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 _____ Бабенко Н.С.

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 _____ Сахаров В.М.

Преподаватель _____ Середа А.-В.И.

Санкт-Петербург

2022

ЗАДАНИЕ

НА КУРСОВУЮ РАБОТУ

Студент Бабенко Н.С.

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: 09.04.2022

Дата защиты реферата: 09.04.2022

Студент _____ Бабенко Н.С.

Преподаватель _____ Середа А.-В.И.

ЗАДАНИЕ

НА КУРСОВУЮ РАБОТУ

Студент Сахаров В.М.

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: 09.04.2022

Дата защиты реферата: 09.04.2022

Студент _____ Сахаров В.М.

Преподаватель _____ Середа А.-В.И.

АННОТАЦИЯ

В данной курсовой работе исследуется двумерная выборка, состоящая из данных наблюдений относительно объемного веса nu $\left(\text{г}/_{\text{см}^3}\right)$ при влажности 10% и модуля упругости E $\left(\text{кг}/_{\text{см}^2}\right)$ при сжатии вдоль волокон древесины резонансной ели. Исследование состоит из таких этапов, как выравнивание статистических рядов, нахождение точечных и интервальных статистических оценок, построение регрессионных кривых, проверка статистических гипотез о нормальном распределении выборки и о равенстве коэффициента корреляции нулю, а также корреляционный анализ, регрессионный анализ и кластерный анализ, представленный методами k-средних и поиска сгущений.

ABSTRACT

In this course work, a two-dimensional sample is investigated, consisting of observational data on bulk density nu 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 consists of such stages as alignment of statistical series, finding point and interval statistical estimates, constructing regression curves, testing statistical hypotheses about the normal distribution of the sample and the equality of the correlation coefficient to zero, as well as correlation analysis, regression analysis and cluster analysis, represented by methods k -averages and search for condensations.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В ходе курсовой работы необходимо ознакомиться с основными правилами формирования выборки и подготовки выборочных данных к статистическому анализу.

Получить практические навыки нахождения точечных статистических оценок и вычисления интервальных статистических оценок параметров распределения выборочных данных и проверки статистических гипотез.

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

Необходимо освоить и реализовать методы кластерного анализа, такие как, метод k-средних 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:

$$x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(m)}$$

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.

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:

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

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

Интервальной оценкой математического ожидания при неизвестном среднем квадратическом отклонении σ генеральной совокупности служит доверительный интервал:

$$\overline{x_{\text{в}}} - \frac{S}{\sqrt{n}} t_\gamma \leq \alpha \leq \overline{x_{\text{в}}} + \frac{S}{\sqrt{n}} t_\gamma,$$

$\overline{x_{\text{в}}}$ – статистическая оценка математического ожидания

S – исправленное СКВО

n – объём выборки

t_γ – из таблицы

Доверительный интервал для оценки СКВО:

$$S(1-q) \leq \sigma \leq S(1+q),$$

S – исправленное СКВО

q – из таблицы

Критерий Пирсона, или критерий χ^2(Хи-квадрат), 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 ($\chi^2_{\text{наб}} > \chi^2_{\text{крит}}$) – гипотезу отвергают.

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

В качестве генеральной совокупности были выбраны данные 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. Далее была сформирована репрезентативная выборка заданного объема из имеющейся генеральной совокупности экспериментальных данных при помощи библиотеки scikit-learn. Объём выборки: 100. Выборка представлена в таблице 1.

Таблица 1

№	nu	E	№	nu	E	№	nu	E	№	nu	E	№	nu	E
1	481	135.2	***21***	418	131.4	***41***	513	159.3	***61***	450	122.3	***81***	475	143.6
2	445	124.7	***22***	378	103.8	***42***	489	149.8	***62***	468	128.9	***82***	518	144.4
3	550	147.9	***23***	521	154.9	***43***	474	132.5	***63***	441	122.8	***83***	566	175.7
4	465	140.9	***24***	394	117.7	***44***	379	94.6	***64***	460	140.7	***84***	464	131.3
5	566	168.5	***25***	504	145.3	***45***	472	135.6	***65***	480	117.7	***85***	394	112.1
6	497	147.3	***26***	440	126.7	***46***	544	169.6	***66***	429	112.9	***86***	480	146.1
7	478	136.6	***27***	465	114.8	***47***	507	142.4	***67***	457	126.4	***87***	321	86.1
8	521	139.6	***28***	418	109.3	***48***	409	116.7	***68***	464	143.2	***88***	502	132.5
9	352	84.9	***29***	418	118.6	***49***	498	164.0	***69***	431	125.0	***89***	460	122.4
10	422	117.9	***30***	465	127.7	***50***	468	142.0	***70***	424	119.0	***90***	458	104.7
11	506	153.5	***31***	447	117.5	***51***	593	187.4	***71***	502	137.2	***91***	362	111.7
12	443	122.9	***32***	433	131.5	***52***	523	152.6	***72***	465	140.7	***92***	503	148.5
13	434	140.4	***33***	460	136.8	***53***	478	126.6	***73***	492	137.5	***93***	446	144.0
14	422	108.6	***34***	382	98.8	***54***	438	122.2	***74***	446	128.4	***94***	421	115.1
15	569	157.4	***35***	532	160.6	***55***	423	115.9	***75***	482	136.4	***95***	407	110.5
16	439	119.2	***36***	482	148.2	***56***	408	110.0	***76***	510	140.6	***96***	448	137.7
17	437	129.4	***37***	472	122.6	***57***	386	105.8	***77***	434	122.3	***97***	490	139.9
18	461	138.6	***38***	532	158.7	***58***	428	130.3	***78***	623	195.7	***98***	482	141.2
19	351	89.0	***39***	473	137.9	***59***	560	169.8	***79***	468	141.2	***99***	463	129.2
20	390	91.4	***40***	525	148.3	***60***	483	130.3	***80***	471	119.7	***100***	459	145.4

Выборка относительно переменной nu представлена в таблице 2.

Таблица 2

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	481	***21***	418	***41***	513	***61***	450	***81***	475
2	445	***22***	378	***42***	489	***62***	468	***82***	518
3	550	***23***	521	***43***	474	***63***	441	***83***	566
4	465	***24***	394	***44***	379	***64***	460	***84***	464
5	566	***25***	504	***45***	472	***65***	480	***85***	394

6	497	***26***	440	***46***	544	***66***	429	***86***	480
7	478	***27***	465	***47***	507	***67***	457	***87***	321
8	521	***28***	418	***48***	409	***68***	464	***88***	502
9	352	***29***	418	***49***	498	***69***	431	***89***	460
10	422	***30***	465	***50***	468	***70***	424	***90***	458
11	506	***31***	447	***51***	593	***71***	502	***91***	362
12	443	***32***	433	***52***	523	***72***	465	***92***	503
13	434	***33***	460	***53***	478	***73***	492	***93***	446
14	422	***34***	382	***54***	438	***74***	446	***94***	421
15	569	***35***	532	***55***	423	***75***	482	***95***	407
16	439	***36***	482	***56***	408	***76***	510	***96***	448
17	437	***37***	472	***57***	386	***77***	434	***97***	490
18	461	***38***	532	***58***	428	***78***	623	***98***	482
19	351	***39***	473	***59***	560	***79***	468	***99***	463
20	390	***40***	525	***60***	483	***80***	471	***100***	459

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	321	***21***	423	***41***	457	***61***	473	***81***	504
2	351	***22***	424	***42***	458	***62***	474	***82***	506
3	352	***23***	428	***43***	459	***63***	475	***83***	507
4	362	***24***	429	***44***	460	***64***	478	***84***	510
5	378	***25***	431	***45***	460	***65***	478	***85***	513
6	379	***26***	433	***46***	460	***66***	480	***86***	518
7	382	***27***	434	***47***	461	***67***	480	***87***	521
8	386	***28***	434	***48***	463	***68***	481	***88***	521
9	390	***29***	437	***49***	464	***69***	482	***89***	523
10	394	***30***	438	***50***	464	***70***	482	***90***	525
11	394	***31***	439	***51***	465	***71***	482	***91***	532
12	407	***32***	440	***52***	465	***72***	483	***92***	532
13	408	***33***	441	***53***	465	***73***	489	***93***	544
14	409	***34***	443	***54***	465	***74***	490	***94***	550
15	418	***35***	445	***55***	468	***75***	492	***95***	560
16	418	***36***	446	***56***	468	***76***	497	***96***	566
17	418	***37***	446	***57***	468	***77***	498	***97***	566

18	421	***38***	447	***58***	471	***78***	502	***98***	569
19	422	***39***	448	***59***	472	***79***	502	***99***	593
20	422	***40***	450	***60***	472	***80***	503	***100***	623

Из таблицы 3 видно, что наименьшее значение в выборке $x_{min} = 321$, а наибольшее значение $x_{max} = 623$.

В таблице 4 представлено преобразование полученной выборки в вариационный ряд с абсолютными $\boldsymbol{n_i}$ и относительными $\boldsymbol{\overline{n_i}}$ частотами соответственно.

Таблица 4

$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\boldsymbol{\overline{n_i}}$	$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\boldsymbol{\overline{n_i}}$	$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\boldsymbol{\overline{n_i}}$	$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\boldsymbol{\overline{n_i}}$
1	321	1	0.01	***26***	439	1	0.01	***51***	481	1	0.01	***76***	593	1	0.01
2	351	1	0.01	***27***	440	1	0.01	***52***	482	3	0.03	***77***	623	1	0.01
3	352	1	0.01	***28***	441	1	0.01	***53***	483	1	0.01				
4	362	1	0.01	***29***	443	1	0.01	***54***	489	1	0.01				
5	378	1	0.01	***30***	445	1	0.01	***55***	490	1	0.01				
6	379	1	0.01	***31***	446	2	0.02	***56***	492	1	0.01				
7	382	1	0.01	***32***	447	1	0.01	***57***	497	1	0.01				
8	386	1	0.01	***33***	448	1	0.01	***58***	498	1	0.01				
9	390	1	0.01	***34***	450	1	0.01	***59***	502	2	0.02				
10	394	2	0.02	***35***	457	1	0.01	***60***	503	1	0.01				
11	407	1	0.01	***36***	458	1	0.01	***61***	504	1	0.01				
12	408	1	0.01	***37***	459	1	0.01	***62***	506	1	0.01				
13	409	1	0.01	***38***	460	3	0.03	***63***	507	1	0.01				
14	418	3	0.03	***39***	461	1	0.01	***64***	510	1	0.01				
15	421	1	0.01	***40***	463	1	0.01	***65***	513	1	0.01				
16	422	2	0.02	***41***	464	2	0.02	***66***	518	1	0.01				
17	423	1	0.01	***42***	465	4	0.04	***67***	521	2	0.02				
18	424	1	0.01	***43***	468	3	0.03	***68***	523	1	0.01				
19	428	1	0.01	***44***	471	1	0.01	***69***	525	1	0.01				
20	429	1	0.01	***45***	472	2	0.02	***70***	532	2	0.02				
21	431	1	0.01	***46***	473	1	0.01	***71***	544	1	0.01				
22	433	1	0.01	***47***	474	1	0.01	***72***	550	1	0.01				
23	434	2	0.02	***48***	475	1	0.01	***73***	560	1	0.01				
24	437	1	0.01	***49***	478	2	0.02	***74***	566	2	0.02				
25	438	1	0.01	***50***	480	2	0.02	***75***	569	1	0.01				

Из таблицы 4 можно увидеть моду выборки, которой является варианта $x_{42} = 465$ с абсолютной частотой равной 4.

Чтобы преобразовать вариационный ряд в интервальный ряд сначала нужно вычислить количество интервалов разбиения с помощью формулы Стерджесса:

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

Далее вычислена ширина интервала с помощью формулы:

$$h = \frac{x_{max} - x_{min}}{k} = \frac{623 - 321}{7} \approx 44$$

В таблице 5 представлен полученный интервальный ряд.

Таблица 5

Границы интервалов	*Середины интервалов*	*Абсолютная частота*	*Относительная частота*
[321, 365)	343	4	0.04
[365, 409)	387	9	0.09
[409, 453)	431	27	0.27
[453, 497)	475	35	0.35
[497, 541)	519	17	0.17
[541, 585)	563	6	0.06
[585, 623)	604	2	0.02

Далее для интервального ряда абсолютных частот были построены полигон и гистограмма.

Полигон представлен на рис. 1.2.1.

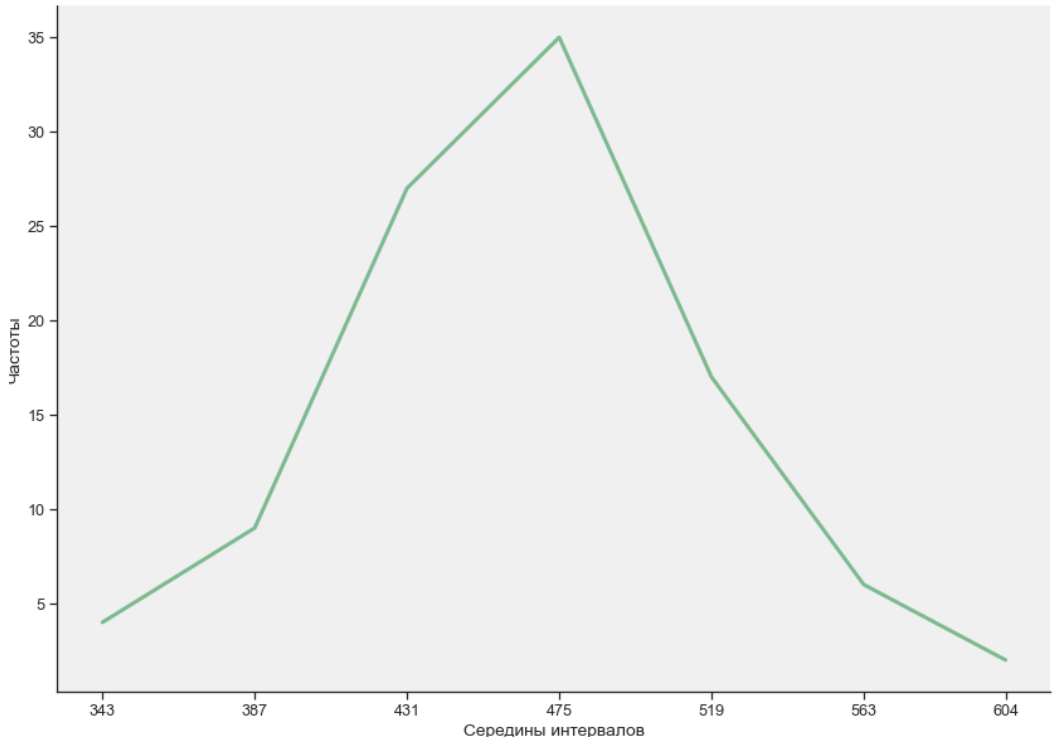

Рисунок 1.2.1 – Полигон для абсолютных частот

Полигон представляет собой ломаную, соединяющую точки, соответствующие срединным значениям интервалов и абсолютным частотам этих интервалов. Видно, что на пике значение равно 35, что сходится с данными таблицы 5.

