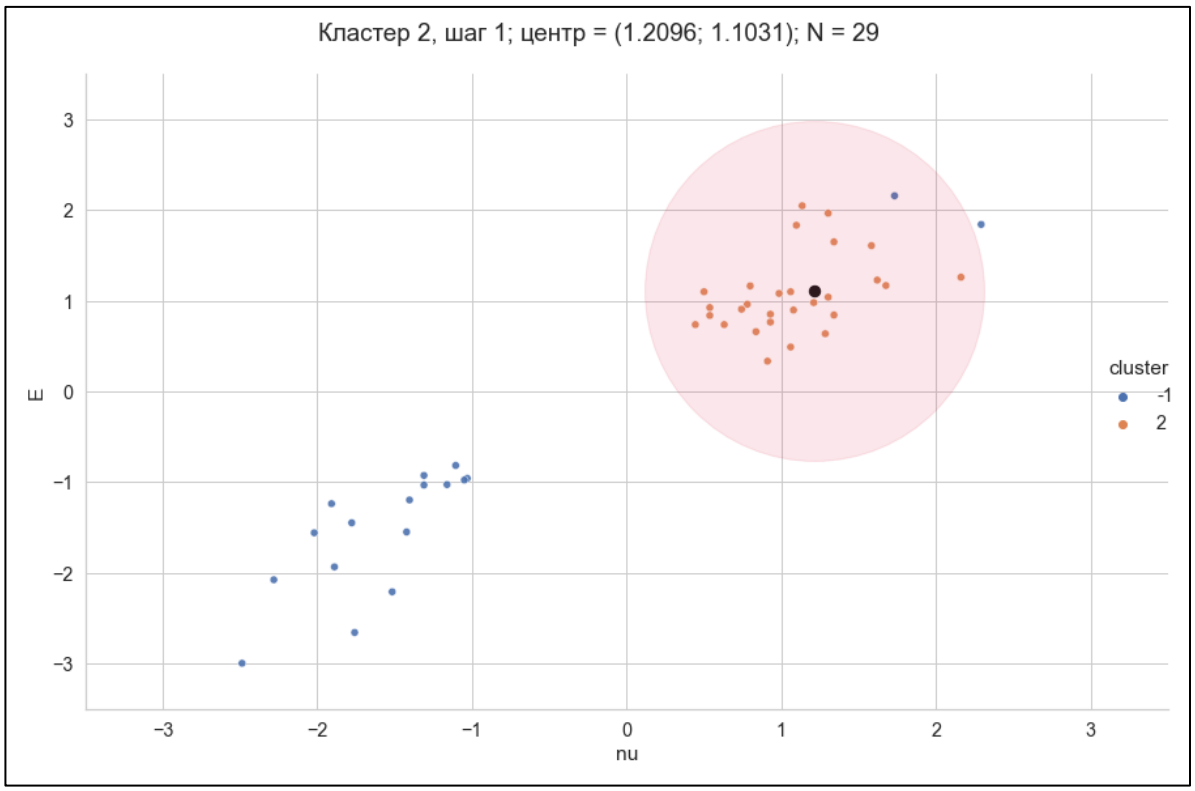

Рисунок 3.3.8

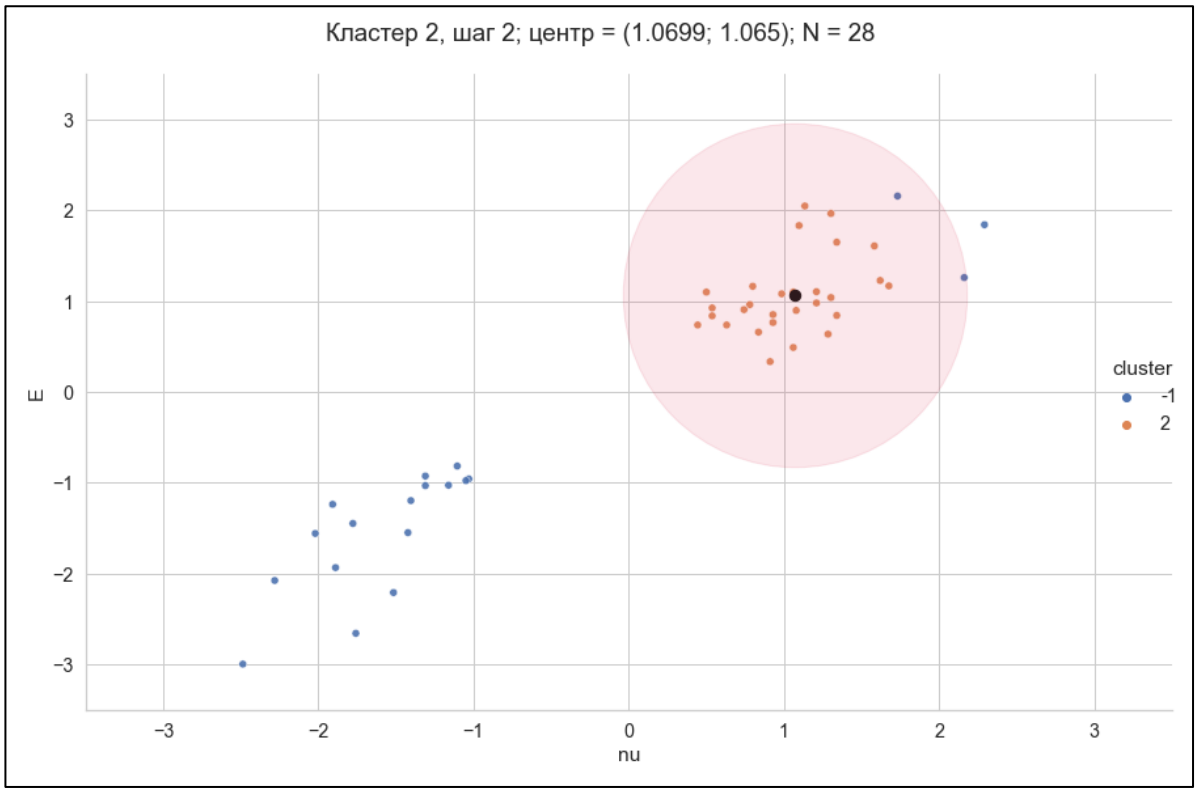

Рисунок 3.3.9

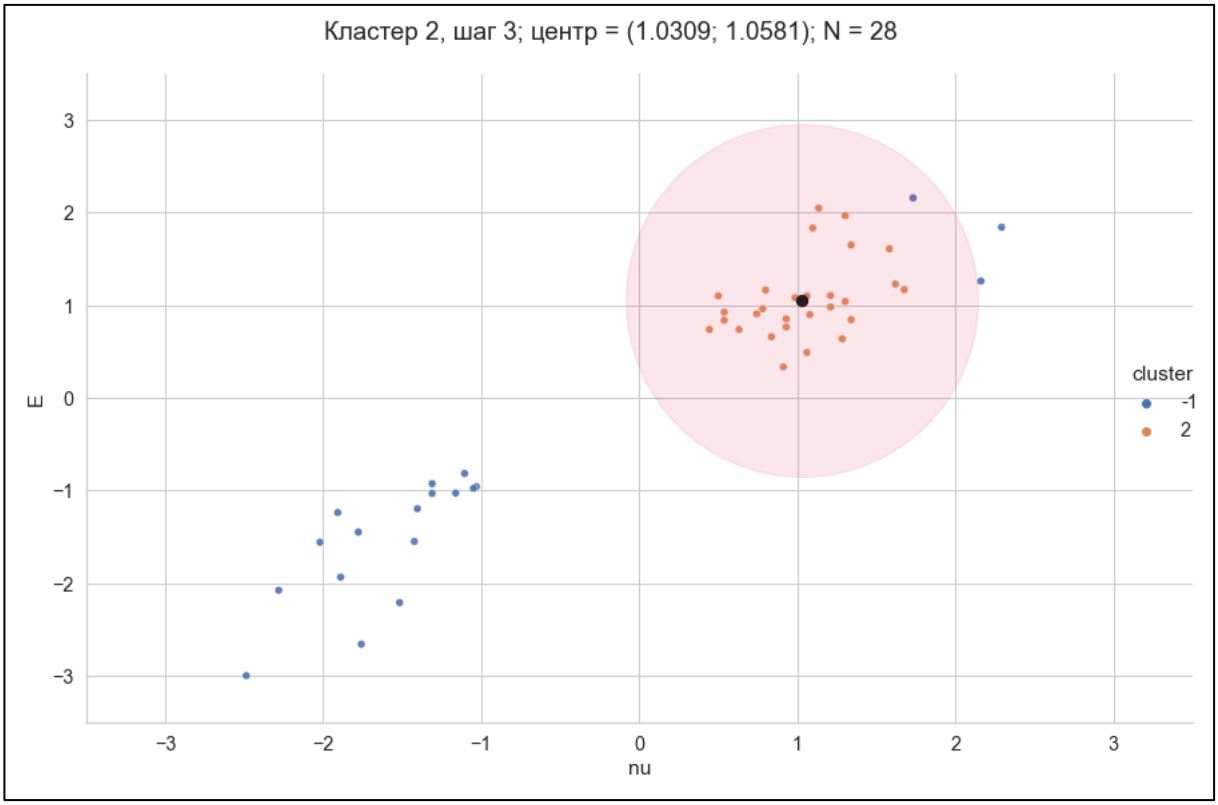

Рисунок 3.3.10

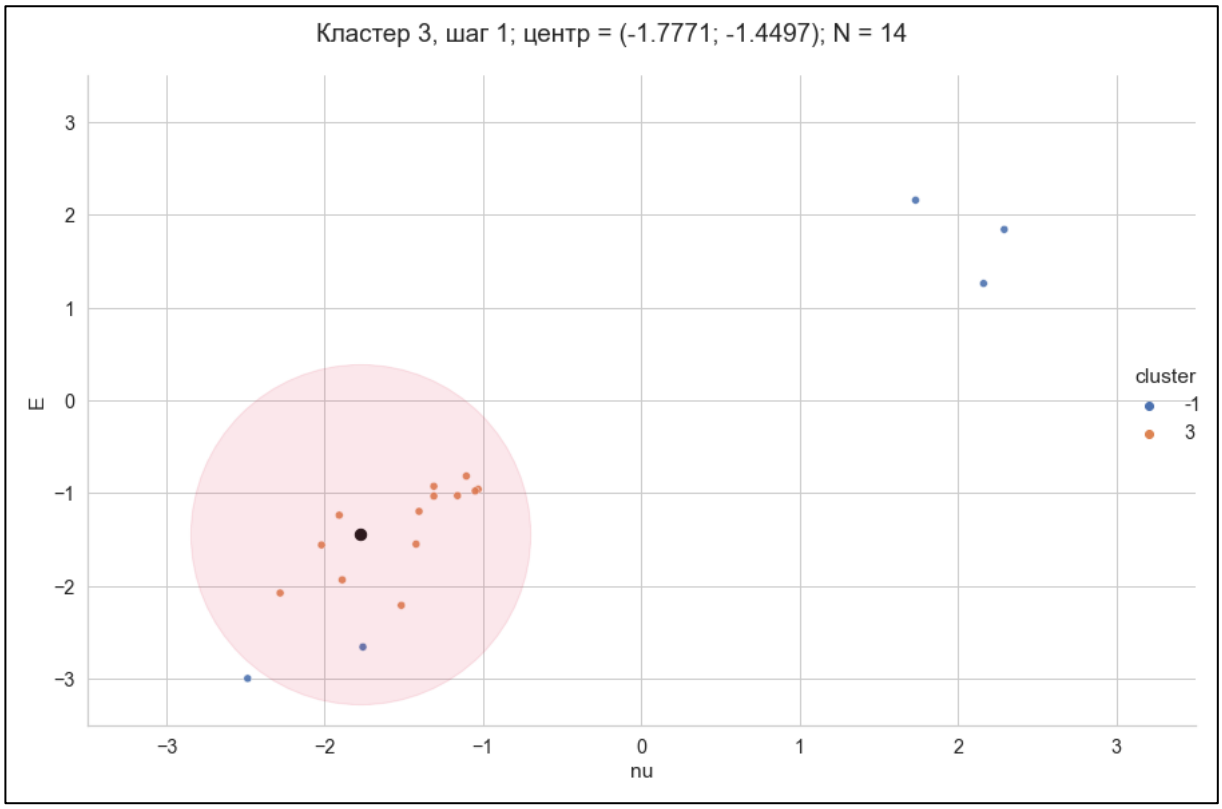

Рисунок 3.3.11

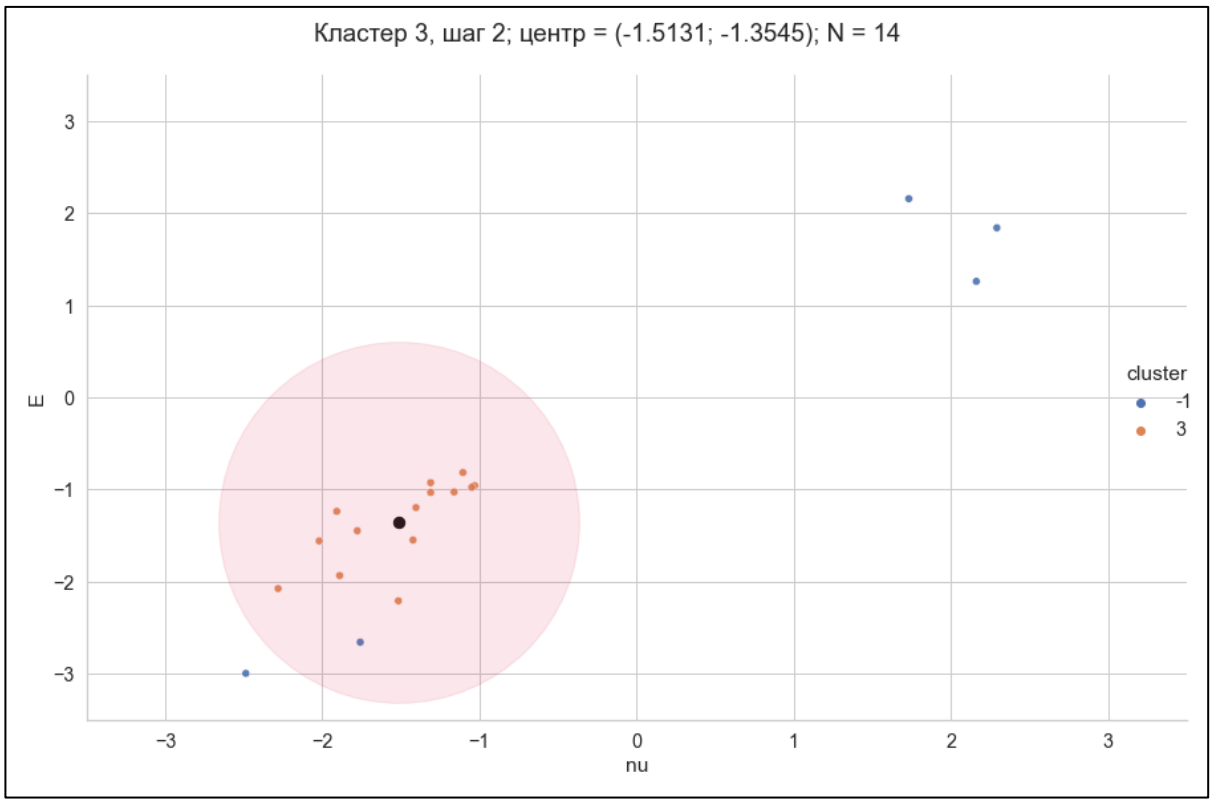

Рисунок 3.3.12

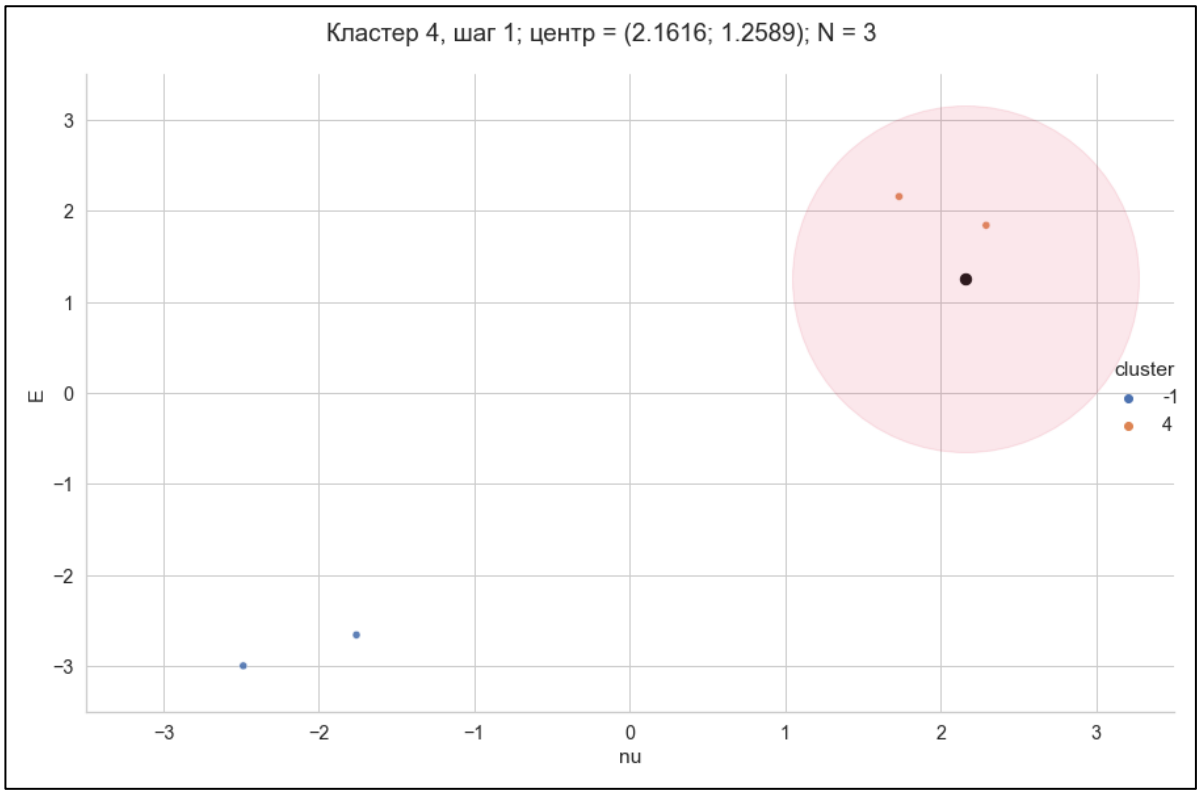

Рисунок 3.3.13

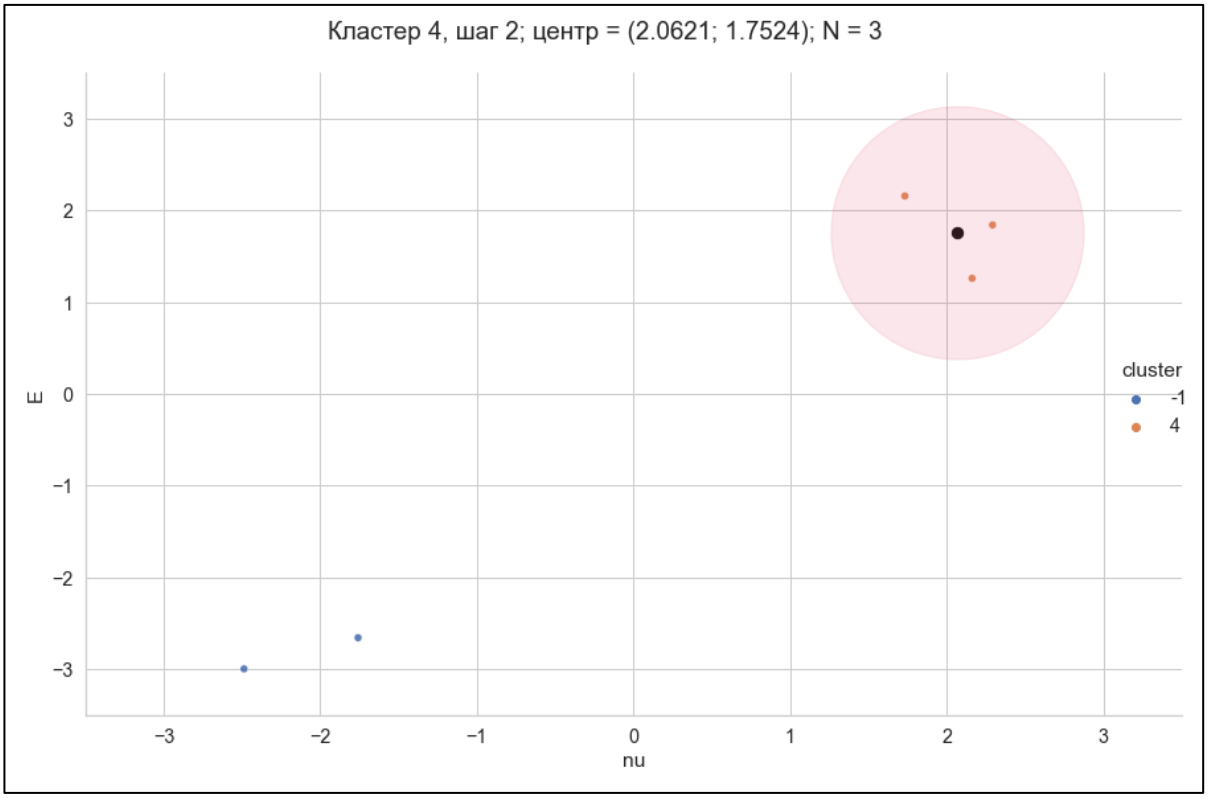

Рисунок 3.3.14

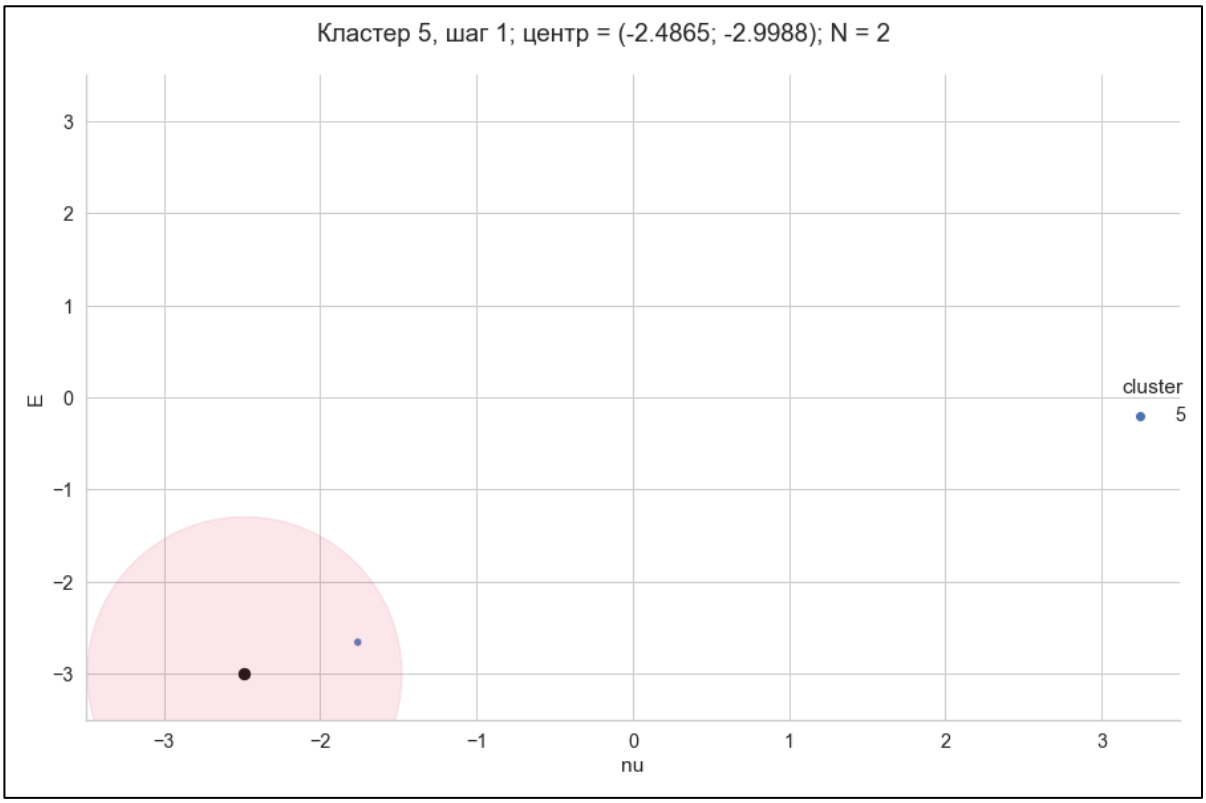