Гистограмма, представлена на рис. 1.2.2.

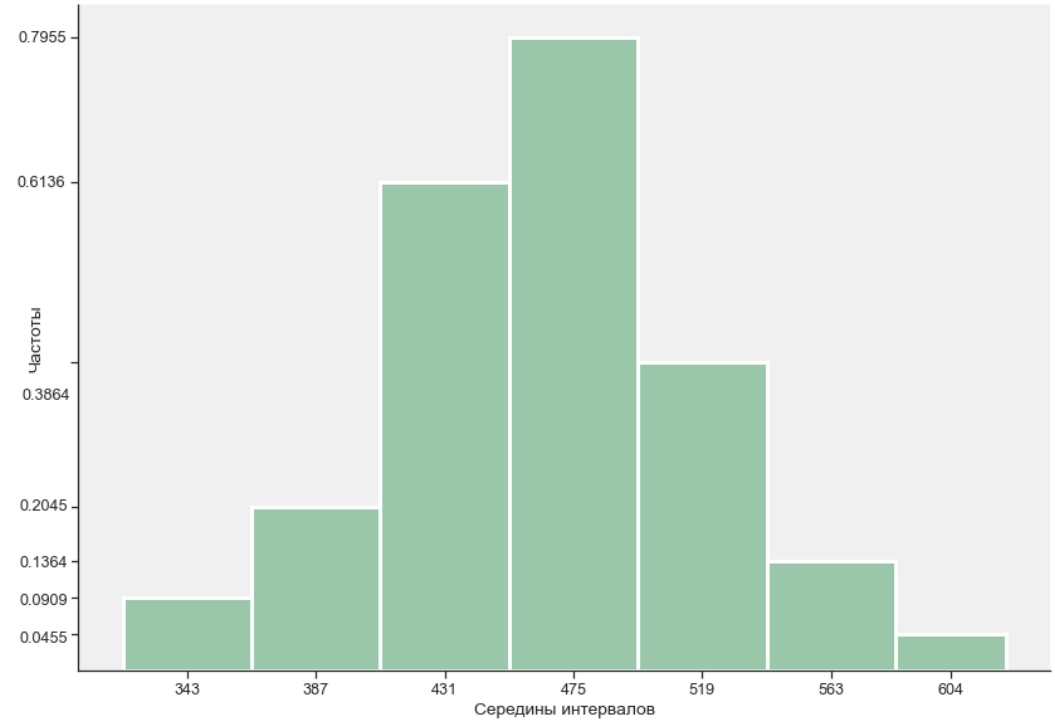

Рисунок 1.2.2 – Гистограмма для абсолютных частот

Гистограмма представляет собой фигуру, состоящую из прямоугольников, основания которых это длина интервалов h, а высота равна отношению частоты к длине интервала, то есть площадь прямоугольника обозначает частоту интервала.

Графики для интервального ряда относительных частот представлены ниже.

Полигон для относительных частот представлен на рис. 1.2.3.

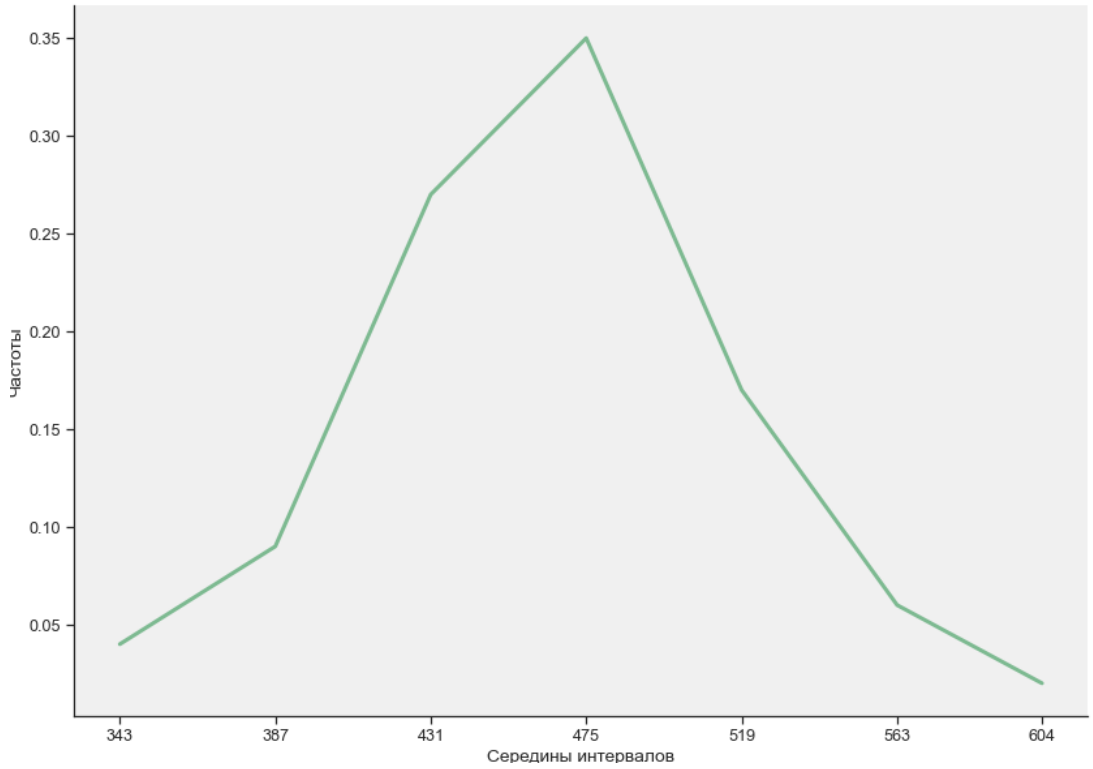

Рисунок 1.2.3 – Полигон для относительных частот

Гистограмма для относительных частот, представлена на рис. 1.2.4.

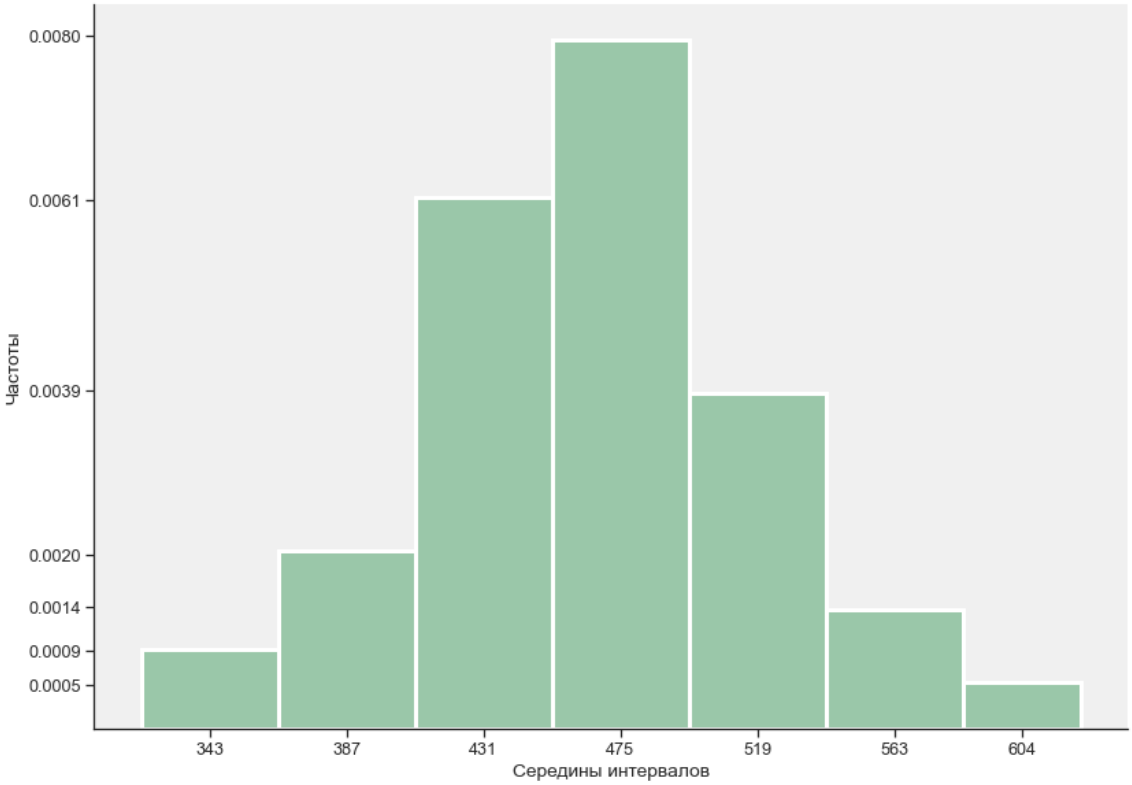

Рисунок 1.2.4 – Гистограмма для относительных частот

Эмпирическая функция распределения, построенная применительно к интервальному ряду для относительных частот представлен на рис. 5.

Функция распределения:

$$F(343) = 0$$

$$F(387) = 0.04$$

$$F(431) = 0.13$$

$$F(475) = 0.40$$

$$F(519) = 0.75$$

$$F(563) = 0.92$$

$$F(604) = 0.98$$

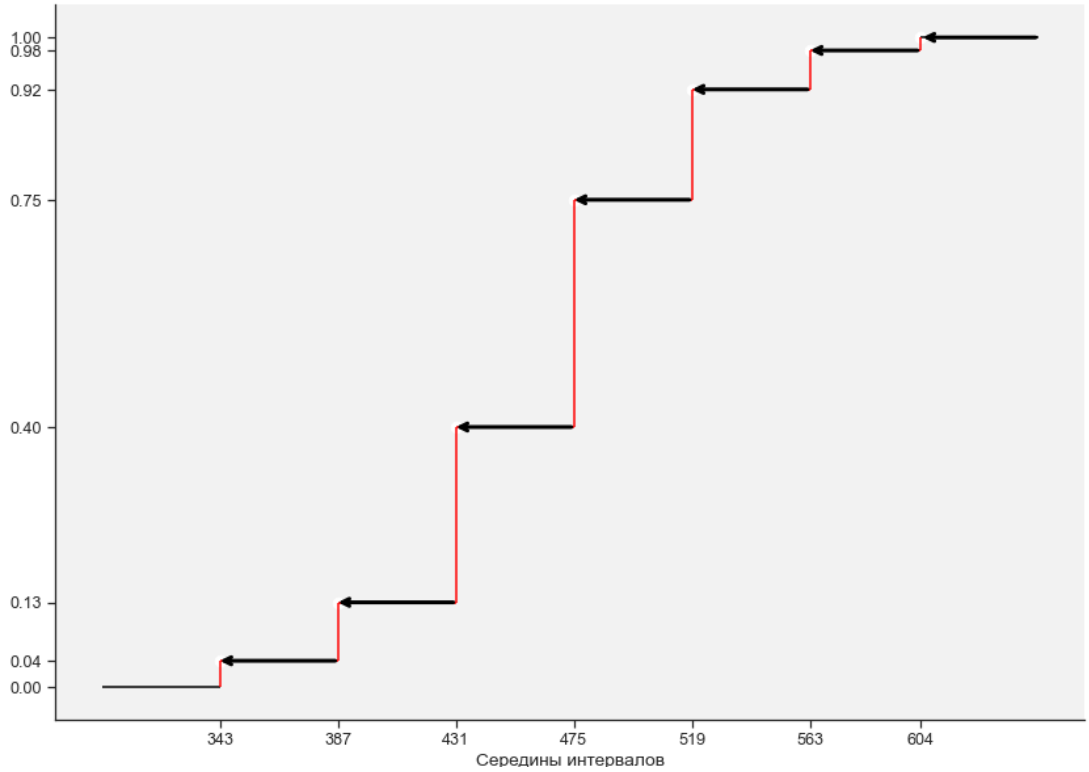

Рисунок 1.2.5 – График эмпирической функции распределения

1.3. Нахождение точечных оценок параметров распределения.

- Переменная nu

Интервальный ряд для переменной nu и с посчитанными накопленными частотами представлен в таблице 5.

Таблица 5

Границы интервалов	Середины интервалов	Абсолютная частота	Относительная частота	Накопленная частота
[321, 365)	343	4	0.04	0.04
[365, 409)	387	9	0.09	0.13
[409, 453)	431	27	0.27	0.4
[453, 497)	475	35	0.35	0.75
[497, 541)	519	17	0.17	0.92
[541, 585)	563	6	0.06	0.98
[585, 623)	604	2	0.02	1

Объем выборки $n = 100$

Условные варианты можно найти с помощью формулы:

$$u_j = \frac{x_j - C}{h}$$

Условные моменты k-го порядка можно найти по формуле:

$$\overline{M_k^*} = \frac{1}{N}\sum n_j u_j^k$$

Результаты вычислений представлены в табл. 6.