Рисунок 3.3.15

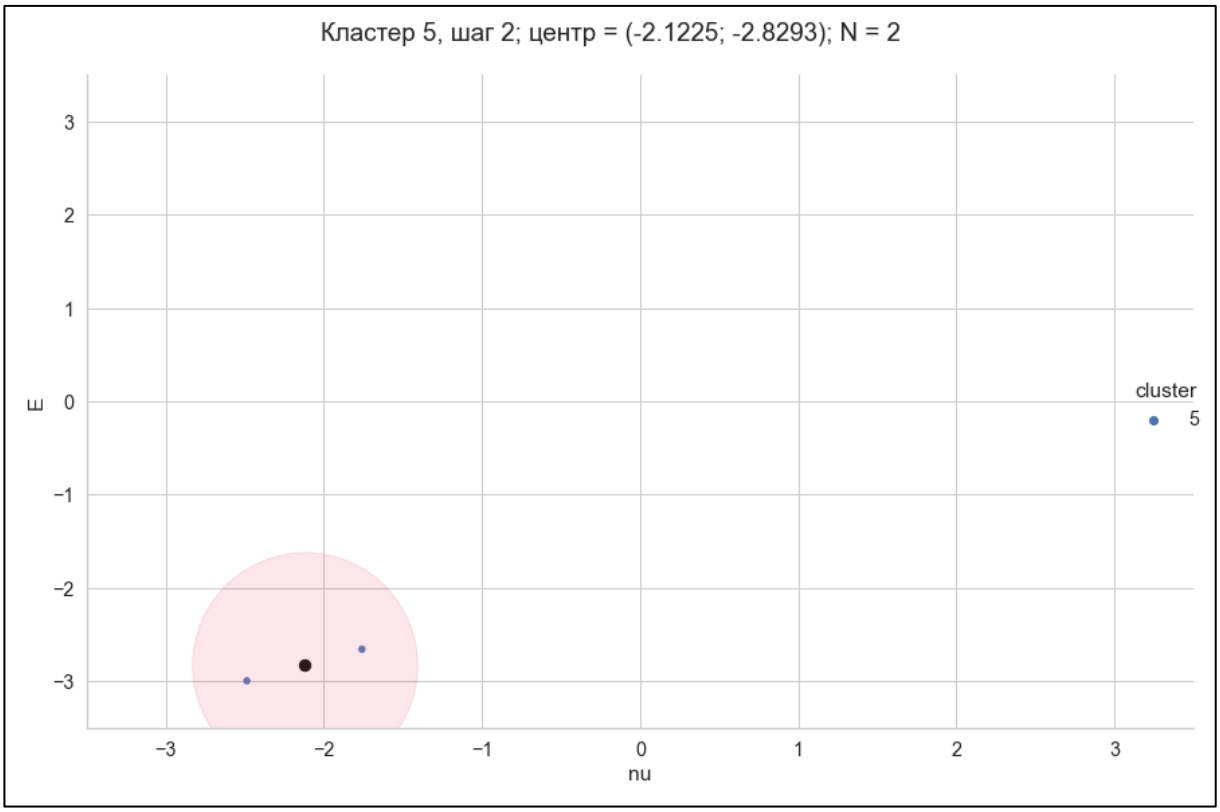

Рисунок 3.3.16

Результат кластеризации представлен на рис. 3.3.17.

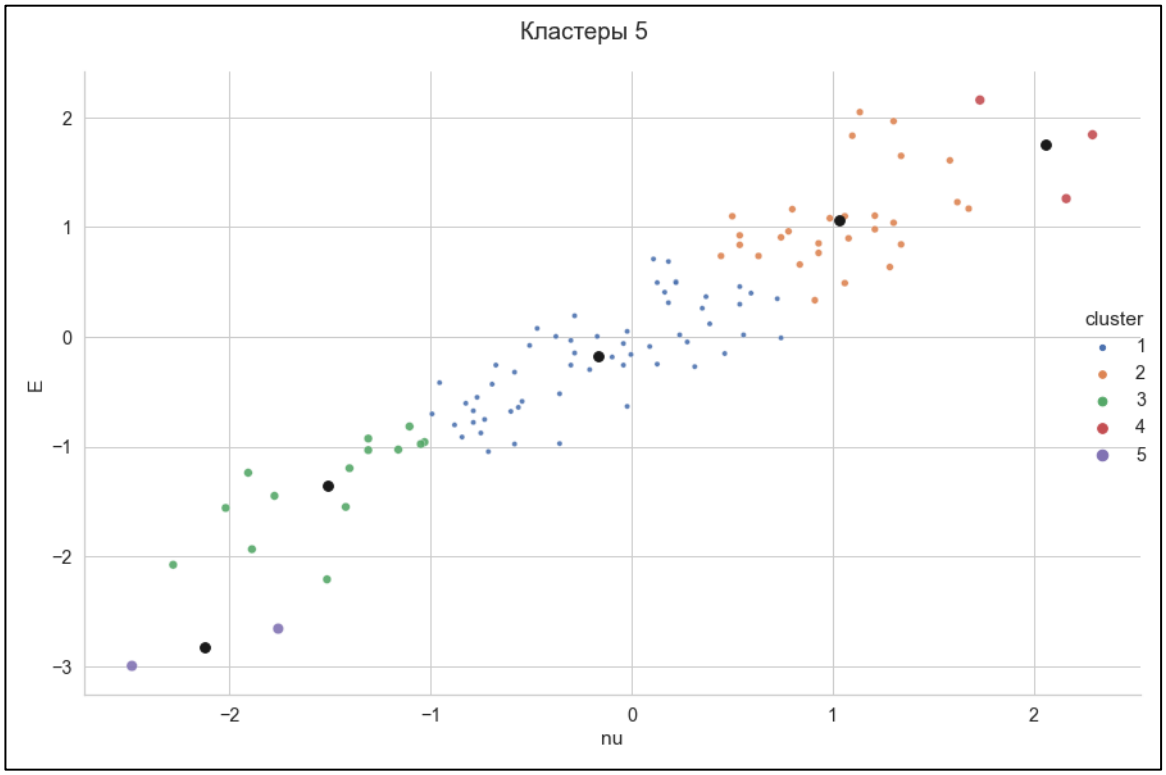

Рисунок 3.3.17

Для проверки чувствительности метода к погрешностям было произведено сравнение значений функционалов качества для первоначального значения R=1.15 и для его изменений на $\delta = 0.02$. Функционалы F_1, F_2, F_3 определены из прошлой лабораторной работы. Значения представлены в таблице 36.

Таблица 36

Радиус	$\boldsymbol{F_1}$	$\boldsymbol{F_2}$	$\boldsymbol{F_3}$
$\boldsymbol{R}$	41.0773	1928.5299	1.6819
$\boldsymbol{R-\delta}$	41.1502	1944.8672	1.6772
$\boldsymbol{R+\delta}$	43.6907	2144.739	1.8447

На основании данных таблицы можно увидеть, что значения функционалов качества растут (хоть и немного) при изменении радиуса на

небольшую дельту. Можно сделать вывод, что метод чувствителен к погрешностям.