Таблица 6

$\boldsymbol{\upsilon}$	$\mathbf{n}$	$\mathbf{u}$	$\mathbf{n} * \mathbf{u}$	$\mathbf{n} * \mathbf{u^2}$	$\mathbf{n} * \mathbf{u^3}$	$\mathbf{n} * \mathbf{u^4}$	$\mathbf{n} * (\mathbf{u}+\mathbf{1})^4$
343	0.04	-3	-0.12	0.36	-1.08	3.24	0.64
387	0.09	-2	-0.18	0.36	-0.72	1.44	0.09
431	0.27	-1	-0.27	0.27	-0.27	0.27	0.0
475	0.35	0	0.0	0.0	0.0	0.0	0.35
519	0.17	1	0.17	0.17	0.17	0.17	2.72
563	0.06	2	0.12	0.24	0.48	0.96	4.86
604	0.02	3	0.06	0.18	0.54	1.62	5.12
$\sum$	1	–	-0.22	1.58	-0.88	7.7	13.78

Проверить вычисления можно с помощью последнего столбца:

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

$$= 7.7 + 4 * -0.88 + 6 * 1.58 + 4 * -0.22 + 1 = 13.78$$

Число совпадает с суммой элементов последнего столбца, следовательно вычисления правильные.

Был посчитан первый начальный эмпирический момент с помощью условных вариант, который обозначает выборочное среднее:

$$\overline{x_в} = \overline{M_1} = \overline{M_1^*}h + C = 465.32$$

Также был посчитан второй центральный эмпирический момент с помощью условных вариант, который обозначает выборочную дисперсию:

$$D_в = \overline{m_2} = \left(\overline{M_2^*} - \left(\overline{M_1^*}\right)^2\right)h^2 = 2965.1776$$

Далее были найдены выборочное среднее и дисперсия с помощью стандартных формул.

$$\bar{x_{\text{в}}} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = 465.26$$

$$D_{\text{в}} = \frac{1}{N}\sum_{i=1}^{k} (x_i - \bar{x_{\text{в}}})^2 n_i = 2968.35$$

Исправленная оценка дисперсии:

$$s^2 = \frac{N}{N-1} D_{\text{в}} = \frac{100}{99} * 2865.503 = 2978.31$$

Были найдены статистические оценки СКО:

$$\sigma_{\text{в}} = \sqrt{D_{\text{В}}} = \sqrt{2968.35} = 54.3$$

$$s = \sqrt{s^2} = \sqrt{2978.31} = 54.72$$

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

Были найдены статистические оценки коэффициентов асимметрии и эксцесса:

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

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

$$\overline{m_3} = \left(\overline{M_3^*} - 3\overline{M_2^*}\,\overline{M_1^*} + 2\left(\overline{M_1^*}\right)^3\right) h^3 = 12053.88$$

$$\overline{m_4} = \left(\overline{M_4^*} - 4\overline{M_3^*}\,\overline{M_1^*} + 6\overline{M_2^*}\left(\overline{M_1^*}\right)^2 + 2\left(\overline{M_1^*}\right)^4\right) h^4 = 27651219.62$$

Статистическая оценка коэффициента асимметрии:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3} = 0.00000045$$

Статистическая оценка коэффициента эксцесса:

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

Коэффициент асимметрии положительный, следовательно, это правосторонняя асимметрия, и $\bar{x_{\text{в}}} > M_o$, но полученный коэффициент незначительный и скос распределения небольшой. Коэффициент эксцесса же

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

Вычислим моду и медиану заданного распределения для интервального ряда. Мода заданного распределения:

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

$$M_o = 453 + 44\frac{35 - 27}{(35 - 27) + (35 - 17)} = 464.53$$

Медиана заданного распределения:

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

$$M_e = 453 + \frac{0.5 * 100 - 40}{35} 44 = 463.57$$

- Переменная E

Интервальный ряд из первой лабораторной работы для переменной E и с посчитанными накопленными частотами представлен в таблице 7.

Таблица 7

Границы интервалов	**Середины интервалов**	**Абсолютная частота**	**Относительная частота**	**Накопленная частота**
[84.9, 100.9)	92	6	0.06	0.06
[100.9, 116.9)	108	14	0.14	0.2
[116.9, 132.9)	124	32	0.32	0.52
[132.9, 148.9)	140	33	0.33	0.85
[148.9, 164.9)	156	9	0.09	0.94
[164.9, 180.9)	172	4	0.04	0.98
[180.9, 195.7)	188	2	0.02	1

Результаты вычислений условных моментов представлены в табл. 8.

Таблица 8

$\boldsymbol{\upsilon}$	$\mathbf{n}$	$\mathbf{u}$	$\mathbf{n} * \mathbf{u}$	$\mathbf{n} * \mathbf{u^2}$	$\mathbf{n} * \mathbf{u^3}$	$\mathbf{n} * \mathbf{u^4}$	$\mathbf{n} * (\mathbf{u} + \mathbf{1})^4$
92	0.06	-3	-0.18	0.54	-1.62	4.86	0.96

108	0.14	-2	-0.28	0.56	-1.12	2.24	0.14
124	0.32	-1	-0.32	0.32	-0.32	0.32	0.0
140	0.33	0	0.0	0.0	0.0	0.0	0.33
156	0.09	1	0.09	0.09	0.09	0.09	1.44
172	0.04	2	0.08	0.16	0.32	0.64	3.24
188	0.02	3	0.06	0.18	0.54	1.62	5.12
∑	1	–	-0.55	1.85	-2.11	9.77	11.23

Проверим вычисления с помощью последнего столбца:

$$\sum n_j * u_j^4 + 4 * \sum n_j * u_j^3 + 6 * \sum n_j * u_j^2 + 4 * \sum n_j * u_j + 1 =$$
$$= 9.77 + 4 * -2.11 + 6 * 1.85 + 4 * -0.55 + 1 = 11.23$$

Число совпадает с суммой элементов последнего столбца, следовательно вычисления правильные.

Был посчитан первый начальный эмпирический момент с помощью условных вариант, который обозначает выборочное среднее:

$$\overline{x_в} = \overline{M_1} = \overline{M_1^*}h + C = 133.7$$

Также был посчитан второй центральный эмпирический момент с помощью условных вариант, который обозначает выборочную дисперсию:

$$D_в = \overline{m_2} = \left(\overline{M_2^*} - \left(\overline{M_1^*}\right)^2\right)h^2 = 396.16$$

Далее были найдены выборочное среднее и дисперсия с помощью стандартных формул.

$$\overline{x_в} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = 133.8$$

$$D_в = \frac{1}{N}\sum_{i=1}^{k} (x_i - \overline{x_в})^2 n_i = 396.16$$

Исправленная оценка дисперсии:

$$s^2 = \frac{N}{N-1} D_в = \frac{100}{99} * 396.16 = 400.16$$

Были найдены статистические оценки СКО:

$$\sigma_{\text{в}} = \sqrt{D_{\text{В}}} = \sqrt{396.16} = 19.9$$

$$s = \sqrt{s^2} = \sqrt{400.16} = 20$$

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

Были найдены статистические оценки коэффициентов асимметрии и эксцесса:

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

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

$$\overline{m_3} = \left(\overline{M_3^*} - 3\overline{M_2^*}\,\overline{M_1^*} + 2\left(\overline{M_1^*}\right)^3\right) h^3 = 2497.536$$

$$\overline{m_4} = \left(\overline{M_4^*} - 4\overline{M_3^*}\,\overline{M_1^*} + 6\overline{M_2^*}\left(\overline{M_1^*}\right)^2 + 2\left(\overline{M_1^*}\right)^4\right) h^4 = 538131.251$$

Статистическая оценка коэффициента асимметрии:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3} = 0.000038$$

Статистическая оценка коэффициента эксцесса:

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

Коэффициент асимметрии положительный, следовательно, это правосторонняя асимметрия, и $\bar{x}_{\text{в}} > M_o$, но полученный коэффициент незначительный и скос распределения небольшой. Коэффициент эксцесса же отрицателен, следовательно, эмпирическое распределение является более низким и пологим относительно нормального распределения.

Вычислим моду и медиану заданного распределения для интервального ряда. Мода заданного распределения:

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

$$M_o = 131.9 + 44\frac{1}{25} = 133.54$$

Медиана заданного распределения:

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

$$M_e = 116.9 + \frac{0.5 * 100 - 20}{32} 16 = 131.9$$

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

Вычислим доверительный интервал для оценки математического ожидания по формуле ниже:

$$\bar{x_в} - t_\gamma \frac{S}{\sqrt{n}} \leq \alpha \leq \bar{x_в} + t_\gamma \frac{S}{\sqrt{n}}, \text{где}$$

$\bar{x_в}$ – выборочное среднее

s – исправленное СКО

$t_\gamma = 2.627$ – из таблицы (при $\gamma = 0.99, N = 100$)

$$\bar{x_в} - t_\gamma \frac{s}{\sqrt{N}} = 465.26 - 2.627 * \frac{54.572}{10} = 450.92$$

$$\bar{x_в} + t_\gamma \frac{s}{\sqrt{N}} = 465.26 + 2.627 * \frac{54.572}{10} = 479.6$$

Доверительный интервал $(450.92; 479.6)$ покрывает истинное значение математического ожидания α с надежностью $\gamma = 0.99$.

Построим доверительный интервал для среднеквадратического отклонения.

Доверительный интервал для оценки СКО:

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

s – исправленное СКО

$q = 0.198$ –из таблицы (при $\gamma = 0.99, N = 100$)

$$s(1 - q) = 54.572 * 0.802 = 43.767$$

$$s(1 + q) = 54.572 * 1.198 = 66.377$$

Доверительный интервал $(43.767; 66.377)$ покрывает истинное значение среднеквадратического отклонения σ с надежностью $\gamma = 0.99$.

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:

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

$$\chi^2_{\text{крит}} = \chi^2(\alpha, k) - \text{из таблицы}$$

Гипотеза H_0 принимается при условии:

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

Вычислим теоретические частоты. Вычисления представлены в табл. 9.

Таблица 9

$\mathbf{x_i}$	$\mathbf{x_{i+1}}$	$\mathbf{n_i}$	$\mathbf{z_i}$	$\mathbf{z_{i+1}}$	$\Phi(\mathbf{z_i})$	$\Phi(\mathbf{z_{i+1}})$	$\mathbf{p_i}$	$\mathbf{n_i}'$
321.0	365.0	4	$-\infty$	-1.84	-0.5	-0.4671	0.0329	3.29
365.0	409.0	9	-1.84	-1.03	-0.4671	-0.3485	0.1186	11.86
409.0	453.0	27	-1.03	-0.22	-0.3485	-0.0871	0.2614	26.14
453.0	497.0	35	-0.22	0.58	-0.0871	0.219	0.3061	30.61
497.0	541.0	17	0.58	1.39	0.219	0.4177	0.1987	19.87
541.0	585.0	6	1.39	2.19	0.4177	0.4858	0.0681	6.81
585.0	623.0	2	2.19	$+\infty$	0.4858	0.5	0.0142	1.42

Вычислим наблюдаемое значение критерия $\chi^2_{\text{набл}}$ с помощью полученных частот по формуле ниже. Отдельные вычисления представлены в табл. 10.

Таблица 10

$\mathbf{n_i}$	$\mathbf{n_i}'$	$\mathbf{n_i - n_i'}$	$\mathbf{(n_i - n_i')^2}$	$\mathbf{(n_i - n_i')^2/n_i'}$
4	3.29	0.71	0.5041	0.1532
9	11.86	-2.86	8.1796	0.6897
27	26.14	0.86	0.7396	0.0283

35	30.61	4.39	19.2721	0.6296
17	19.87	-2.87	8.2369	0.4145
6	6.81	-0.81	0.6561	0.0963
2	1.42	0.58	0.3364	0.2369

$$\chi^2_{\text{набл}} = 2.2485$$

Определим табличное значение $\chi^2_{\text{крит}}$ при $\alpha = 0.05$ и $k = 7 - 3 = 4$:

$$\chi^2_{\text{крит}} = 9.5$$

Сравним полученные значения:

$$\chi^2_{\text{набл}} = 2.2485 \leq \chi^2_{\text{крит}} = 9.5$$

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

1.5. Выводы.

Была выбрана выборка, которая представляет собой данные 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. Выборка была преобразована в ранжированный, вариационный и интервальный ряды.

С помощью ранжированного ряда удалось определить минимальный и максимальный элемент выборки $x_{min} = 321$, $x_{max} = 623$, так как его элементы находятся в порядке возрастания. Далее при преобразовании ряда в вариационный ряд (объединение одинаковых элементов) удалось определить моду – значение в выборке, которое встречается наиболее часто, для данной выборки это $x_{42} = 465$ с абсолютной $n_{42} = 4$ и относительной частотой $\overline{n_{42}} = 0.04$. Далее при преобразовании интервального ряда из вариационного с помощью высчитанных значений количества интервалов $k = 7$ (нечетное) и последующего $h = 44$ можно было заметить, что наибольшая частота попаданий в интервал равная $n = 35$ находится в интервале $[453, 497)$.

Построенные графики также помогают увидеть наглядное представление ряда распределения. Видно, например, что в интервале $[453, 497)$ больше всего значений. Полигон строится как ломаная, которая соединяет точки, соответствующие срединным значениям интервалов и частотам этих интервалов, поэтому его форма не меняется для абсолютных и относительных частот, а меняется ось ординат, где как раз откладывают соответствующие абсолютные или относительные частоты. Гистограмма же — это фигура, состоящая из прямоугольников, площадь которых как раз и обозначает соответствующие частоты. Можно проверить, что для гистограммы абсолютных частот общая площадь прямоугольников равна объему выборки, а для гистограммы относительных частот она равна единице. Эмпирическая функция распределения же показывает отношение накопленных частот до середины интервалов к объему выборки $n = 100$, где опять же видно, как на интервале $[497, 541)$ с серединой равной 519 накопленная частота резко увеличивается.