Сравним метод k-means с методом поиска сгущений. Количество кластеров равно 5. Визуальное сравнение представлено на рис. 3.3.18 и 3.3.19.

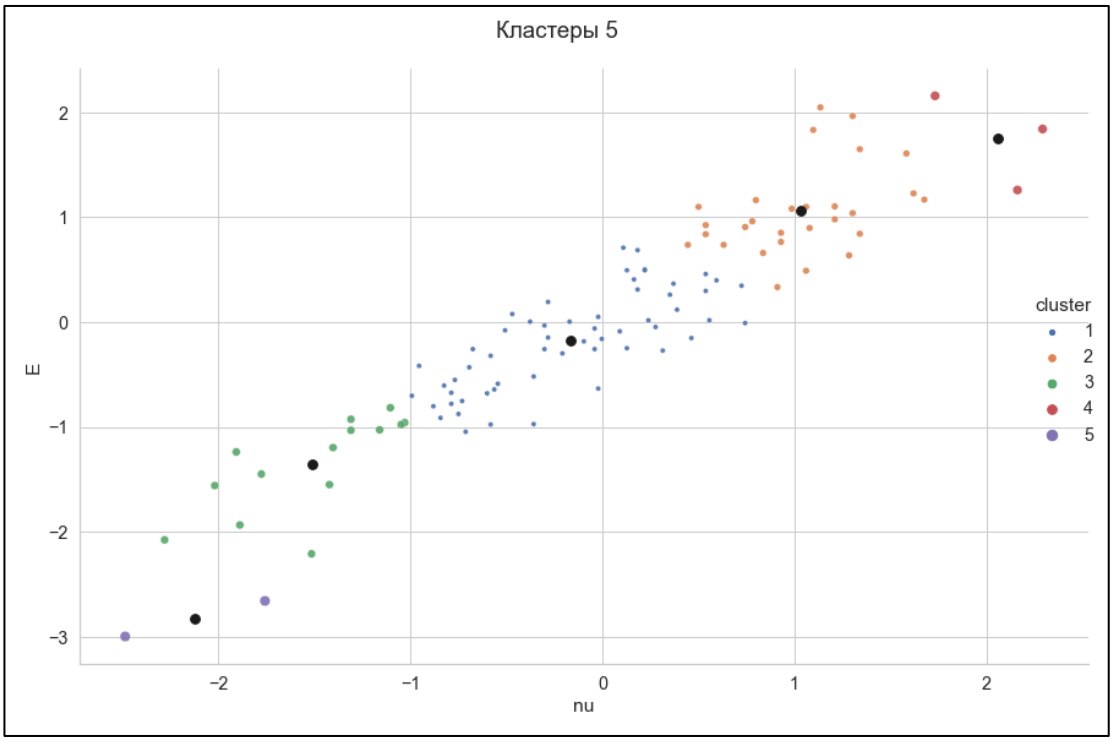

Рисунок 3.3.18 – Метод поиска сгущений

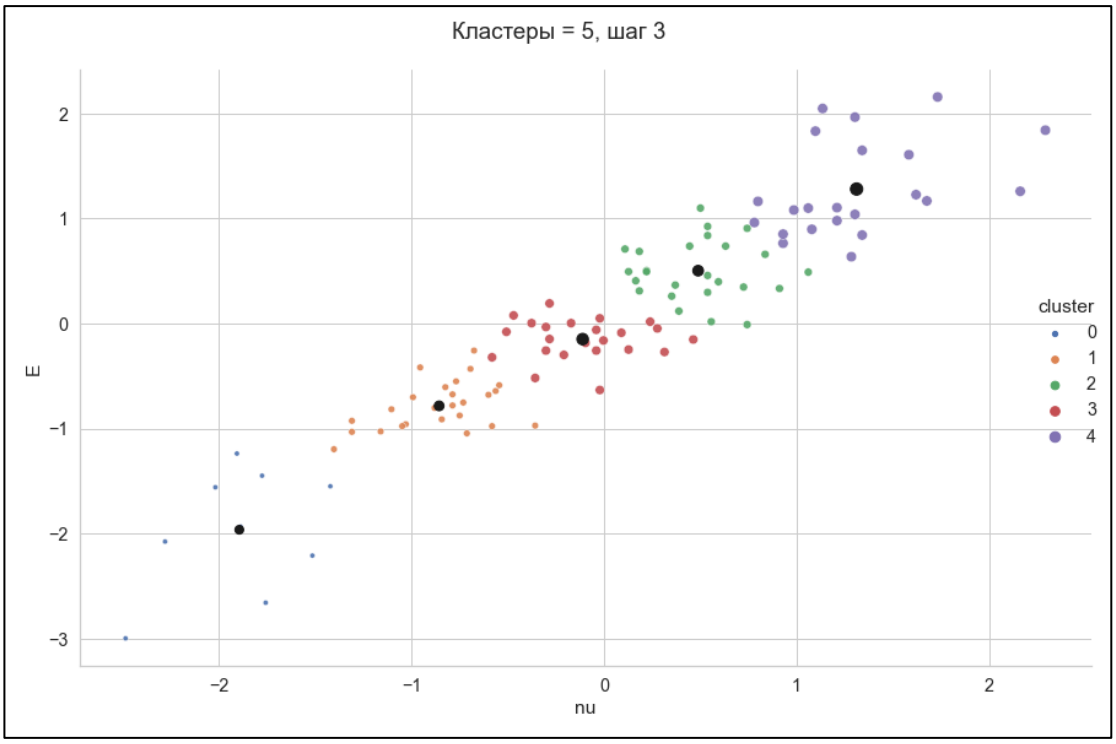

Рисунок 3.3.19 – Метод k-средних

Визуально можно увидеть, что в методе k-средних количество элементов в кластерах примерно одинаковое в отличие от метода поиска сгущений.

В таблице 37 приведены значения функционалов качества для методов.

Таблица 37

Метод	F_1	F_2	F_3
k-means	24.084	654.669	1.259
Поиск сгущений	41.0773	1928.5299	1.6819

Видно, что значения функционалов качества метода k-средних намного меньше, можно сделать вывод, что использование метода k-средних предпочтительнее.

3.4. Выводы.

В ходе выполнения лабораторной работы были освоены основные понятия кластерного анализа, в частности, метода k-средних. Исходная двумерная выборка была нормализована и отображена на рисунке. Была определена грубая верхняя оценка количества кластеров $\bar{k} = 7$.

Реализован алгоритм k-means в двух вариантах: пересчет центра осуществляется после каждого изменения состава кластера, либо же после просмотра всех данных. Первый вариант имеет меньшее число шагов процедуры, так как центр корректируется больше раз, что лучше минимизирует функционал качества.

Разбиение проводилось для разного количества кластеров, от 2 до 7. Можно заметить, что при увеличении количества кластеров увеличивается число итераций алгоритма и минимизируются значения функционалов качества.

Было найдено оптимальное значение количества кластеров с помощью метода локтя и метода силуэтов. В первом случае значение равно шести, во втором же пяти.

Освоены основные понятия кластерного анализа и метода поиска сгущений, в частности. Было нормализовано множество точек.

Были найдены границы радиуса сферы.

$$R_{min} = mind_{ij} = 0.0092$$

$$R_{max} = maxd_{ij} = 6.8015$$

Был реализован алгоритм поиска сгущений, с помощью которого выборка была разбита на 5 кластеров для $R = 1.15$. Кластеры были отображены, выделены цветом, отмечены центроиды.

Была проведена проверка чувствительности метода к погрешностям. На основании данных можно увидеть, что значения функционалов качества растут при изменении радиуса на дельту. Можно сделать вывод, что метод чувствителен к погрешностям.

Было проведено сравнение метода k-means с методом поиска сгущений. Визуально можно увидеть, что в методе k-средних количество элементов в кластерах примерно одинаковое в отличие от метода поиска сгущений. А также значения функционалов качества метода k-средних намного меньше, следовательно можно сделать вывод, что использование метода k-means предпочтительнее.

ЗАКЛЮЧЕНИЕ

В ходе выполнения курсовой работы были выполнены все поставленные цели: построены выборки из генеральной совокупности заданного объёма, построены ранжированные, вариационные и интервальные ряды, графически построены полигоны частот, гистограммы, эмпирические функции распределения двумерной выборки.

Найдены выборочные оценки: среднего, дисперсии, СКО, асимметрии, эксцесса, медианы и моды, построены доверительные интервалы для математического ожидания и СКО, проверена гипотеза о нормальном законе с помощью критерия Пирсона.

Построена корреляционная таблица, найдена оценка коэффициента корреляции, проверена гипотеза о равенстве коэффициента корреляции нулю, построены уравнения выборочных прямых среднеквадратической регрессии, найдены оценки корреляционных отношений

Нормализовано множество точек, реализован алгоритм k-means, отображены полученные кластеры, реализован метод поиска сгущений, произведена оценка качества кластеризации, проверена чувствительность метода поиска сгущений к погрешностям, произведено сравнение методов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Методические указания по выполнению курсовой работы: учеб.-метод. пособие / сост.: А.-В.И. Середа. СПб. 2016. 15 с.

2. Белоногов А.М., Попов Ю.И., Посредник О.В. Статистическая обработка результатов физического эксперимента [Комплект] : учеб. пособие: - СПб. : Изд-во СП6ГЭТУ "ЛЭТИ", 2009.

3. Морозов В.В., Соботковский Б.Е., Шейнман И.Л. Методы обработки результатов физического эксперимента: учеб. пособие: — СПб.: Изд-во СПБГЭТУ «ЛЭТИ», 2004.

4. Егоров В.А. и др. Анализ однородных статистически данных: учеб. пособие: — СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2005.

5. Буре В.М., Парилина Е.М., Свиркин М.В. Математическая статистика. СПб.: факультет ПМ ПУ СПбГУ, 2007.

6. Котельников Р.Б. Анализ результатов наблюдений.

7. Митин И.В., Русаков В.С. Анализ и обработка экспериментальных данных. М.: Физический факультет МГУ, 2006.

8. Смирнов Н.А., Экало А.В. Методы обработки экспериментальных данных: учеб. пособие: — СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2009.

9. Пособие по практическим занятиям: учеб.-метод. пособие / сост.: А-В.И. Середа.СПб. 2016. 12 с.

10. Регрессия // bstudy.ru URL: https://bstudy.net/672105/sotsiologiya/regressiya_vide_stepennoy_funktsii (дата обращения: 05.04.2022).

11. Метод k-means // statistica.ru URL: http://statistica.ru/theory/klasterizatsiya-metod-k-srednikh/ (дата обращения: 05.04.2022).

ПРИЛОЖЕНИЕ А
ПРОГРАММА ДЛЯ ФОРМИРОВАНИЯ И ПЕРВИЧНОЙ ОБРАБОТКИ ВЫБОРКИ, ПОСТРОЕНИЯ, РАНЖИРОВАННОГО И ИНТЕРВАЛЬНОГО РЯДОВ

```
# %%
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

# %% [markdown]
# ## Выборка

# %%
raw = pd.read_csv('c:/Users/gandh/dev/unv/smoed/data/sample.csv')
df = pd.read_csv('c:/Users/gandh/dev/unv/smoed/data/main_data.csv')
df.to_csv('data/data1.csv', index=False)
n = len(df)
n

# %% [markdown]
# ## Распределение

# %%
sns.set_theme(style="whitegrid", palette='deep', context='notebook',
font_scale=1.3)
ax = sns.catplot(data=raw, kind='box', height=8.27, aspect=11.7/8.27)
plt.savefig('pics/1.png')
# %%
sns.set_theme(style="whitegrid", palette='deep', context='notebook',
font_scale=1.3)
ax = sns.catplot(data=df, kind='box', height=8.27, aspect=11.7/8.27)
plt.savefig('pics/2.png')

# %% [markdown]
# ## Одна переменная
# %%
df2 = df.drop('E', axis=1)
df2.to_csv('data/data2.csv', index=False)
df2.head()
# %% [markdown]
# ## Ранжированный ряд
# %%
df2 = df2.sort_values(by=['nu'], ignore_index = True)
df2.to_csv('data/data3.csv', index=False)
df2.head()
```

```
# %%
df2.min()
df2.max()
# %%
X = df2['nu']
# %% [markdown]
# ## Вариационный ряд
# %%
table_af = X.value_counts().sort_index()
table_rf = X.value_counts(normalize=True).sort_index()
table_af = pd.DataFrame({'nu': table_af.index, 'af': table_af.values})
table_rf = pd.DataFrame({'nu': table_rf.index, 'rf': table_rf.values})
table_rf2 = table_rf.copy()
table_rf2['rf'] = np.round(table_rf2['rf'], 4)
table_af.to_csv('data/data4.csv', index=False)
table_rf2.to_csv('data/data5.csv', index=False)

# %% [markdown]
# ## Интервальный ряд

# %%
k = 1+3.31*np.log10(n)
k = int(np.floor(k))
k

# %%
min(X)
max(X)

# %%
h = (max(X)-min(X))/k
h = int(np.ceil(h))
h

# %%
data_interval = pd.concat([table_af, table_rf], ignore_index=True, ax-
is=1).drop(2, axis=1)
data_interval.columns = ['nu', 'af', 'rf']
data_interval.to_csv('data/data6.csv', index=False)

# %%
ivs = np.hstack((np.arange(min(X), max(X), h), np.array(max(X))))

# %%
data_interval['inter'] = pd.cut(data_interval['nu'], bins=ivs,
                                right=False)
data_interval['inter'].value_counts().sort_index()
data_interval.iloc[83, 3] = data_interval.iloc[82, 3]

# %%
```

```
f_inter = data_interval.groupby(['inter'])[['af',
'rf']].apply(sum).reset_index()
f_inter['avg_inter'] = np.array([np.mean([ivs[i], ivs[i+1]], axis=0) for
i in range(k)])
f_inter = f_inter[['inter', 'avg_inter', 'af', 'rf']]
f_inter.to_csv('data/data7.csv', index=False)

# %% [markdown]
# ## Графики абсолют

# %%
ax = sns.relplot(data=f_inter, x='avg_inter', y='af', kind='line',
height=8.27, aspect=11.7/8.27)
ax.set_axis_labels('Середина интервала', 'Частота')
ax.set(ylim=[0,28], xticks=f_inter['avg_inter'])
ax.fig.suptitle('Полигон для абсолютных частот')
plt.savefig('pics/3.png')

# %%
ax = sns.displot(data=df, x='nu', bins=ivs, kind='hist', height=8.27, as-
pect=11.7/8.27, stat='probability')
ax.set_axis_labels('Середина интервала', 'Частота')
ax.set(ylim=[0,.26],xticks=f_inter['avg_inter'])
ax.fig.suptitle('Гистограмма для абсолютных частот')
plt.savefig('pics/4.png')

# %%
f_inter['sum_rf'] = f_inter['rf'].cumsum()
f_inter

# %%
ax = sns.relplot(data=f_inter, x='avg_inter', y='sum_rf', s=80,
                 kind='scatter', height=8.27, aspect=11.7/8.27, col-
or='b')
for i in range(6):
    plt.hlines(f_inter['sum_rf'][i], f_inter['avg_inter'][i],
f_inter['avg_inter'][i+1], color='b')
plt.hlines(1, 559, 589, color='b')
for i in range(6):
    plt.vlines(f_inter['avg_inter'][i+1], f_inter['sum_rf'][i],
f_inter['sum_rf'][i+1], color='b', linestyle='--')
plt.vlines(338.5, 0, 0.048, color='b', linestyle='--')
ax.set_axis_labels('Середина интервала', '')
ax.set(xticks=f_inter['avg_inter'])
ax.fig.suptitle('Эмпирическая фукнция распределения')
plt.savefig('pics/5.png')

# %% [markdown]
# ## Графики относительно

# %%
```

```
ax = sns.relplot(data=f_inter, x='avg_inter', y='rf', kind='line',
height=8.27, aspect=11.7/8.27)
ax.set_axis_labels('Середина интервала', 'Частота')
ax.set(ylim=[0,0.26], xticks=f_inter['avg_inter'])
ax.fig.suptitle('Полигон для относительных частот')
plt.savefig('pics/6.png')

# %%
ax = sns.displot(data=df, x='nu', bins=ivs, kind='hist', height=8.27, as-
pect=11.7/8.27, stat='density')
ax.set_axis_labels('Середина интервала', 'Частота')
ax.set(xticks=f_inter['avg_inter'])
ax.fig.suptitle('Гистограмма для относительных частот')
plt.savefig('pics/7.png')

# %%
f_inter['af']/h
f_inter['rf']/h
```

ПРИЛОЖЕНИЕ Б

ПРОГРАММА ДЛЯ НАХОЖДЕНИЯ ТОЧЕЧНЫХ ОЦЕНОК ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

```
# %%
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

# %%
original =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/lab1/data/data2.csv')
var_row =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/lab1/data/data4.csv')
var_row.to_csv('data/var_row.csv', index=False)
n = 104
h = 37

# %%
int_row =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/interval.csv')
int_row['cum_sum'] = np.round(np.cumsum(int_row['rf']), 3)
int_row['rf'] = np.round(int_row['rf'], 3)
int_row.to_csv('data/int_row.csv', index=False)

# %%
usl_mom = int_row.copy()
usl_mom = usl_mom.iloc[:, [1,3]]
usl_mom['u'] = np.arange(-3,4,1)
usl_mom['nu'] = usl_mom['rf']*usl_mom['u']
```

```
usl_mom['nu2'] = usl_mom['rf']*pow(usl_mom['u'], 2)
usl_mom['nu3'] = usl_mom['rf']*pow(usl_mom['u'], 3)
usl_mom['nu4'] = usl_mom['rf']*pow(usl_mom['u'], 4)
usl_mom['nu4+'] = usl_mom['rf']*pow(usl_mom['u']+1, 4)

# %%
usl_mom_f = usl_mom.append(usl_mom.sum(), ignore_index=True)
usl_mom_f.to_csv('data/usl_mom.csv', index=False)

# %%
moms = usl_mom_f.iloc[7, [3,4,5,6]]
moms[3]+4*moms[2]+6*moms[1]+4*moms[0]+1

# %%
int_mean = (int_row['avg_inter']*int_row['af']).sum()/n
int_var = (((int_row['avg_inter']-int_mean)**2)*int_row['af']).sum()/n
s = int_var*(n/(n-1))
std_s = np.sqrt(s)
std_var = np.sqrt(int_var)
std_s
std_var

# %%
np.mean(original, axis=0)
np.std(original, axis=0)
np.var(original, axis=0)*(n/(n-1))

# %%
M1 = moms[0]*h+449.5
m2 = (moms[1] - pow(moms[0],2))*pow(h,2)
m3 = (moms[2] - 3*moms[1]*moms[0] + 2*pow(moms[0],3))*pow(h,3)
m4 = (moms[3] - 4*moms[2]*moms[0] + 6*moms[1]*pow(moms[0],2) -
3*pow(moms[0],4))*pow(h,4)
```

```
# %%
As = m3/(pow(s, 3))
Ex = (m4/(pow(s, 4))) - 3

# %%
As, Ex

# %%
original.mean()
np.asarray(original.mode())
original.median()

# %%
raw_mode = 431+h*(2/3)
raw_median = 431+(((0.5*n)-36)/25)*h
raw_mode
raw_median
int_mean

# %%
original.head()

# %%
sns.set_theme(style="whitegrid",   palette='deep',   context='notebook',
font_scale=1.3)
ax = sns.displot(data=original, x='nu', bins=np.array([320, 357, 394, 431,
468, 505, 542, 576]),
                 kind='hist',       height=8.27,       aspect=11.7/8.27,
stat='density')
plt.vlines(raw_mode, 0, int_row.loc[3, 'rf']/h, colors='orange', lin-
estyles='solid', label='$M_o$')
plt.vlines(raw_median, 0, int_row.loc[3, 'rf']/h, colors='r', lin-
estyles='solid', label='$M_e$')
```

```
plt.vlines(int_mean, 0, int_row.loc[3, 'rf']/h, colors='k', lin-
estyles='solid', label='$x_в$')
ax.set_axis_labels('Середина интервала', 'Частота')
ax.set(xticks=int_row['avg_inter'], yticks=round((int_row['rf']/h), 4))
ax.fig.suptitle('Гистограмма для относительных частот')
plt.legend()
plt.savefig('pics/1.png')

# %%
```