Для интервального ряда для обеих переменных были вычислены условные эмпирические моменты через условные варианты. Была проведена корректность вычислений через контрольную сумму, вычисления оказались верны для обеих переменных. 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 для обеих переменных.

Были найдены коэффициенты асимметрии и эксцесса. Для обеих переменных коэффициент асимметрии получился положительным (правосторонняя асимметрия), то есть присутствует удлиненный правый хвост и $\overline{x_в} > M_o$, но полученное значение незначительно и скос распределения небольшой. Коэффициент эксцесса для обеих переменных получился уже отрицательным, следовательно, эмпирическое распределение является более низким и пологим относительно нормального распределения.

Был вычислен доверительный интервал для математического ожидания при неизвестном СКО с доверительной точностью $\gamma = 0.99$. Исходя из

полученных результатов можно сделать вывод, что доверительный интервал $(450.92;\ 479.6)$ покрывает истинное значение математического ожидания α с надежностью $\gamma = 0.99$. Вычислен доверительный интервал для среднеквадратического отклонения. Можно сделать вывод, что доверительный интервал $(43.767;\ 66.377)$ покрывает истинное значение среднеквадратического отклонения σ с надежностью $\gamma = 0.99$.

Выполнена проверка гипотезы о нормальности заданного распределения с помощью критерия χ^2 (Пирсона). Определено, что $\chi^2_{\text{набл}} \leq \chi^2_{\text{крит}}$, следовательно, нулевая гипотеза принимается, то есть можно предположить, что случайная величина распределена по нормальному закону распределения.

2. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

2.1. Основные теоретические положения.

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:

$$\phi(x/y) = \frac{f(x,y)}{f_2(y)}; \phi(y/x) = \frac{f(x,y)}{f_1(x)}$$

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.

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:

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

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 $H_0: r_{xy} = 0$. 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 $k = N - 2$ 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 — 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.

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

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:

$$\overline{y_x} = \overline{y}_{\text{в}} + \overline{r_{xy}}\frac{S_y}{S_x}(x - \overline{x}_{\text{в}});\ \overline{x_y} = \overline{x}_{\text{в}} + \overline{r_{xy}}\frac{S_x}{S_y}(y - \overline{y}_{\text{в}})$$

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

$$D_{\text{общ}} = D_{\text{межгр}} + D_{\text{внгр}}$$

Внутригрупповая дисперсия вычисляется, как взвешенная по объемам групп средняя арифметическая групповых дисперсий, межгрупповая – как дисперсия условных средних $\overline{x_{y_i}}$ относительно выборочной средней $\overline{x}_{\text{в}}$.

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

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

Запишем выборочное уравнение регрессии Y на X в параболическом виде:

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

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

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

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

- Статистическая обработка второй выборки

Выборка, сформированная из генеральной совокупности, представлена в таблице 11. Объём выборки: 100.

Таблица 11

№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***
1	481	135.2	***21***	418	131.4	***41***	513	159.3	***61***	450	122.3	***81***	475	143.6
2	445	124.7	***22***	378	103.8	***42***	489	149.8	***62***	468	128.9	***82***	518	144.4
3	550	147.9	***23***	521	154.9	***43***	474	132.5	***63***	441	122.8	***83***	566	175.7
4	465	140.9	***24***	394	117.7	***44***	379	94.6	***64***	460	140.7	***84***	464	131.3
5	566	168.5	***25***	504	145.3	***45***	472	135.6	***65***	480	117.7	***85***	394	112.1
6	497	147.3	***26***	440	126.7	***46***	544	169.6	***66***	429	112.9	***86***	480	146.1
7	478	136.6	***27***	465	114.8	***47***	507	142.4	***67***	457	126.4	***87***	321	86.1
8	521	139.6	***28***	418	109.3	***48***	409	116.7	***68***	464	143.2	***88***	502	132.5
9	352	84.9	***29***	418	118.6	***49***	498	164.0	***69***	431	125.0	***89***	460	122.4
10	422	117.9	***30***	465	127.7	***50***	468	142.0	***70***	424	119.0	***90***	458	104.7

11	506	153.5	***31***	447	117.5	***51***	593	187.4	***71***	502	137.2	***91***	362	111.7
12	443	122.9	***32***	433	131.5	***52***	523	152.6	***72***	465	140.7	***92***	503	148.5
13	434	140.4	***33***	460	136.8	***53***	478	126.6	***73***	492	137.5	***93***	446	144.0
14	422	108.6	***34***	382	98.8	***54***	438	122.2	***74***	446	128.4	***94***	421	115.1
15	569	157.4	***35***	532	160.6	***55***	423	115.9	***75***	482	136.4	***95***	407	110.5
16	439	119.2	***36***	482	148.2	***56***	408	110.0	***76***	510	140.6	***96***	448	137.7
17	437	129.4	***37***	472	122.6	***57***	386	105.8	***77***	434	122.3	***97***	490	139.9
18	461	138.6	***38***	532	158.7	***58***	428	130.3	***78***	623	195.7	***98***	482	141.2
19	351	89.0	*39*	473	137.9	*59*	560	169.8	*79*	468	141.2	*99*	463	129.2
20	390	91.4	***40***	525	148.3	***60***	483	130.3	***80***	471	119.7	***100***	459	145.4

Выборка для переменной E представлена в таблице 12.

Таблица 12

i	y_i	i	y_i	i	y_i	i	y_i	i	y_i
1	135.2	***21***	131.4	***41***	159.3	***61***	122.3	***81***	143.6
2	124.7	***22***	103.8	***42***	149.8	***62***	128.9	***82***	144.4
3	147.9	***23***	154.9	***43***	132.5	***63***	122.8	***83***	175.7
4	140.9	***24***	117.7	***44***	94.6	***64***	140.7	***84***	131.3
5	168.5	***25***	145.3	***45***	135.6	***65***	117.7	***85***	112.1
6	147.3	***26***	126.7	***46***	169.6	***66***	112.9	***86***	146.1
7	136.6	***27***	114.8	***47***	142.4	***67***	126.4	***87***	86.1
8	139.6	***28***	109.3	***48***	116.7	***68***	143.2	***88***	132.5
9	84.9	***29***	118.6	***49***	164.0	***69***	125.0	***89***	122.4
10	117.9	***30***	127.7	***50***	142.0	***70***	119.0	***90***	104.7
11	153.5	***31***	117.5	***51***	187.4	***71***	137.2	***91***	111.7
12	122.9	***32***	131.5	***52***	152.6	***72***	140.7	***92***	148.5
13	140.4	***33***	136.8	***53***	126.6	***73***	137.5	***93***	144.0
14	108.6	***34***	98.8	***54***	122.2	***74***	128.4	***94***	115.1
15	157.4	***35***	160.6	***55***	115.9	***75***	136.4	***95***	110.5
16	119.2	***36***	148.2	***56***	110.0	***76***	140.6	***96***	137.7
17	129.4	***37***	122.6	***57***	105.8	***77***	122.3	***97***	139.9
18	138.6	***38***	158.7	***58***	130.3	***78***	195.7	***98***	141.2
19	89.0	***39***	137.9	***59***	169.8	***79***	141.2	***99***	129.2
20	91.4	***40***	148.3	***60***	130.3	***80***	119.7	***100***	145.4

В таблице 13 представлено преобразование выборки в ранжированный ряд.

Таблица 13

i	y_i	i	y_i	i	y_i	i	y_i	i	y_i
1	84.9	***21***	117.5	***41***	127.7	***61***	137.9	***81***	147.3
2	86.1	***22***	117.7	***42***	128.4	***62***	138.6	***82***	147.9
3	89.0	***23***	117.7	***43***	128.9	***63***	139.6	***83***	148.2
4	91.4	***24***	117.9	***44***	129.2	***64***	139.9	***84***	148.3
5	94.6	***25***	118.6	***45***	129.4	***65***	140.4	***85***	148.5
6	98.8	***26***	119.0	***46***	130.3	***66***	140.6	***86***	149.8
7	103.8	***27***	119.2	***47***	130.3	***67***	140.7	***87***	152.6
8	104.7	***28***	119.7	***48***	131.3	***68***	140.7	***88***	153.5
9	105.8	***29***	122.2	***49***	131.4	***69***	140.9	***89***	154.9
10	108.6	***30***	122.3	***50***	131.5	***70***	141.2	***90***	157.4
11	109.3	***31***	122.3	***51***	132.5	***71***	141.2	***91***	158.7
12	110.0	***32***	122.4	***52***	132.5	***72***	142.0	***92***	159.3
13	110.5	***33***	122.6	***53***	135.2	***73***	142.4	***93***	160.6
14	111.7	***34***	122.8	***54***	135.6	***74***	143.2	***94***	164.0
15	112.1	***35***	122.9	***55***	136.4	***75***	143.6	***95***	168.5
16	112.9	***36***	124.7	***56***	136.6	***76***	144.0	***96***	169.6
17	114.8	***37***	125.0	***57***	136.8	***77***	144.4	***97***	169.8
18	115.1	***38***	126.4	***58***	137.2	***78***	145.3	***98***	175.7
19	115.9	***39***	126.6	***59***	137.5	***79***	145.4	***99***	187.4
20	116.7	***40***	126.7	***60***	137.7	***80***	146.1	***100***	195.7

Видно, что $y_{min} = 84.9$, а $y_{max} = 195.7$

В таблице 14 представлено преобразование полученной выборки в вариационный ряд с абсолютными $\boldsymbol{n_i}$ и относительными $\overline{\boldsymbol{n_i}}$ частотами.

Таблица 14

i	y_i	n_i	$\overline{n_i}$	i	y_i	n_i	$\overline{n_i}$	i	y_i	n_i	$\overline{n_i}$	i	y_i	n_i	$\overline{n_i}$
1	84.9	1	0.01	***26***	119.2	1	0.01	***51***	136.4	1	0.01	***76***	147.9	1	0.01
2	86.1	1	0.01	***27***	119.7	1	0.01	***52***	136.6	1	0.01	***77***	148.2	1	0.01
3	89.0	1	0.01	***28***	122.2	1	0.01	***53***	136.8	1	0.01	***78***	148.3	1	0.01
4	91.4	1	0.01	***29***	122.3	2	0.02	***54***	137.2	1	0.01	***79***	148.5	1	0.01
5	94.6	1	0.01	***30***	122.4	1	0.01	***55***	137.5	1	0.01	***80***	149.8	1	0.01
6	98.8	1	0.01	***31***	122.6	1	0.01	***56***	137.7	1	0.01	***81***	152.6	1	0.01

7	103.8	1	0.01	***32***	122.8	1	0.01	***57***	137.9	1	0.01	***82***	153.5	1	0.01
8	104.7	1	0.01	***33***	122.9	1	0.01	***58***	138.6	1	0.01	***83***	154.9	1	0.01
9	105.8	1	0.01	***34***	124.7	1	0.01	***59***	139.6	1	0.01	***84***	157.4	1	0.01
10	108.6	1	0.01	***35***	125.0	1	0.01	***60***	139.9	1	0.01	***85***	158.7	1	0.01
11	109.3	1	0.01	***36***	126.4	1	0.01	***61***	140.4	1	0.01	***86***	159.3	1	0.01
12	110.0	1	0.01	***37***	126.6	1	0.01	***62***	140.6	1	0.01	***87***	160.6	1	0.01
13	110.5	1	0.01	***38***	126.7	1	0.01	***63***	140.7	2	0.02	***88***	164.0	1	0.01
14	111.7	1	0.01	***39***	127.7	1	0.01	***64***	140.9	1	0.01	***89***	168.5	1	0.01
15	112.1	1	0.01	***40***	128.4	1	0.01	***65***	141.2	2	0.02	***90***	169.6	1	0.01
16	112.9	1	0.01	***41***	128.9	1	0.01	***66***	142.0	1	0.01	***91***	169.8	1	0.01
17	114.8	1	0.01	***42***	129.2	1	0.01	***67***	142.4	1	0.01	***92***	175.7	1	0.01
18	115.1	1	0.01	***43***	129.4	1	0.01	***68***	143.2	1	0.01	***93***	187.4	1	0.01
19	115.9	1	0.01	***44***	130.3	2	0.02	***69***	143.6	1	0.01	***94***	195.7	1	0.01
20	116.7	1	0.01	***45***	131.3	1	0.01	***70***	144.0	1	0.01				
21	117.5	1	0.01	***46***	131.4	1	0.01	***71***	144.4	1	0.01				
22	117.7	2	0.02	***47***	131.5	1	0.01	***72***	145.3	1	0.01				
23	117.9	1	0.01	***48***	132.5	2	0.02	***73***	145.4	1	0.01				
24	118.6	1	0.01	***49***	135.2	1	0.01	***74***	146.1	1	0.01				
25	119.0	1	0.01	***50***	135.6	1	0.01	***75***	147.3	1	0.01				

Количество интервалов разбиения вычислено с помощью формулы Стерджесса:

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

Ширина интервала:

$$h = \frac{y_{max} - y_{min}}{k} = \frac{195.7 - 84.9}{7} = 16$$

В таблице 15 представлен полученный интервальный ряд.

Таблица 15

Границы интервалов	**Середины интервалов**	**Абсолютная частота**	**Относительная частота**
[84.9, 100.9)	92.9	6	0.06
[100.9, 116.9)	108.9	14	0.14
[116.9, 132.9)	124.9	32	0.32
[132.9, 148.9)	140.9	33	0.33
[148.9, 164.9)	156.9	9	0.09

[164.9, 180.9)	172.9	4	0.04
[180.9, 195.7)	188.3	2	0.02

Далее для интервального ряда абсолютных частот были построены полигон и гистограмма. Полигон представлен на рис. 2.2.1.

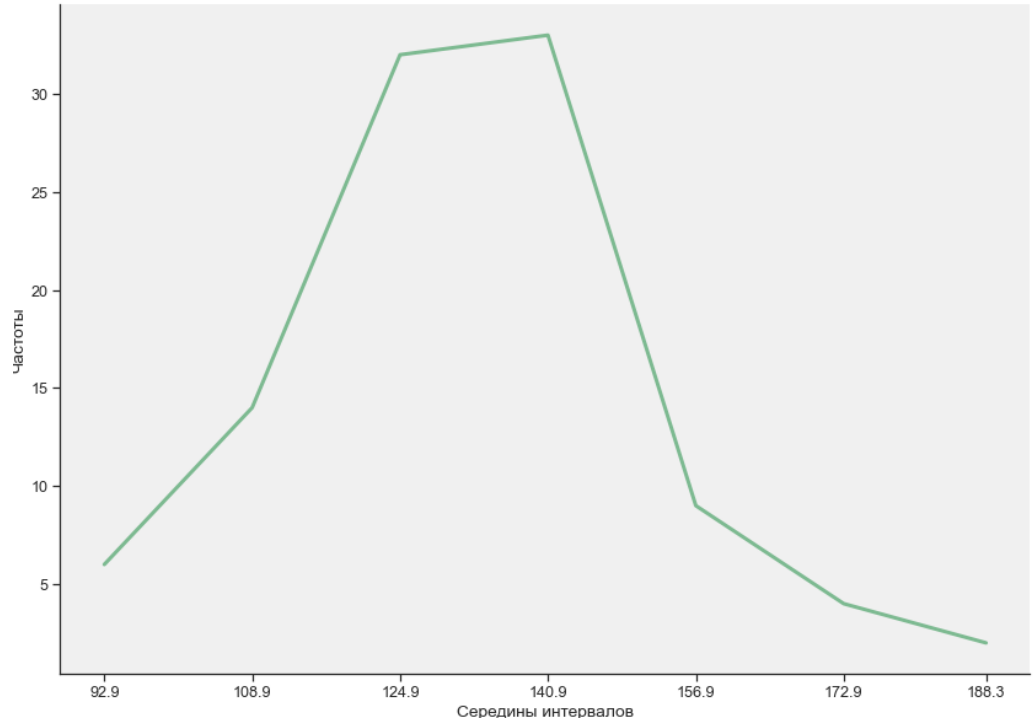

Рисунок 2.2.1 – Полигон для абсолютных частот

Полигон представляет собой ломаную, соединяющую точки, соответствующие срединным значениям интервалов и абсолютным частотам этих интервалов. Гистограмма, представлена на рис. 2.2.2.

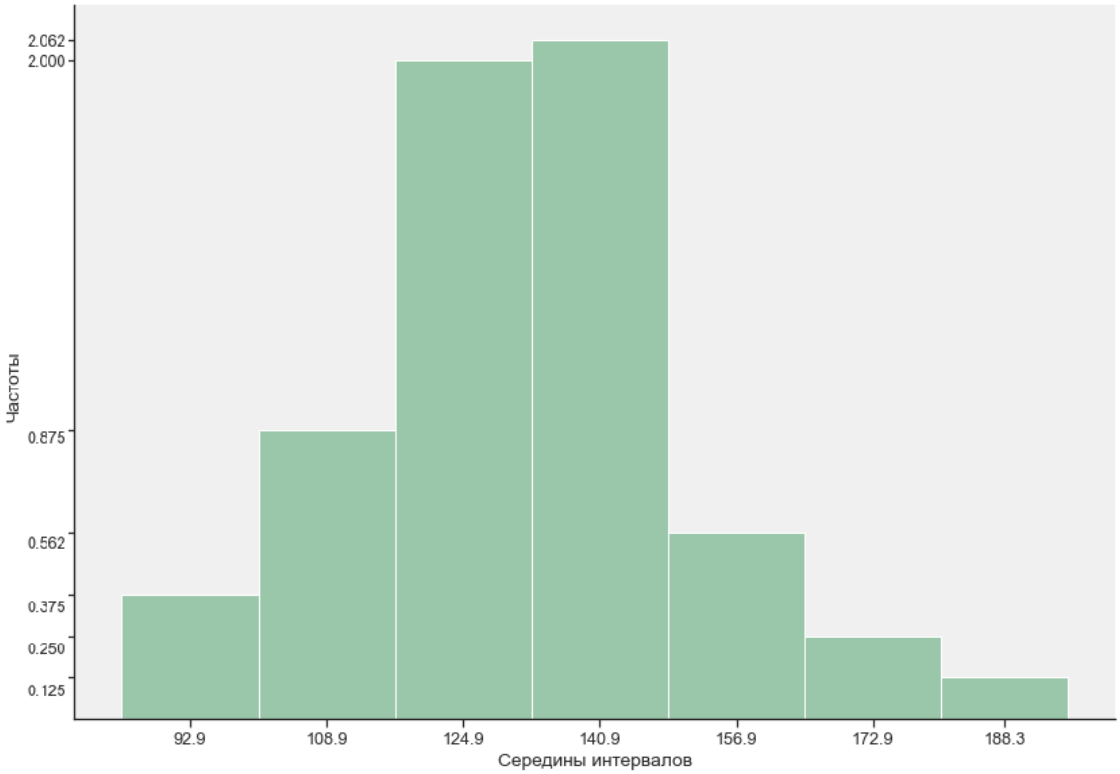

Рисунок 2.2.2 – Гистограмма для абсолютных частот

Гистограмма представляет собой фигуру, состоящую из прямоугольников, основания которых это длина интервалов h, а высота равна отношению частоты к длине интервала, то есть площадь прямоугольника обозначает частоту интервала.

Графики для интервального ряда относительных частот представлены ниже. Эмпирическая функция распределения, построенная применительно к интервальному ряду для относительных частот представлен на рис. 2.2.3.

Функция распределения:

$$F(92.9) = 0$$

$$F(108.9) = 0.06$$

$$F(124.9) = 0.20$$

$$F(140.9) = 0.52$$

$$F(156.9) = 0.85$$

$$F(172.9) = 0.94$$

$$F(188.3) = 0.98$$

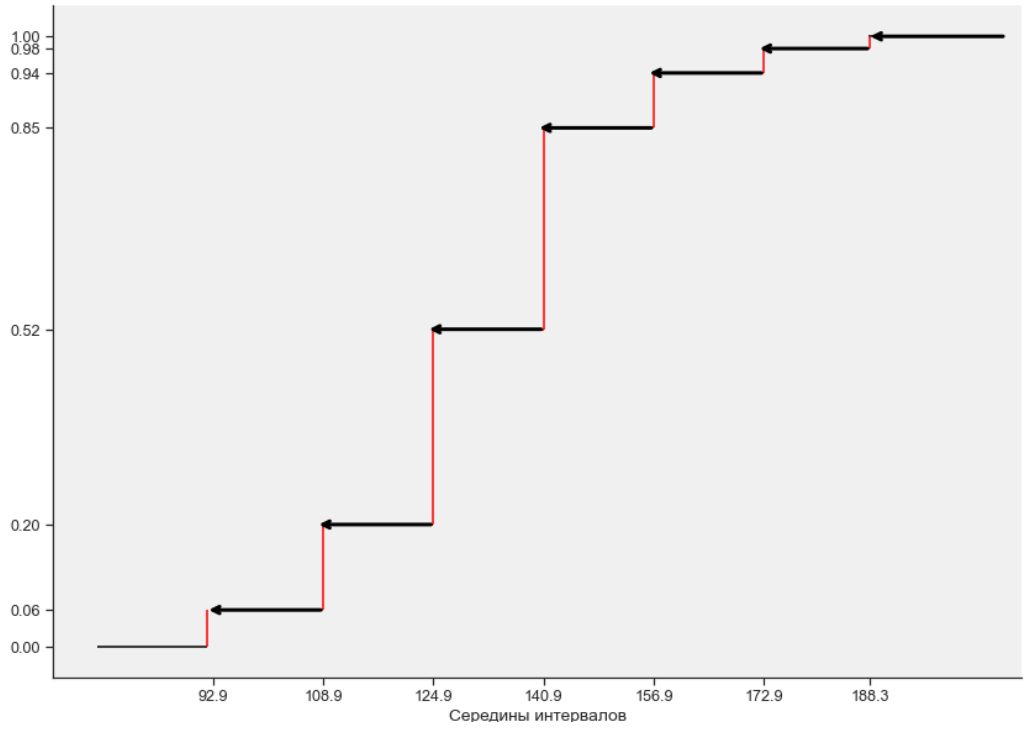

Рисунок 2.2.3 – График эмпирической функции распределения

Полигон для относительных частот представлен на рис. 2.2.4.

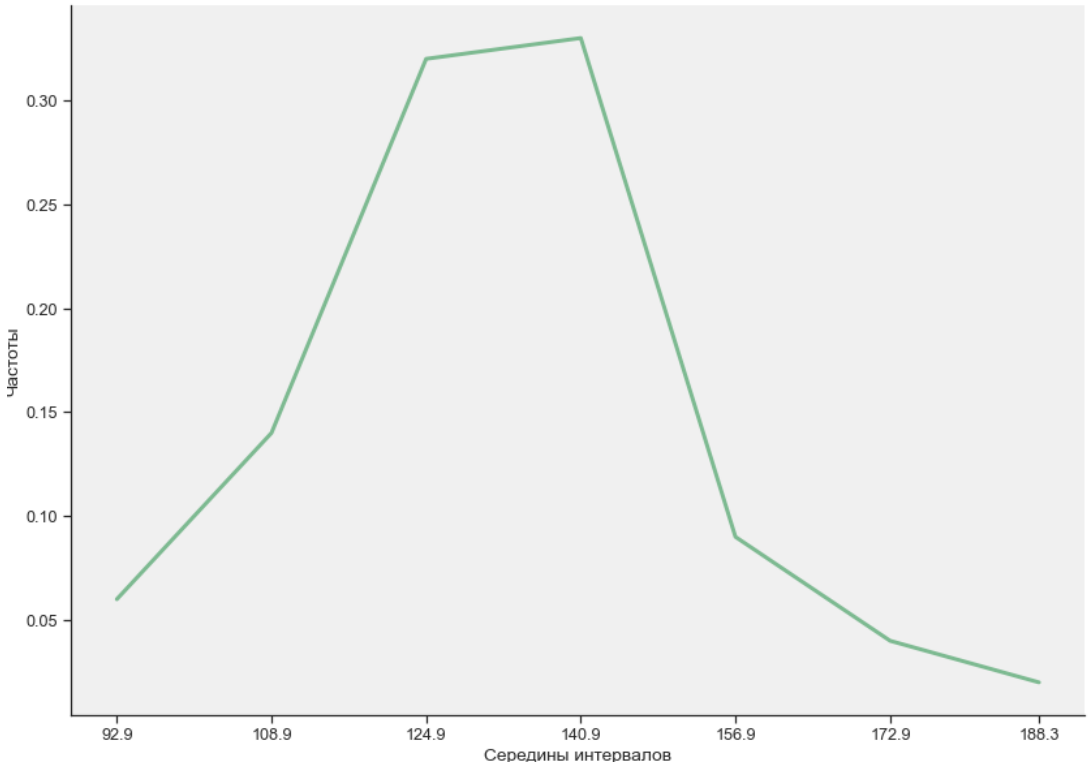

Рисунок 2.2.4 – Полигон для относительных частот

Гистограмма для относительных частот, представлена на рис. 2.2.5.

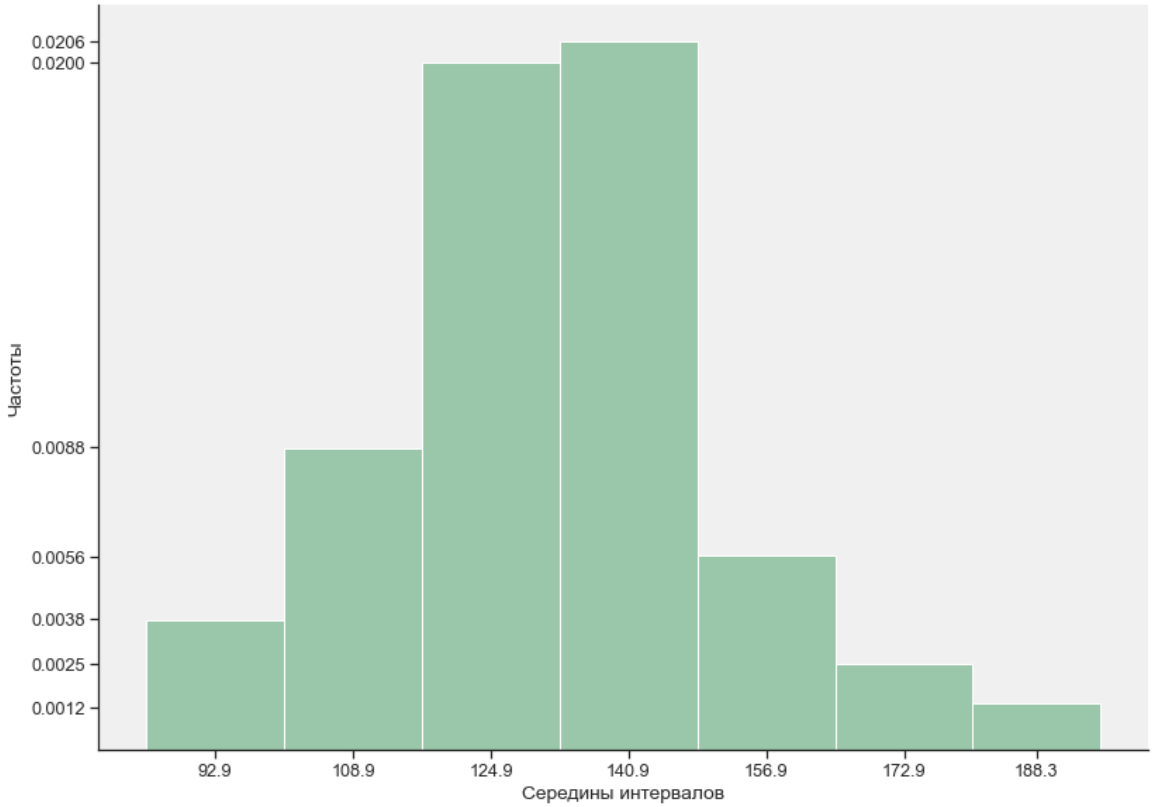

Рисунок 2.2.5 – Гистограмма для относительных частот

Интервальный ряд для переменной E и с посчитанными накопленными частотами представлен в таблице 16.