ПРИЛОЖЕНИЕ В

ПРОГРАММА ДЛЯ НАХОЖДЕНИЯ ИНТЕРВАЛЬНЫХ ОЦЕНОК ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ И ПРОВЕРКИ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ О НОРМАЛЬНОМ РАСПРЕДЕЛЕНИИ

```
#!/usr/bin/env python
# coding: utf-8

# In[127]:

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
sns.set_theme(style="whitegrid",   palette='deep',   con-text='notebook',
font_scale=1.3)

# ## Переменная $\nu$

# In[68]:

int_row                                                                =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/interval.csv')
N = int_row['af'].sum()
h = 37
N

# In[52]:

xv = (np.dot(int_row['avg_inter'], int_row['af'])/N).round(2)
dv = (np.dot((int_row['avg_inter']-xv)**2, int_row['af'])/N)
s = np.sqrt(dv*(N/(N-1))).round(2)
```

```
# In[53]:

k = N-1
gamma = 0.95
tg = 1.984

# In[54]:

di_a = (xv-tg*s/np.sqrt(N), xv+tg*s/np.sqrt(N))
xv
di_a

# In[58]:

q = 0.141
di_s = (s*(1-q), s*(1+q))
s
di_s

# In[241]:

alpha = 0.05

# In[242]:

df = int_row.copy().drop(['avg_inter', 'inter', 'rf'], axis=1)
df['xi'] = int_row['avg_inter']-h/2
df['xi+1'] = int_row['avg_inter']+h/2
df = df[['xi', 'xi+1', 'af']]
df = df.rename(columns={'af': 'ni'})
df.iloc[6, 0], df.iloc[6, 1] = 542, 576
df['zi'] = np.round((df['xi']-xv)/s, 2)
df['zi+1'] = np.round((df['xi+1']-xv)/s, 2)
```

```
df.loc[0, 'zi'], df.loc[6, 'zi+1'] = -np.inf, np.inf


# In[258]:

df['F(zi)'] = np.array([-5000,-4641,-3665,-1628,1064,3289,4495])/10000
df['F(zi+1)'] = np.array([-4641,-3665,-1628,1064,3289,4495,5000])/10000
df['pi'] = np.round(df['F(zi+1)'] - df['F(zi)'], 4)
df['ni*'] = np.round(df['pi']*N, 4)
df.to_csv('data/data1.csv', index=False)
df

# In[261]:

k = len(df)-3
(k, alpha)
hi_crit = 9.5
hi_nabl = np.dot((df['ni']-df['ni*'])**2, 1/df['ni*']).round(4)
(hi_nabl, hi_crit)
'True' if hi_nabl <= hi_crit else 'False'

# In[ ]:
```

ПРИЛОЖЕНИЕ Г

ПРОГРАММА ДЛЯ ЭЛЕМЕНТОВ КОРРЕЛЯЦИОННОГО АНАЛИЗА И ПРОВЕРКИ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ О РАВЕНСТВЕ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ НУЛЮ

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)


# In[2]:


df = pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/main_data.csv')
X = df['nu']
Y = df['E']


# In[3]:


h1, h2 = 37, 16.1
ivs_X = np.hstack((np.arange(min(X), max(X), h1), np.array(max(X))))
ivs_Y = np.hstack((np.arange(min(Y), max(Y), h2), np.array(max(Y))))


# ## Двумерный интервальный ряд

# In[4]:


df_int = df.copy()
```

```
df_int['intX'] = pd.cut(df_int['nu'], bins=ivs_X, right=False)
df_int['intXl'] = pd.cut(df_int['nu'], bins=ivs_X,
                         labels=[1,2,3,4,5,6,7], right=False)
df_int['intY'] = pd.cut(df_int['E'], bins=ivs_Y, right=False)
df_int['intYl'] = pd.cut(df_int['E'], bins=ivs_Y,
                         labels=[1,2,3,4,5,6,7], right=False)


# In[5]:


df_int.iloc[64, 4] = df_int.iloc[63, 4]
df_int.iloc[64, 5] = df_int.iloc[63, 5]
df_int.iloc[97, 2] = df_int.iloc[99, 2]
df_int.iloc[97, 3] = df_int.iloc[99, 3]
# df_int['intXl'].value_counts().sort_index()
# df_int['intYl'].value_counts().sort_index()
# df_int.sort_values(by=['nu'], ignore_index = True).head()
# df_int.value_counts(['intYl', 'intXl']).sort_index()


# ## Корреляционная таблица

# In[6]:


N = 104
xv = 453.71
sx = 53.79
yv = 129.98
sy = 22.06


# In[9]:


df_kor =
pd.DataFrame(columns=['yi','x1','x2','x3','x4','x5','x6','x7','Xi','yX'])
df_kor['yi'] =
[np.NaN,72.55,88.65,104.75,120.85,136.95,153.05,169.05,np.NaN,np.NaN]
df_kor['x1'] = [338.5,1,3,1,0,0,0,0,np.NaN,np.NaN]
df_kor['x2'] = [375.5,1,2,5,0,0,0,0,np.NaN,np.NaN]
df_kor['x3'] = [412.5,0,0,8,14,1,0,0,np.NaN,np.NaN]
df_kor['x4'] = [449.5,0,0,1,11,12,1,0,0,np.NaN]
```

```
df_kor['x5'] = [486.5,0,0,0,2,12,10,0,np.NaN,np.NaN]
df_kor['x6'] = [523.5,0,0,0,0,2,8,5,np.NaN,np.NaN]
df_kor['x7'] = [559,0,0,0,0,0,2,2,np.NaN,np.NaN]
# df_kor['yi']*df_kor['x1']
df_curr1 = pd.DataFrame()
df_curr2 = pd.DataFrame()
for i in range(7):
    df_curr1[i] = df_kor.iloc[0,1:8]*df_kor.iloc[i+1,1:8]
    df_kor.loc[i+1,'Xi'] =
np.dot(df_kor.iloc[0,1:8],df_kor.iloc[i+1,1:8])
    df_curr2[i] = df_kor.iloc[1:8,0]*df_kor.iloc[1:8,i+1]
    df_kor.iloc[8,i+1] = np.dot(df_kor.iloc[1:8,0],df_kor.iloc[1:8,i+1])

df_kor['yX'] = df_kor['yi']*df_kor['Xi']
df_kor.iloc[9,:] = df_kor.iloc[0,:]*df_kor.iloc[8,:]
df_kor.loc[8,'yX'] = df_kor['yX'].sum()
df_kor.loc[9,'Xi'] = df_kor.iloc[9,:].sum()

df_curr1.transpose() # желт
df_curr2 # зелен
df_kor


# ### Коэффициент корреляции

# In[14]:


r = ((df_kor.loc[8,'yX']-N*xv*yv)/(N*sx*sy)).round(4)
r


# ### Оценка кк

# In[20]:


((r-3*((1-r**2)/np.sqrt(N))).round(4),
(r+3*((1+r**2)/np.sqrt(N))).round(4))


# ## Доверительный интервал для кк

# In[31]:
```

```
z = (0.5*np.log((1+r)/(1-r))).round(3)
z


# In[30]:


sz = (1/np.sqrt(N-3)).round(4)
sz

# In[35]:

gamma = 0.95
F = gamma/2
l = 1.96
z1 = (z-l*sz).round(4)
z2 = (z+l*sz).round(4)
(z1,z2)

# In[38]:

r1 = ((np.exp(2*z1)-1)/(np.exp(2*z1)+1)).round(4)
r2 = ((np.exp(2*z2)-1)/(np.exp(2*z2)+1)).round(4)
(r1, r2)

# ## Гипотеза о значимости выборочного коэффициента корреляции

# In[50]:

K = 7
Tn = ((r*np.sqrt(N-2))/np.sqrt(1-r**2)).round(3)
tk = 1.985

# In[51]:

'True' if np.abs(Tn) <= tk else 'False'