Таблица 16

Границы интервалов	**Середины интервалов**	**Абсолютная частота**	**Относительная частота**	**Накопленная частота**
[84.9, 100.9)	92.9	6	0.06	0.06
[100.9, 116.9)	108.9	14	0.14	0.2
[116.9, 132.9)	124.9	32	0.32	0.52
[132.9, 148.9)	140.9	33	0.33	0.85
[148.9, 164.9)	156.9	9	0.09	0.94
[164.9, 180.9)	172.9	4	0.04	0.98
[180.9, 195.7)	188.3	2	0.02	1

Результаты вычислений условных моментов представлены в табл. 17.

Таблица 7

$\boldsymbol{v}$	$\boldsymbol{n}$	$\boldsymbol{u}$	$\boldsymbol{n} * \boldsymbol{u}$	$\boldsymbol{n} * \boldsymbol{u^2}$	$\boldsymbol{n} * \boldsymbol{u^3}$	$\boldsymbol{n} * \boldsymbol{u^4}$	$\boldsymbol{n} * (\boldsymbol{u}+\boldsymbol{1})^{\boldsymbol{4}}$

92.9	0.06	-3	-0.18	0.54	-1.62	4.86	0.96
108.9	0.14	-2	-0.28	0.56	-1.12	2.24	0.14
124.9	0.32	-1	-0.32	0.32	-0.32	0.32	0.0
140.9	0.33	0	0.0	0.0	0.0	0.0	0.33
156.9	0.09	1	0.09	0.09	0.09	0.09	1.44
172.9	0.04	2	0.08	0.16	0.32	0.64	3.24
188.3	0.02	3	0.06	0.18	0.54	1.62	5.12
∑	1	–	-0.55	1.85	-2.11	9.77	11.23

Проверим вычисления с помощью последнего столбца:

$$\sum n_j * u_j^4 + 4 * \sum n_j * u_j^3 + 6 * \sum n_j * u_j^2 + 4 * \sum n_j * u_j + 1 =$$
$$= 9.77 + 4 * -2.11 + 6 * 1.85 + 4 * -0.55 + 1 = 11.23$$

Число совпадает с суммой элементов последнего столбца, следовательно вычисления правильные.

Был посчитан первый начальный эмпирический момент с помощью условных вариант, который обозначает выборочное среднее:

$$\bar{x_в} = \overline{M_1} = \overline{M_1^*}h + C = 132.1$$

Также был посчитан второй центральный эмпирический момент с помощью условных вариант, который обозначает выборочную дисперсию:

$$D_в = \overline{m_2} = \left(\overline{M_2^*} - \left(\overline{M_1^*}\right)^2\right)h^2 = 395.16$$

Далее были найдены выборочное среднее и дисперсия с помощью стандартных формул.

$$\bar{x_в} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = 132.09$$

$$D_в = \frac{1}{N}\sum_{i=1}^{k} (x_i - \bar{x_в})^2 n_i = 394.8$$

Исправленная оценка дисперсии:

$$s^2 = \frac{N}{N-1} D_в = \frac{100}{99} * 394.8 = 398.79$$

Были найдены статистические оценки СКО:

$$\sigma_{\text{в}} = \sqrt{D_{\text{В}}} = \sqrt{394.8} = 19.87$$

$$s = \sqrt{s^2} = \sqrt{398.79} = 19.97$$

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

Были найдены статистические оценки коэффициентов асимметрии и эксцесса:

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

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

$$\overline{m_3} = \left(\overline{M_3^*} - 3\overline{M_2^*}\,\overline{M_1^*} + 2\left(\overline{M_1^*}\right)^3\right) h^3 = 2497.536$$

$$\overline{m_4} = \left(\overline{M_4^*} - 4\overline{M_3^*}\,\overline{M_1^*} + 6\overline{M_2^*}\left(\overline{M_1^*}\right)^2 + 2\left(\overline{M_1^*}\right)^4\right) h^4 = 538131.251$$

Статистическая оценка коэффициента асимметрии:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3} = 0.000039$$

Статистическая оценка коэффициента эксцесса:

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

Коэффициент асимметрии положительный – это правосторонняя асимметрия. Коэффициент эксцесса отрицательный – пик распределения около математического ожидания гладкий.

- Двумерный интервальный вариационный ряд

В таблице 18 представлен построенный двумерный интервальный вариационный ряд (корреляционная таблица).

Таблица 18

Y	**X**							
	343	387	431	475	519	563	604	n_y

92.9	3	3	0	0	0	0	0	6
108.9	1	5	6	2	0	0	0	14
124.9	0	1	18	12	1	0	0	32
140.9	0	0	3	20	9	1	0	33
156.9	0	0	0	1	7	1	0	9
172.9	0	0	0	0	0	4	0	4
188.3	0	0	0	0	0	0	2	2
$\mathbf{n_x}$	4	9	27	35	17	6	2	100

Как видно из таблицы суммы частот по столбцам совпадают с абсолютными частотами интервального вариационного ряда по признаку nu, то же самое можно сказать и для строк (переменная E), таблица составлена корректно.

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

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

Чтобы удобно посчитать двойную сумму, можно воспользоваться преобразованием ниже, данные вычисления представлены в таблице 19.

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

Таблица 19

Y	***X***								
	343	*387*	*431*	*475*	*519*	*563*	*604*	X_i	y_iX_i
92.9	1029 **3** 278.7	1161 **3** 278.7						2190	203451
108.9	343 **1** 108.9	1935 **5** 544.5	2586 **6** 653.4	950 **2** 217.8				5814	633144.6
124.9		387 **1** 124.9	7758 **18** 2248.2	5700 **12** 1498.8	519 **1** 124.9			14364	1794063.6

140.9			1293 **3** 422.7	9500 **20** 2818	4671 **9** 1268.1	563 **1** 140.9		16027	2258204.3
156.9				475 **1** 156.9	3633 **7** 1098.3	563 **1** 156.9		4671	732879.9
172.9						2252 **4** 691.6		2252	389370.8
188.3							1208 **2** 376.6	1208	227466.4
Y_j	387.6	948.1	3324.3	4691.5	2491.3	989.4	376.6	**6238580.6**	
x_jY_j	132946.8	366914.7	1432773.3	2228462.5	1292984.7	557032.2	227466.4		

Вычислен выборочный коэффициент корреляции:

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}y_i x_j - N\bar{x}_в\bar{y}_в}{NS_xS_y} = \frac{6238580.6 - 100 * 465.26 * 132.09}{100 * 54.57 * 19.97} = 0.853$$

Выборочный коэффициент корреляции не равен нулю и положителен, значит X и Y коррелированы и зависимы, а также это положительная корреляционная зависимость.

Также по аналогии было посчитано значение выборочного коэффициента корреляции с помощью условных вариант.

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}u_i v_j - N\overline{u_в}\overline{v_в}}{NS_uS_v} = \frac{145 - 100 * -0.2115 * -0.529}{100 * 1.214 * 1.224} = 0.8525$$

Коэффициенты корреляции, рассчитанные с помощью основной формулы и условных вариант совпали.

Оценим значение r_{xy} в случае нормального распределения:

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

$$0.853 - 3\frac{1-0.853^2}{\sqrt{100}} \le r_{xy} \le 0.853 + 3\frac{1+0.853^2}{\sqrt{100}}$$

$$0.7713 \le r_{xy} \le 1$$

- Доверительный интервал для коэффициента корреляции

  Построим доверительный интервал для коэффициента корреляции. Перейдём к случайной величине z:

$$\bar{z} = 0.5 \ln \frac{1 + \bar{r}_{xy}}{1 - \bar{r}_{xy}} = 0.5 \ln \frac{1 + 0.853}{1 - 0.853} = 1.267$$

  Среднеквадратическое отклонение:

$$\bar{\sigma}_z = \frac{1}{\sqrt{N - 3}} = \frac{1}{\sqrt{100 - 3}} = 0.1015$$

  Доверительный интервал:

$$(\bar{z} - \lambda(\gamma)\bar{\sigma}_z; \bar{z} + \lambda(\gamma)\bar{\sigma}_z), \Phi\big(\lambda(\gamma)\big) = \frac{\gamma}{2}$$

  При уровне значимости $\gamma = 0.99$:

$$\Phi\big(\lambda(\gamma)\big) = 0.495 \Rightarrow \lambda(\gamma) = 2.58$$

$$(1.267 - 2.58 * 0.1015; 1.267 + 2.58 * 0.1015)$$

$$(1.0051; 1.5289)$$

Для построения доверительного интервала для коэффициента корреляции воспользуемся обратным преобразованием Фишера:

$$r_{xy} \in \left(\frac{e^{2z_l} - 1}{e^{2z_l} + 1}; \frac{e^{2z_r} - 1}{e^{2z_r} + 1}\right)$$

$$\frac{e^{2z_l} - 1}{e^{2z_l} + 1} = 0.7637; \frac{e^{2z_r} - 1}{e^{2z_r} + 1} = 0.9102$$

Доверительный интервал $(0.7637; 0.9102)$ покрывает истинное значение коэффициента корреляции с надежностью $\gamma = 0.99$.

- Гипотеза о равенстве коэффициента корреляции нулю

  Проверим гипотезу $H_0: r_{xy} = 0$; $H_1: r_{xy} \neq 0$.

  В качестве критерия проверки гипотезы примем случайную величину:

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

  Найдём $T_{\text{набл}}$ по формуле:

$$T_{\text{набл}} = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}} = \frac{0.853 * \sqrt{98}}{\sqrt{1-0.853^2}} = 16.18$$

Для уровня значимости $\alpha = 0.05$ и $k = 102$ было определено $t_{\text{крит}} = 1.986$.

Определено, что $|T_{\text{набл}}| > t_{\text{крит}}$, то есть основная гипотеза H_0 должна быть отвергнута, это означает, что выборочный коэффициент корреляции $\bar{r}_{xy}$ значим.

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

Для заданной двумерной выборки были построены уравнения средней квадратичной регрессии x на y и y на x. Далее полученные прямые были отображены на множестве выборки.

Выборочные прямые средней квадратичной регрессии x на y и y на x:

$$\overline{x_y} = \bar{x}_{\text{в}} + \overline{r_{xy}}\,\frac{S_x}{S_y}(y - \bar{y}_{\text{в}})$$

$$\overline{y_x} = \bar{y}_{\text{в}} + \overline{r_{xy}}\,\frac{S_y}{S_x}(x - \bar{x}_{\text{в}})$$

$$x(y) = 2.331 * y + 157.371$$

$$y(x) = 0.3122 * x - 13.1442$$

Полученные прямые, отображенные на множестве выборки представлены на рис. 2.3.1.

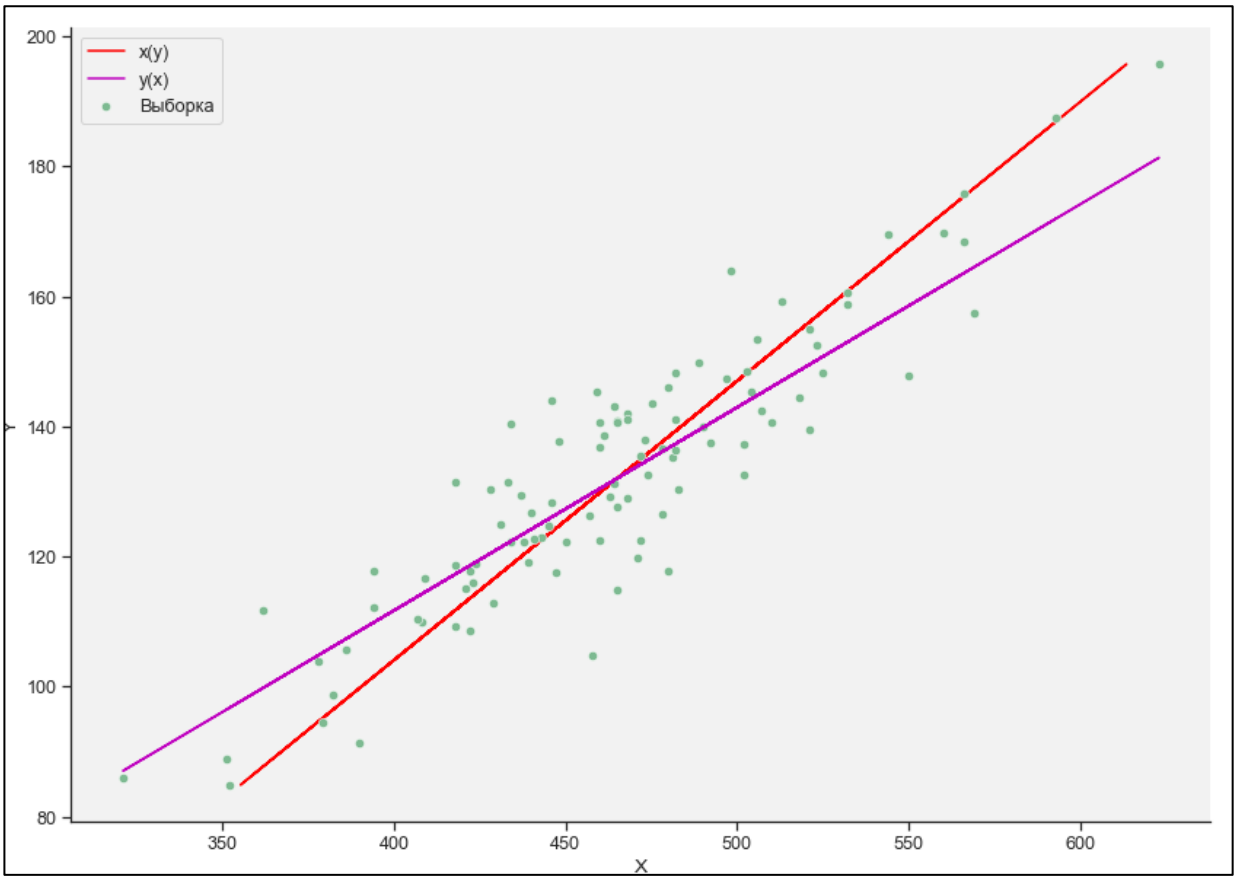

Рисунок 2.3.1 - Выборочные прямые средней квадратичной регрессии x на y и y на x

Были найдены статистические оценки остаточной дисперсии для полученных выборочных прямых средней квадратичной регрессии x на y и y на x:

$$D_{\text{ост } x} = S_x^2\left(1 - \overline{r_{xy}}^2\right) = 811.149$$

$$D_{\text{ост } y} = S_y^2\left(1 - \overline{r_{xy}}^2\right) = \ 108.63$$

- Нахождение выборочного корреляционного отношения

Была составлена корреляционная таблица для нахождения выборочного корреляционного отношения, которая представлена в таблице 20. Были посчитаны условные выборочные средние и дисперсии.

Таблица 20 - Корреляционная таблица

Y	**X**									
	343	**387**	**431**	**475**	**519**	**563**	**604**	$\mathbf{n_y}$	$\mathbf{\bar{x}_{гр}}$	$\mathbf{D_{x_{гр}}}$
92.9	3	3	0	0	0	0	0	6	365	484

108.9	1	5	6	2	0	0	0	14	415.29	1270.64
124.9	0	1	18	12	1	0	0	32	448.88	704.5
140.9	0	0	3	20	9	1	0	33	485.67	821.65
156.9	0	0	0	1	7	1	0	9	519	430.22
172.9	0	0	0	0	0	4	0	4	563	0
188.3	0	0	0	0	0	0	2	2	604	0
$\mathbf{n}_x$	4	9	27	35	17	6	2	100		
$\bar{\mathbf{y}}_{\text{гр}}$	96.9	105.34	123.12	134.04	146.55	164.9	188.3			
$\mathbf{D}_{\mathbf{y}_{\text{гр}}}$	48	102.07	82.72	107.35	87.72	149.33	0			

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

$$\overline{\eta_{yx}} = \frac{\overline{\sigma_{\bar{y}_x}}}{\overline{\sigma_y}} = \sqrt{\frac{D_{\text{межгр}}}{D_{\text{общ}}}}$$

Аналогично для выборочного корреляционного отношения X к Y.

Для этого были рассчитаны внутригрупповая, межгрупповая и общая дисперсии. Выборочное корреляционное отношение Y к X:

$$D_{\text{внгр } yx} = \frac{1}{n}\sum_{i=1}^{k1} \mathrm{D}_{\mathrm{y}_{\text{гр}\,i}} n_{x_i} = 94.8854$$

$$D_{\text{межгр } yx} = \frac{1}{n}\sum_{i=1}^{k1} \left(\bar{\mathrm{y}}_{\text{гр i}} - \bar{y}_{\text{в}}\right)^2 n_{x_i} = 300.316$$

$$D_{\text{общ } yx} = D_{\text{внгр } yx} + D_{\text{межгр } yx} = 395.2014$$

$$\overline{\eta_{yx}} = \sqrt{\frac{D_{\text{межгр } yx}}{D_{\text{общ } yx}}} = 0.8717$$

Выборочное корреляционное отношение X к Y:

$$D_{\text{внгр } xy} = \frac{1}{n}\sum_{i=1}^{k2} \mathrm{D}_{\mathrm{x_{гр\,i}}} n_{y_i} = 742.2339$$

$$D_{\text{межгр } xy} = \frac{1}{n}\sum_{i=1}^{k2} \left(\bar{\mathrm{x}}_{\text{гр i}} - \bar{x_{\text{в}}}\right)^2 n_{y_i} = 2203.048$$

$$D_{\text{общ } xy} = D_{\text{внгр } xy} + D_{\text{межгр } xy} = 2945.2819$$

$$\overline{\eta_{xy}} = \sqrt{\frac{D_{\text{межгр } xy}}{D_{\text{общ } xy}}} = 0.8649$$

Убедимся, что неравенства $\overline{\eta_{xy}} \geq \left|\overline{r_{xy}}\right|$ и $\overline{\eta_{yx}} \geq \left|\overline{r_{xy}}\right|$ выполняются:

$$\overline{r_{xy}} = 0.853$$

$$\overline{\eta_{xy}} = 0.8649$$

$$\overline{\eta_{yx}} = 0.8717$$

Неравенство $\overline{\eta_{xy}} \geq \left|\overline{r_{xy}}\right|$ выполняется, так же, как и неравенство $\overline{\eta_{yx}} \geq \left|\overline{r_{xy}}\right|$.

- Построение корреляционных кривых

1. Параболический вид

Для заданной выборки была построена корреляционная кривая параболического вида $y = \beta_2 x^2 + \beta_1 x^2 + \beta_0$.

Запишем выборочное уравнение регрессии Y на X в параболическом виде:

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

Значения коэффициентов определим с помощью МНК. Была решена следующая система уравнений:

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

Чтобы удобно рассчитать приведенные суммы была построена таблица 21.

Таблица 21 – Таблица сумм МНК

$\mathbf{x}$	$\mathbf{n_x}$	$\mathbf{\bar{y}_x}$	$\mathbf{n_x x}$	$\mathbf{n_x x^2}$	$\mathbf{n_x x^3}$	$\mathbf{n_x x^4}$	$\mathbf{n_x \bar{y}_x}$	$\mathbf{n_x \bar{y}_x x}$	$\mathbf{n_x \bar{y}_x x^2}$
343	4	96.9	1372	470596	161414428	55365148804	387.6	132946.8	45600752.4
387	9	105.34	3483	1347921	521645427	201876780249	948	366899.22	141989998.14
431	27	123.12	11637	5015547	2161700757	931693026267	3324.24	1432747.44	617514146.64
475	35	134.04	16625	7896875	3751015625	1781732421875	4691.4	2228415	1058497125
519	17	146.55	8823	4579137	2376572103	1233440921457	2491.35	1293010.65	671072527.35
563	6	164.9	3378	1901814	1070721282	602816081766	989.4	557032.2	313609128.6
604	2	188.3	1208	729632	440697728	266181427712	376.6	227466.4	137389705.6
∑	100		46526	21941522	10483767350	5073105808130	13208.65	6238517.7	2985673383.73

Система была решена с помощью написанной программы. В результате были получены следующие значения коэффициентов:

$$a = 0.0003$$

$$b = 0.0021$$

$$c = 57.4223$$

Тогда, выборочное уравнение регрессии Y на X:

$$y(x) = 0.0003 * x^2 + 0.0021 * x + 57.4223$$

Корреляционная кривая параболического вида на множестве выборки представлена на рис. 2.

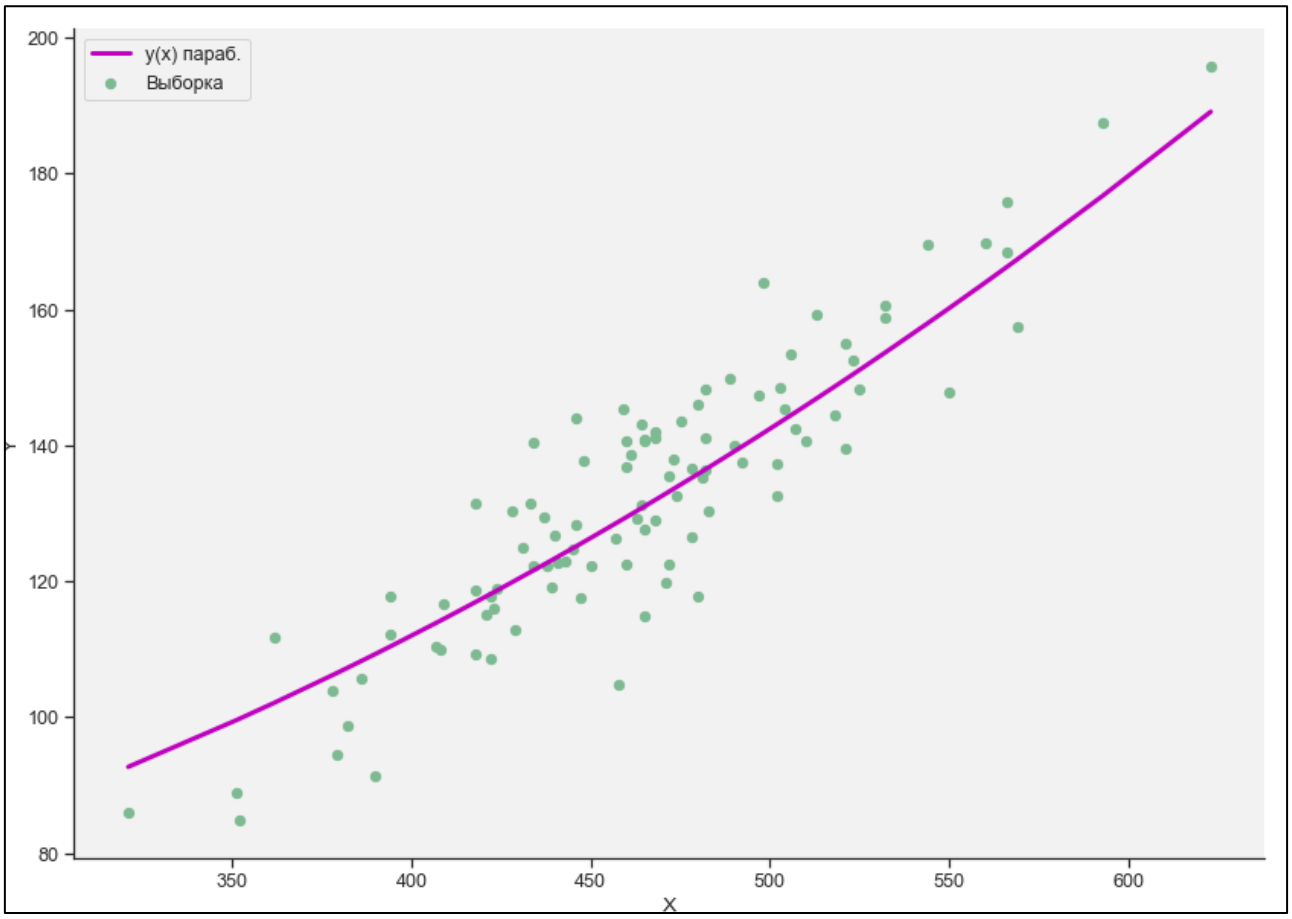

Рисунок 2.3.2 – Корреляционная кривая параболического вида

2. Логарифмическая функция

Для заданной выборки построим корреляционную кривую логарифмической функции $y = \beta_1 \ln x + \beta_0$. Выборочное уравнение регрессии Y на X:

$$\overline{y_x} = a + b \ln x$$

Применяя МНК, можно получить формулы для расчета значений коэффициентов a и b:

$$\begin{cases} b = \dfrac{n \sum_{i=1}^{K} (\overline{y_{x_i}} * \ln x_i) - \sum_{i=1}^{K} \ln \overline{y_{x_i}} * \sum_{i=1}^{K} \ln x_i}{n \sum_{i=1}^{K} (\ln x_i)^2 - (\sum_{i=1}^{K} \ln x_i)^2} \\ a = \dfrac{\sum_{i=1}^{K} \overline{y_{x_i}} - (\sum_{i=1}^{K} \ln x_i) b}{n} \end{cases}$$

С помощью написанной программы на языке Python были найдены данные коэффициенты:

$$a = -845.15$$

$$b = 159.4$$

Тогда, выборочное уравнение регрессии Y на X:

$$y(x) = -845.15 + 159.4 * \ln x$$

Корреляционная кривая логарифмической функции представлена на рисунке 2.3.3.

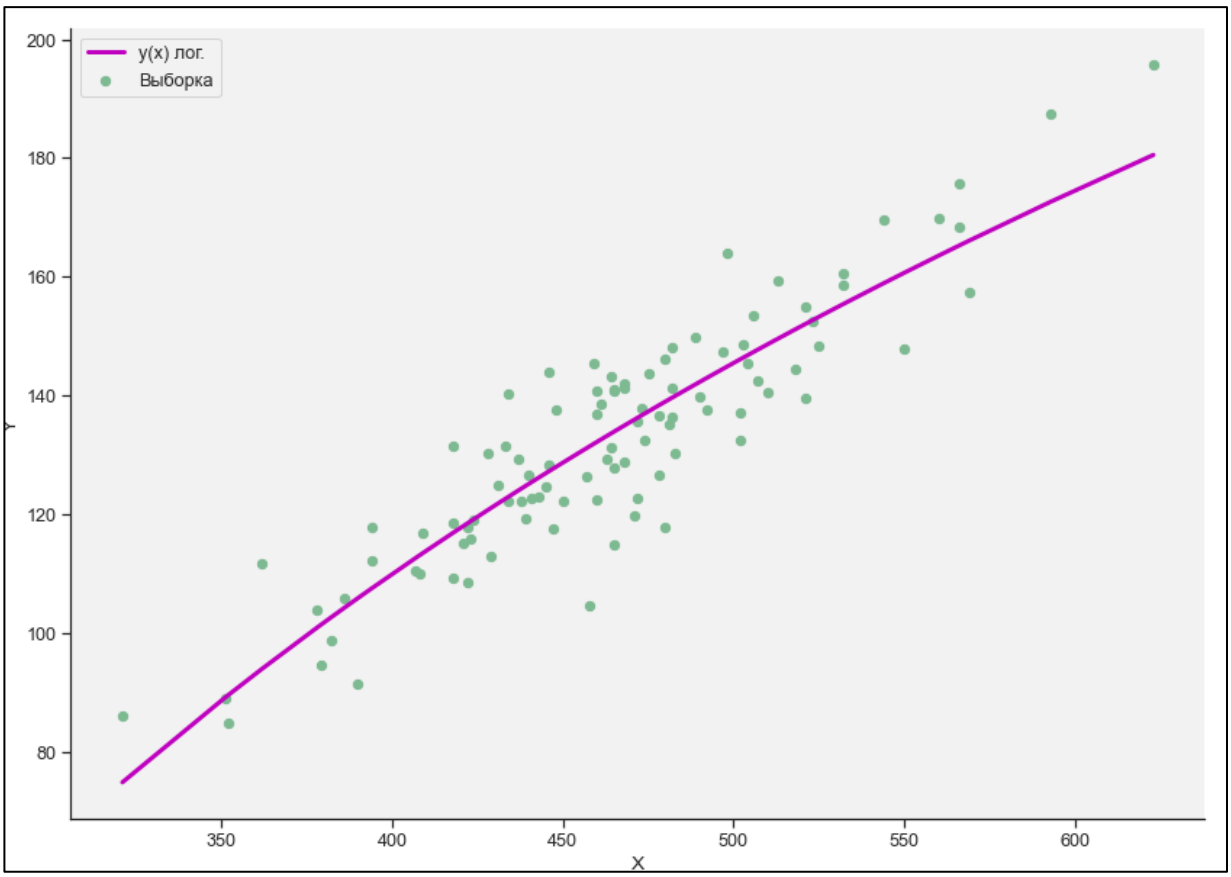

Рисунок 2.3.3 – Корреляционная кривая логарифмической функции

2.4. Выводы.