# In[ ]:
```

ПРИЛОЖЕНИЕ Д
ПРОГРАММА ДЛЯ ЭЛЕМЕНТОВ РЕГРЕССИОННОГО АНАЛИЗА И ПОСТРОЕНИЯ ВЫБОРОЧНЫХ ПРЯМЫХ СРЕДНЕКВАДРАТИЧЕСКОЙ РЕГРЕССИИ, ПОИСКА КОРРЕЛЯЦИОННОГО ОТНОШЕНИЯ

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
# pd.set_option('display.max_columns', None)
# pd.set_option('display.max_rows', None)

# ## Выборка

# In[2]:

df = pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/main_data.csv')
X = df['nu']
Y = df['E']
int_rowX                                                              =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/interval.csv')
int_rowY                                                              =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/interval2.csv')
kor = pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/kor.csv')

# In[3]:
```

```
sns.set_theme(style="whitegrid",    palette='deep',    context='notebook',
font_scale=1.3)
ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27, as-
pect=11.7/8.27)
ax.set_axis_labels('nu', 'E')
ax.fig.suptitle('Двумерная выборка')
plt.tight_layout()
plt.savefig('pics/1.png')

# ## Прямые регрессии

# In[4]:

N = 104
xv, yv = 453.71, 129.98
sx, sy = 53.79, 22.06
r = 0.8765

# ### Прямая x на y

# In[5]:

regr_xy = lambda y: xv + r*(sx/sy)*(y-yv)

# In[6]:

ost_var_xy = (sx**2)*(1-r**2)

# ### Прямая y на x

# In[7]:

regr_yx = lambda x: yv + r*(sy/sx)*(x-xv)
```

```
# In[8]:

ost_var_yx = (sy**2)*(1-r**2)

# ### График

# In[9]:

# Регрессия Y на X
# ax = sns.lmplot(data=df, x='nu', y='E', height=8.27, aspect=11.7/8.27)


# In[10]:

ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27,
                 aspect=11.7/8.27, s=50, label='Выборка')
plt.plot(regr_xy(df['E']), df['E'], label='x(y)', zorder=0)
plt.plot(df['nu'], regr_yx(df['nu']), label='y(x)', zorder=1)
plt.scatter(xv, yv, s=60, c='crimson', zorder=2)
ax.set_axis_labels('nu', 'E')
ax.fig.suptitle('Прямые регрессии')
plt.legend()
plt.tight_layout()
plt.savefig('pics/2.png')

# In[11]:

ost_var_xy
ost_var_yx

# ## Выборочное корреляционное отношение

# ### Таблица
```

```
# In[12]:

kor.loc[1:7,'Xi'] = [np.sum(kor.iloc[i,1:8]) for i in range(1,8)]
kor.iloc[8,1:8] = [np.sum(kor.iloc[1:8,i]) for i in range(1,8)]
kor.iloc[8,8] = 104

# #### Средний x для данного y (условный выборочный x)

# In[13]:

kor.loc[1:7,'yX']
=[(np.dot(kor.iloc[0,1:8],kor.iloc[i,1:8])/kor.loc[i,'Xi']).round(2)  for
i in range(1,8)]

# #### Средний y для данного x (условный выборочный y)

# In[14]:

kor.iloc[9,1:8]
=[(np.dot(kor.iloc[1:8,0],kor.iloc[1:8,i])/kor.iloc[8,i]).round(2)  for  i
in range(1,8)]

# #### Групповая выборочная дисперсия X

# In[15]:

kor['D_grX'] = np.NaN
for i in range(1,8):
    x0_arg_kv = kor.iloc[0,1:8]**2
    dt = np.dot(x0_arg_kv,kor.iloc[i,1:8])/kor.loc[i,'Xi']
    dt -= kor.loc[i,'yX']**2
    kor.loc[i,'D_grX'] =(dt).round(2)
```

```
# #### Групповая выборочная дисперсия Y

# In[16]:
kor = kor.append(pd.Series(dtype='float64'), ignore_index=True)
for i in range(1,8):
    y0_arg_kv = kor.iloc[1:8,0]**2
    dt2 = np.dot(y0_arg_kv,kor.iloc[1:8,i])/kor.iloc[8,i]
    dt2 -= kor.iloc[9,i]**2
    kor.iloc[10,i] =(dt2).round(2)

# In[17]:

kor

# ### Дисперсии X к Y

# #### Внутригрупповая дисперсия X к Y

# In[18]:

D_vngr_xy = np.dot(kor.loc[1:7,'Xi'],kor.loc[1:7,'D_grX'])/kor.iloc[8,8]
D_vngr_xy.round(4)

# #### Межгрупповая дисперсия X к Y

# In[19]:

kv_mezh_xy = (kor.loc[1:7,'yX']-xv)**2
D_mezh_xy = np.dot(kor.loc[1:7,'Xi'],kv_mezh_xy)/kor.iloc[8,8]
D_mezh_xy.round(4)

# #### Общая дисперсия X к Y

# In[20]:
```

```
D_obsh_xy = D_vngr_xy + D_mezh_xy
D_obsh_xy.round(4)
# #### Выборочное корреляционное отношение X к Y

# In[21]:

eta_xy = np.sqrt(D_mezh_xy/D_obsh_xy)
eta_xy.round(4)
r

# ### Дисперсии Y к X

# #### Внутригрупповая дисперсия Y к X

# In[22]:

D_vngr_yx = np.dot(kor.iloc[8,1:8],kor.iloc[10,1:8])/kor.iloc[8,8]
D_vngr_yx

# #### Межгрупповая дисперсия Y к X

# In[23]:

kv_mezh_yx = (kor.iloc[9,1:8]-yv)**2
D_mezh_yx = np.dot(kor.iloc[8,1:8],kv_mezh_yx)/kor.iloc[8,8]
D_mezh_yx.round(4)

# #### Общая дисперсия Y к X

# In[24]:

D_obsh_yx = D_vngr_yx + D_mezh_yx
D_obsh_yx.round(4)
```

```
# #### Выборочное корреляционное отношение Y к X
# In[25]:

eta_yx = np.sqrt(D_mezh_yx/D_obsh_yx)
eta_yx.round(4)
r

# ## Корреляционные кривые

# In[26]:

kor

# ### Параболическая регерессия Y на X

# In[27]:

df_prbl_x = pd.DataFrame({'x': kor.iloc[0,1:8], 'n': kor.iloc[8,1:8], 'y':
kor.iloc[9,1:8]})

# In[28]:

for i in range(1,5):
    df_prbl_x[f'nx{i}'] = df_prbl_x['n']*(df_prbl_x['x']**i)
df_prbl_x['ny'] = df_prbl_x['n']*df_prbl_x['y']
df_prbl_x['nyx1'] = df_prbl_x['nx1']*df_prbl_x['y']
df_prbl_x['nyx2'] = df_prbl_x['nx2']*df_prbl_x['y']

df_prbl_xf = df_prbl_x.append(df_prbl_x.sum(), ignore_index=True)
df_prbl_xf.iloc[-1,[0,2]] = np.NaN
df_prbl_xf.to_csv('data/parabolxy.csv', index=False)
df_prbl_xf
```

```
M1 =
np.array([[df_prbl_xf.loc[7,'nx4'],df_prbl_xf.loc[7,'nx3'],df_prbl_xf.loc
[7,'nx2']],

[df_prbl_xf.loc[7,'nx3'],df_prbl_xf.loc[7,'nx2'],df_prbl_xf.loc[7,'nx1']]
,

[df_prbl_xf.loc[7,'nx2'],df_prbl_xf.loc[7,'nx1'],df_prbl_xf.loc[7,'n']]])
v1 =
np.array([df_prbl_xf.loc[7,'nyx2'],df_prbl_xf.loc[7,'nyx1'],df_prbl_xf.lo
c[7,'ny']])
a, b, c = np.linalg.solve(M1, v1)
parab_regr = lambda x: a*x*x+b*x+c
a, b, c

ax = sns.relplot(data=df, x='nu', y=parab_regr(df['nu']), kind='line',
linewidth=2.5,
                 height=8.27,   aspect=11.7/8.27,   label='y(x)',   col-
or='crimson')
plt.scatter(df['nu'], df['E'], s=50, label='Выборка')
ax.set_axis_labels('nu', 'E')
ax.fig.suptitle('Параболическая регрессия')
plt.legend()
plt.tight_layout()
plt.savefig('pics/3.png')
# ### Степенная регерессия Y на X
df_step_x = pd.DataFrame({'x': kor.iloc[0,1:8], 'n': kor.iloc[8,1:8], 'y':
kor.iloc[9,1:8]})

df_step_x['log_x'] = np.log(df_step_x['x'])
df_step_x['log_x2'] = np.log(df_step_x['x'])**2
df_step_x['log_y'] = np.log(df_step_x['y'])
df_step_x['log_x_log_y'] = df_step_x['log_x']*df_step_x['log_y']
```

```
df_step_xf = df_step_x.append(df_step_x.sum(), ignore_index=True)
df_step_xf.iloc[-1,[0,2]] = np.NaN
df_step_xf.round(3).to_csv('data/stepxy.csv', index=False)
df_step_xf
M1 = np.array([[df_step_xf.loc[7,'n'],df_step_xf.loc[7,'log_x']],
               [df_step_xf.loc[7,'log_x'],df_step_xf.loc[7,'log_x2']]])
v1                                                                    =
np.array([df_step_xf.loc[7,'log_y'],df_step_xf.loc[7,'log_x_log_y']])
a2, b2 = np.linalg.solve(M1, v1)
step_regr = lambda x: np.exp(a2)*(x**b2)
np.exp(a2), b2, a2
dfst = df.copy()
dfst['1'] = parab_regr(dfst['nu'])
dfst['2'] = step_regr(dfst['nu'])
dfstm = dfst.melt(id_vars='nu', value_vars=['1','2'])
dfstm
ax   =   sns.relplot(data=dfstm,   x='nu',   y='value',   hue='variable',
kind='line', linewidth=2.5,
                height=8.27, aspect=11.7/8.27)
plt.scatter(df['nu'], df['E'], s=50, label='Выборка')
ax.set_axis_labels('nu', 'E')
plt.legend()
ax  =  sns.relplot(data=df,  x='nu',  y=step_regr(df['nu']),  kind='line',
linewidth=2.5,
                height=8.27,   aspect=11.7/8.27,   label='y(x)',   col-
or='crimson')
plt.scatter(df['nu'], df['E'], s=50, label='Выборка')
ax.set_axis_labels('nu', 'E')
ax.fig.suptitle('Степенная регрессия')
plt.legend()
plt.tight_layout()
plt.savefig('pics/4.png')
```

ПРИЛОЖЕНИЕ Е

ПРОГРАММА ДЛЯ МЕТОДА K-MEANS

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:



import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
from scipy.spatial import distance
from sklearn.metrics import silhouette_score


# In[2]:



df0 = pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/main_data.csv')
X = df0['nu']
Y = df0['E']


# In[3]:
```

```
sns.set_theme(style="whitegrid", palette='deep', context='notebook',
font_scale=1.3)

ax = sns.relplot(data=df0, x='nu', y='E', kind='scatter', height=8.27,
aspect=11.7/8.27)

ax.set_axis_labels('nu', 'E')

ax.fig.suptitle('Двумерная выборка')

plt.tight_layout()

plt.savefig('pics/0.png')



# In[4]:



X_norm = StandardScaler().fit_transform(df0)

df = pd.DataFrame(data=X_norm, columns=['nu','E'])

df.to_csv('data/df_norm.csv', index=False)


# In[5]:


sns.set_theme(style="whitegrid", palette='deep', context='notebook',
font_scale=1.3)

ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27,
aspect=11.7/8.27)

ax.set_axis_labels('nu', 'E')

ax.fig.suptitle('Нормализованная выборка')

plt.tight_layout()

plt.savefig('pics/1.png')


# In[6]:


N = len(df)

sup_k = np.floor(np.sqrt(N/2))
```

```
sup_k


# ## Кластеры = 3


# In[7]:


clusterN = 3
k_means = KMeans(init='k-means++', n_clusters=clusterN, n_init=15)
k_means.fit(X_norm)
labels = k_means.labels_
df['cluster'] = labels
means = k_means.cluster_centers_


# In[8]:


pd.DataFrame(np.concatenate((means,
df.groupby('cluster')['nu'].count().values.reshape(-1,1)), axis=1),
             columns=['nu_mean', 'E_mean', 'num'])


# In[9]:


ax = sns.relplot(data=df, x='nu', y='E', hue='cluster', kind='scatter',
palette='deep', alpha=0.9,
                 size='cluster', height=8.27, aspect=11.7/8.27)
plt.scatter(means[:,0],means[:,1], c='crimson', s=70)


ax.set_axis_labels('nu', 'E')
ax.fig.suptitle('Кластеры 3')
plt.tight_layout()
# plt.savefig('pics/2.png')
df = df.drop('cluster', axis=1)
```

```
# ## Алгоритм


# In[10]:


def sc_plots(data, means, Ncl, step):
    if Ncl > 6:
        ax = sns.relplot(data=data, x='nu', y='E', hue='cluster',
kind='scatter', alpha=0.9,
                 size='cluster', height=8.27, aspect=11.7/8.27)
        plt.scatter(means[:,0],means[:,1], c='crimson',
s=np.linspace(50,100,Ncl))
    else:
        ax = sns.relplot(data=data, x='nu', y='E', hue='cluster',
kind='scatter', palette='deep', alpha=0.9,
                 size='cluster', height=8.27, aspect=11.7/8.27)
        plt.scatter(means[:,0],means[:,1], c='k',
s=np.linspace(50,100,Ncl))


    ax.set_axis_labels('nu', 'E')
    ax.fig.suptitle(f'Кластеры = {Ncl}, шаг {step}')
    plt.tight_layout()
    plt.savefig(f'pics/2_{Ncl}.png')
    plt.close()


# In[11]:


def nearest_center(data, cts):
    distl = np.array([], dtype=np.float64)
    for i in cts:
        distl = np.append(distl, np.linalg.norm(i[:-1]-data)) # евклидово
расстояние
    min_dist = np.argmin(distl)
    return min_dist
```

```
# In[12]:


def Fs(data):
    curr_data = data.copy()
    cts = curr_data.groupby('cluster').mean()
    F1,F2,F3 = 0,0,0

    # F1 - сумма кв. расст. точек до центров соотв. кластеров
    for i in range(len(curr_data)):
        dist_F1 = np.linalg.norm(curr_data.iloc[i,:-1].values-
cts.values[curr_data.iloc[i,2]])
        F1 += dist_F1**2

    # F2 - сумма кв. расст. до всех точек соотв. кластеров
    for i in range(len(cts)):
        coords = curr_data[curr_data['cluster']==i].iloc[:,:2].values
        dist_F2 = distance.cdist(coords, coords, 'euclidean')
        F2 += (np.triu(dist_F2,0)**2).sum()

    # F3 - сумма внутрикластерных дисперсий
    F3 = curr_data.groupby('cluster').var().values.sum()

    return F1,F2,F3

# In[13]:


def custKM(dataf, n_clusters, chng_ctr=1, max_iter=30, tol=0.01):
    data = dataf.copy()
    centers = data.sample(n_clusters) # рандомные центры
    data['cluster'] = -1 # нет принадлежности кластерам
```

```
    cts = np.array([], dtype=np.float64)
    F1,F2,F3 = 0,0,0
    df_Fs = pd.DataFrame(columns=['F1', 'F2', 'F3'])

    for i in range(n_clusters):
        data.loc[centers.index[i],'cluster'] = i # кластеры для центров
        cts = np.append(cts, [data.loc[centers.index[i]].values])
    centers = cts.reshape((n_clusters,3))

    for j in range(max_iter):
        for i in range(len(data)): # ближ. центр для каждой точки
            curr_clust = nearest_center(data.iloc[i,:-1].values, centers)
            data.loc[i,'cluster'] = curr_clust # соотносим кластер
            if chng_ctr: # пересчет центра при новой точке
                centers[curr_clust][:2] =
data[data['cluster']==curr_clust].iloc[:,:2].mean()

        if chng_ctr == 0: # пересчет центра на каждой итерации
            for i in range(n_clusters):
                centers[i][:2] =
data[data['cluster']==i].iloc[:,:2].mean()

        cur_F1,cur_F2,cur_F3 = Fs(data) # функционалы
        df_Fs = df_Fs.append({'F1':cur_F1,'F2':cur_F2,'F3':cur_F3},
ignore_index=True)

        if np.abs(F1-cur_F1) < tol:
            data['cluster'].astype('int')
            sc_plots(data, centers, n_clusters, j+1)
            break
        F1,F2,F3 = cur_F1,cur_F2,cur_F3
        data['cluster'] = -1
```

```
    df_ctrs = pd.DataFrame(np.concatenate((centers[:,:2],

data.groupby('cluster')['nu'].count().values.reshape(-1,1)), axis=1),
                           columns=['nu_mean', 'E_mean', 'num'])
    silhouette_avg = silhouette_score(data.values[:,:2],
data.values[:,2])


    return df_Fs, df_ctrs, silhouette_avg


# In[71]:


sse = []
sils = []
for i in range(2,8):
    F, ctrs, sil = custKM(df, n_clusters=i, chng_ctr=1)
#     print(len(F))
    sse.append(F.iloc[-1,0])
    sils.append(sil)
    F.round(3).to_csv(f'data/Fs_{i}c.csv', index=False)
    ctrs.round(4).to_csv(f'data/centers_{i}c.csv', index=False)


# ### Локоть


# In[53]:


# sse = [76.855,43.857,35.379,24.084,15.096,14.481]
ax = sns.relplot(x=range(2,8), y=sse, kind='line', height=8.27,
aspect=11.7/8.27)
ax.set(ylim=[10,80])
ax.set_axis_labels('Кластеры', '$F_1$')
ax.fig.suptitle('Метод локтя')
```

```
plt.tight_layout()

plt.savefig(f'pics/elbow.png')

plt.show()


# ### Силуэт


# In[54]:


ax = sns.relplot(x=range(2,8), y=sils, kind='line', height=8.27,
aspect=11.7/8.27)

ax.set_axis_labels('Кластеры', 'Коэффициент')

ax.fig.suptitle('Коэффициент силуэт')

plt.tight_layout()

plt.savefig(f'pics/silhouette.png')

plt.show()
```

ПРИЛОЖЕНИЕ Ж

ПРОГРАММА ДЛЯ МЕТОДА ПОИСКА СГУЩЕНИЙ

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:



import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
from scipy.spatial import distance
import functools



# In[2]:



df0 = pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/main_data.csv')
X = df0['nu']
Y = df0['E']



# In[3]:



sns.set_theme(style="whitegrid", palette='deep', context='notebook',
font_scale=1.3)
ax = sns.relplot(data=df0, x='nu', y='E', kind='scatter', height=8.27,
as-pect=11.7/8.27)
```

```
ax.set_axis_labels('nu', 'E')
ax.fig.suptitle('Двумерная выборка')
plt.tight_layout()
plt.savefig('pics/0.png')


# In[4]:


X_norm = StandardScaler().fit_transform(df0)
df = pd.DataFrame(data=X_norm, columns=['nu','E'])
df.to_csv('data/df_norm.csv', index=False)


# In[5]:


sns.set_theme(style="whitegrid", palette='deep', context='notebook',
font_scale=1.3)
ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27, as-
pect=11.7/8.27)
ax.set_axis_labels('nu', 'E')
ax.fig.suptitle('Нормализованная выборка')
plt.tight_layout()
plt.savefig('pics/1.png')


# ## Алгоритм

# In[6]:


def sc_plots(data, center, R, step, itera):
```

```
    ax = sns.relplot(data=data, x='nu', y='E', hue='cluster',
kind='scatter', palette='deep',
                     alpha=0.9, height=8.27, aspect=11.7/8.27)
    for j in [center]:
        plt.scatter(j[0],j[1], c='k', s=70)
#     print(center.values[:2])
#     print(data[data['cluster']!=1]['nu'].count())

    circle = np.array([], dtype=np.float64)
    for i in data[data['cluster']!=-1].values:
        circle = np.append(circle, np.linalg.norm(i[:-1]-
center.values[:2]))

    plt.scatter(j[0], j[1], linewidths=1, facecolors='crimson', edgecol-
ors='crimson', s=max(circle)*2*35000, alpha=0.1)

    ax.set_axis_labels('nu', 'E')
    ax.fig.suptitle(f'Кластер {itera}, шаг {step}; центр = ({cen-
ter.values[0].round(4)}; {center.values[1].round(4)}); N = {da-
ta[data["cluster"]!=-1]["nu"].count()}')
    ax.set(xlim=[-3.5,3.5], ylim=[-3.5,3.5])
    plt.tight_layout()
    plt.savefig(f'pics/{itera}_{step}.png')
    plt.show()


# In[7]:


def Fs(data):
    curr_data = data.copy()
    cts = curr_data.groupby('cluster').mean()
    F1,F2,F3 = 0,0,0
```

```
    # F1 - сумма кв. расст. точек до центров соотв. кластеров
    for i in range(len(curr_data)):
        dist_F1 = np.linalg.norm(curr_data.iloc[i,:-1].values-
cts.values[curr_data.iloc[i,2]-1])
        F1 += dist_F1**2

    # F2 - сумма кв. расст. до всех точек соотв. кластеров
    for i in range(1,len(cts)+1):
        coords = curr_data[curr_data['cluster']==i].iloc[:,:2].values
        dist_F2 = distance.cdist(coords, coords, 'euclidean')
        F2 += (np.triu(dist_F2,0)**2).sum()

    # F3 - сумма внутрикластерных дисперсий
    F3 =
curr_data.groupby('cluster').var().values.sum(where=~np.isnan(curr_data.g
roupby('cluster').var().values), initial=0)

    return F1,F2,F3


# In[8]:


def custFE(cur_data, R, itera, plots=1, max_iter=20):
    cur_dist = np.array([], dtype=np.float64)
    data = cur_data.copy()
    coords = data.values

    # расстояние между объектами
    dist = distance.cdist(coords, coords, 'euclidean')
    data['cluster'] = -1

    # сколько объектов с растоянием < R для каждого объекта
    for i in dist:
```

```
        cur_dist = np.append(cur_dist, len(i[np.where((i>=0) & (i<=R))]))

    # индекс центра
    center_ind = np.argmax(cur_dist)
    # индексы объектов с расстоянием < R до центра
    cluster_ind = np.where((dist[np.argmax(cur_dist)]>=0) &
                           (dist[np.argmax(cur_dist)]<=R))
    data.iloc[cluster_ind[0],2] = itera
    data.iloc[center_ind,2] = itera
    if plots == 1:
        sc_plots(data, data.iloc[center_ind], R, 1, itera)
    cur_center = data.iloc[center_ind]

    for it in range(max_iter):
        distl = np.array([], dtype=np.float64)
        # новый центр тяжетси
        center = data[data['cluster']==itera].mean()
        data['cluster'] = -1

        # расстояния до нового центра
        for i in data.iloc[:,:2].values:
            distl = np.append(distl, np.linalg.norm(center[:-1].values-
i))
        cluster_ind = np.where((distl>=0) & (distl<=R))

        data.iloc[cluster_ind[0],2] = itera

        if functools.reduce(lambda x, y : x and y, map(lambda p, q: p ==
q,center.values,cur_center.values), True):
            break
        if plots == 1:
            sc_plots(data, center, R, it+2, itera)
        cur_center = center
```

```
    # график
    if plots == 0:
        sc_plots(data, center, R, 'последний', itera)

    return data[data['cluster']==-1], data, np.array(center.values[:2])


# ## Основа

# In[9]:


coords = df.values
dist = np.triu(distance.cdist(coords, coords, 'euclidean'), 0)
rmin = np.amin(dist, where=dist!=0, initial=10)
rmax = np.amax(dist)
rmin.round(4), rmax.round(4)


# In[22]:


upd_df = df.copy()
it = 1
radius = 1.13
df['cluster'] = -1
ctrs = np.array([], dtype=np.float64)


# In[23]:


while len(upd_df):
    upd_df, main, ctr = custFE(upd_df, radius, it, 2)
```

```
    ctrs = np.append(ctrs, [ctr])
    it += 1
    df.loc[main[main['cluster']!=-1].index, :] =
main.loc[main[main['cluster']!=-1].index, :]
df.to_csv('data/result.csv', index=False)


# ### Финальное разбиение

# In[24]:


F1, F2, F3 = Fs(df)
F1, F2, F3


# In[18]:


F1, F2, F3 = Fs(df)
F1, F2, F3


# In[21]:


F1, F2, F3 = Fs(df)
F1, F2, F3


# In[15]:
```

```
ax = sns.relplot(data=df, x='nu', y='E', hue='cluster', kind='scatter',
pal-ette='deep', alpha=0.9,
                 size='cluster', height=8.27, aspect=11.7/8.27)
ctrs = ctrs.reshape((-1,2))
for i in ctrs:
    plt.scatter(i[0], i[1], c='k', s=60)
ax.set_axis_labels('nu', 'E')
ax.fig.suptitle(f'Кластеры {len(df["cluster"].unique())}')
plt.tight_layout()
plt.savefig('pics/result.png')
plt.show()
```