Был построен двумерный интервальный вариационный ряд (корреляционная таблица). На основании результатов корреляционной таблицы был вычислен выборочный коэффициент корреляции $\overline{r_{xy}} = 0.853$. Выборочный коэффициент корреляции не равен нулю и положителен, значит X и Y коррелированы и зависимы, а также это положительная корреляционная зависимость. Также было посчитано значение выборочного коэффициента корреляции с помощью условных вариант. Коэффициенты корреляции, рассчитанные с помощью основной формулы и условных вариант совпали. С

помощью выборочного коэффициента корреляции было оценено значение r_{xy} в случае нормального распределения.

Построен доверительный интервал для коэффициента корреляции при уровне значимости $\gamma = 0.99$. Определено, что доверительный интервал $(0.7637; 0.9102)$ покрывает истинное значение коэффициента корреляции с надежностью $\gamma = 0.99$.

Осуществлена проверка статистической гипотезы о равенстве коэффициента корреляции нулю при заданном уровне значимости $\alpha = 0.05$. Найдены значения $T_{\text{набл}} = 16.18$ и $t_{\text{крит}} = 1.986$. Определено, что $|T_{\text{набл}}| > t_{\text{крит}}$, то есть основная гипотеза H_0 должна быть отвергнута, это означает, что выборочный коэффициент корреляции $\bar{r}_{xy}$ значим.

Для заданной двумерной выборки были получены выборочные прямые средней квадратичной регрессии x на y и y на x. Данные прямые были построены на множестве выборки.

Были найдены условные выборочные средние и дисперсии и посчитаны внутригрупповая, межгрупповая и общая дисперсии для расчёта выборочного корреляционного отношения x к y и y к x.

Найдены выборочные корреляционные отношения $\overline{\eta_{xy}} = 0.8649$ и $\overline{\eta_{yx}} = 0.8717$. Выяснено, что выполняются неравенства $\overline{\eta_{xy}} \geq |\overline{r_{xy}}|$ и $\overline{\eta_{yx}} \geq |\overline{r_{xy}}|$. В результате, на основании полученных значений выборочного корреляционного отношения было выдвинуто предположение о корреляционной зависимости признаков, однако зависимость не линейная корреляционная и не функциональная.

Были построены корреляционные кривые параболического и логарифмического вида. Коэффициенты уравнений были найдены с помощью МНК. Исходя из построенных графиков, можно увидеть, что корреляционная зависимость может быть выражена обеими функциями.

3. КЛАСТЕРНЫЙ АНАЛИЗ

3.1. Основные теоретические положения

Задача кластерного анализа заключается в том, чтобы разбить множества исследуемых объектов и признаков на однородные в соответствующем понимании группы. 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.

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:

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

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

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

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

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

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.

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.

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

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.

Идея метода поиска сгущений 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.

- Нормирование

Исходная выборка представлена в таблице 22.

Таблица 22

№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***
1	481	135.2	***21***	418	131.4	***41***	513	159.3	***61***	450	122.3	***81***	475	143.6
2	445	124.7	***22***	378	103.8	***42***	489	149.8	***62***	468	128.9	***82***	518	144.4
3	550	147.9	***23***	521	154.9	***43***	474	132.5	***63***	441	122.8	***83***	566	175.7
4	465	140.9	***24***	394	117.7	***44***	379	94.6	***64***	460	140.7	***84***	464	131.3
5	566	168.5	***25***	504	145.3	***45***	472	135.6	***65***	480	117.7	***85***	394	112.1
6	497	147.3	***26***	440	126.7	***46***	544	169.6	***66***	429	112.9	***86***	480	146.1
7	478	136.6	***27***	465	114.8	***47***	507	142.4	***67***	457	126.4	***87***	321	86.1
8	521	139.6	***28***	418	109.3	***48***	409	116.7	***68***	464	143.2	***88***	502	132.5
9	352	84.9	***29***	418	118.6	***49***	498	164.0	***69***	431	125.0	***89***	460	122.4
10	422	117.9	***30***	465	127.7	***50***	468	142.0	***70***	424	119.0	***90***	458	104.7
11	506	153.5	***31***	447	117.5	***51***	593	187.4	***71***	502	137.2	***91***	362	111.7
12	443	122.9	***32***	433	131.5	***52***	523	152.6	***72***	465	140.7	***92***	503	148.5
13	434	140.4	***33***	460	136.8	***53***	478	126.6	***73***	492	137.5	***93***	446	144.0

14	422	108.6	***34***	382	98.8	***54***	438	122.2	***74***	446	128.4	***94***	421	115.1
15	569	157.4	***35***	532	160.6	***55***	423	115.9	***75***	482	136.4	***95***	407	110.5
16	439	119.2	***36***	482	148.2	***56***	408	110.0	***76***	510	140.6	***96***	448	137.7
17	437	129.4	***37***	472	122.6	***57***	386	105.8	***77***	434	122.3	***97***	490	139.9
18	461	138.6	***38***	532	158.7	***58***	428	130.3	***78***	623	195.7	***98***	482	141.2
19	351	89.0	***39***	473	137.9	***59***	560	169.8	***79***	468	141.2	***99***	463	129.2
20	390	91.4	***40***	525	148.3	***60***	483	130.3	***80***	471	119.7	***100***	459	145.4

Нормализация была выполнена по методу минимакс:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

После нормализации минимальное и максимальное масштабируемые значения равны 0 и 1 соответственно.

Нормализованная выборка представлена в табл. 23 и отображена на рис. 3.2.1.

Таблица 23

№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***
1	0.53	0.454	***21***	0.321	0.42	***41***	0.636	0.671	***61***	0.427	0.338	***81***	0.51	0.53
2	0.411	0.359	***22***	0.189	0.171	***42***	0.556	0.586	***62***	0.487	0.397	***82***	0.652	0.537
3	0.758	0.569	***23***	0.662	0.632	***43***	0.507	0.43	***63***	0.397	0.342	***83***	0.811	0.819
4	0.477	0.505	***24***	0.242	0.296	***44***	0.192	0.088	***64***	0.46	0.504	***84***	0.474	0.419
5	0.811	0.755	***25***	0.606	0.545	***45***	0.5	0.458	***65***	0.526	0.296	***85***	0.242	0.245
6	0.583	0.563	***26***	0.394	0.377	***46***	0.738	0.764	***66***	0.358	0.253	***86***	0.526	0.552
7	0.52	0.467	***27***	0.477	0.27	***47***	0.616	0.519	***67***	0.45	0.375	***87***	0.0	0.011
8	0.662	0.494	***28***	0.321	0.22	***48***	0.291	0.287	***68***	0.474	0.526	***88***	0.599	0.43
9	0.103	0.0	***29***	0.321	0.304	***49***	0.586	0.714	***69***	0.364	0.362	***89***	0.46	0.338
10	0.334	0.298	***30***	0.477	0.386	***50***	0.487	0.515	***70***	0.341	0.308	***90***	0.454	0.179
11	0.613	0.619	***31***	0.417	0.294	***51***	0.901	0.925	***71***	0.599	0.472	***91***	0.136	0.242
12	0.404	0.343	***32***	0.371	0.421	***52***	0.669	0.611	***72***	0.477	0.504	***92***	0.603	0.574
13	0.374	0.501	***33***	0.46	0.468	***53***	0.52	0.376	***73***	0.566	0.475	***93***	0.414	0.533
14	0.334	0.214	***34***	0.202	0.125	***54***	0.387	0.337	***74***	0.414	0.393	***94***	0.331	0.273
15	0.821	0.654	***35***	0.699	0.683	***55***	0.338	0.28	***75***	0.533	0.465	***95***	0.285	0.231
16	0.391	0.31	***36***	0.533	0.571	***56***	0.288	0.227	***76***	0.626	0.503	***96***	0.421	0.477
17	0.384	0.402	***37***	0.5	0.34	***57***	0.215	0.189	***77***	0.374	0.338	***97***	0.56	0.496
18	0.464	0.485	***38***	0.699	0.666	***58***	0.354	0.41	***78***	1.0	1.0	***98***	0.533	0.508

19	0.099	0.037	***39***	0.503	0.478	***59***	0.791	0.766	***79***	0.487	0.508	***99***	0.47	0.4
20	0.228	0.059	***40***	0.675	0.572	***60***	0.536	0.41	***80***	0.497	0.314	***100***	0.457	0.546

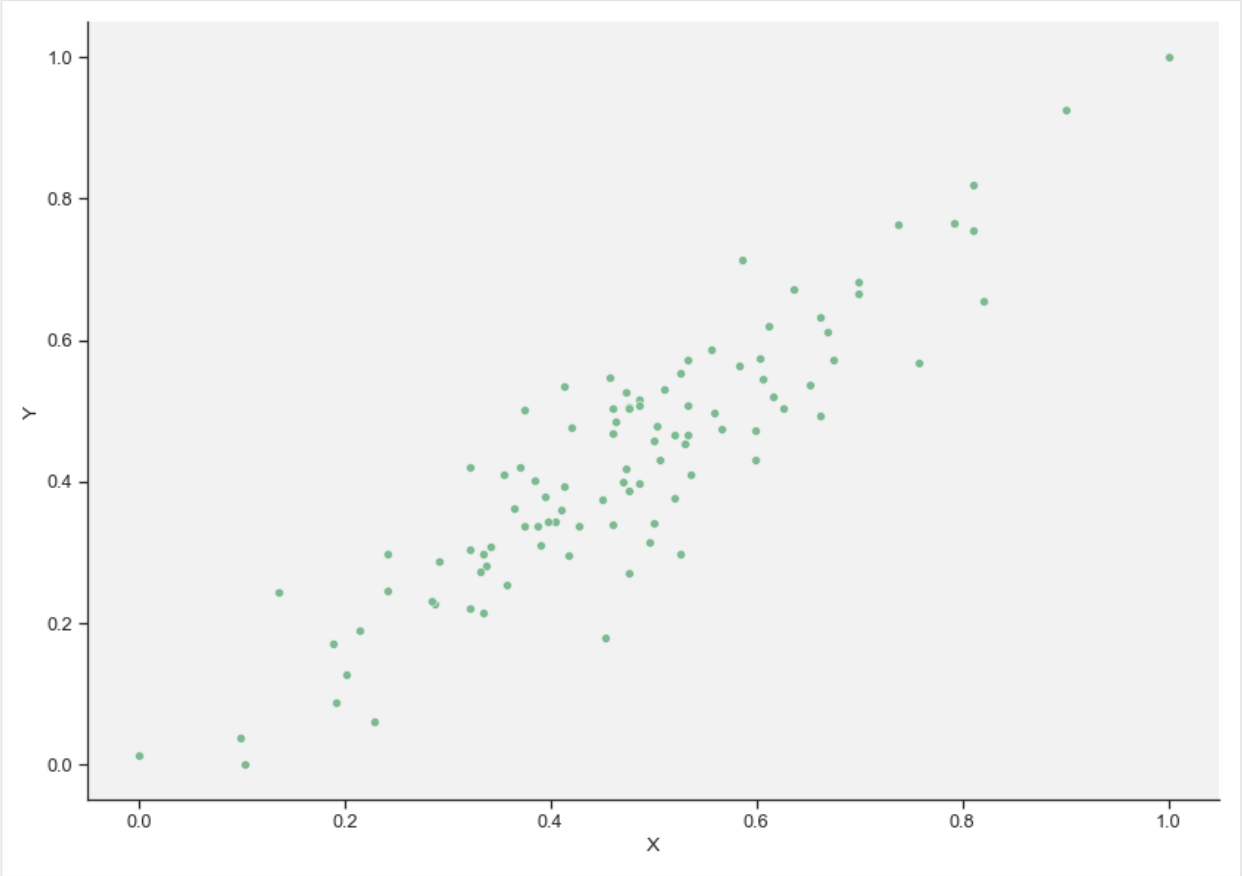

Рисунок 3.2.1 – Нормализованная выборка

- Оценка количества кластеров

  Была найдена верхняя оценка количества кластеров:

$$\bar{k} = \left\lfloor \sqrt{\frac{N}{2}} \right\rfloor = \left\lfloor \sqrt{\frac{104}{2}} \right\rfloor = 7$$

- Метод k-средних

Реализован метод k-средних для количества кластеров $[2;\ 7]$. Полученные кластеры были отображены, выделены разным цветом, были отмечены центроиды, вычислены функционалы качества разбиения. В таблицах представлены количество элементов в кластерах и их центры.

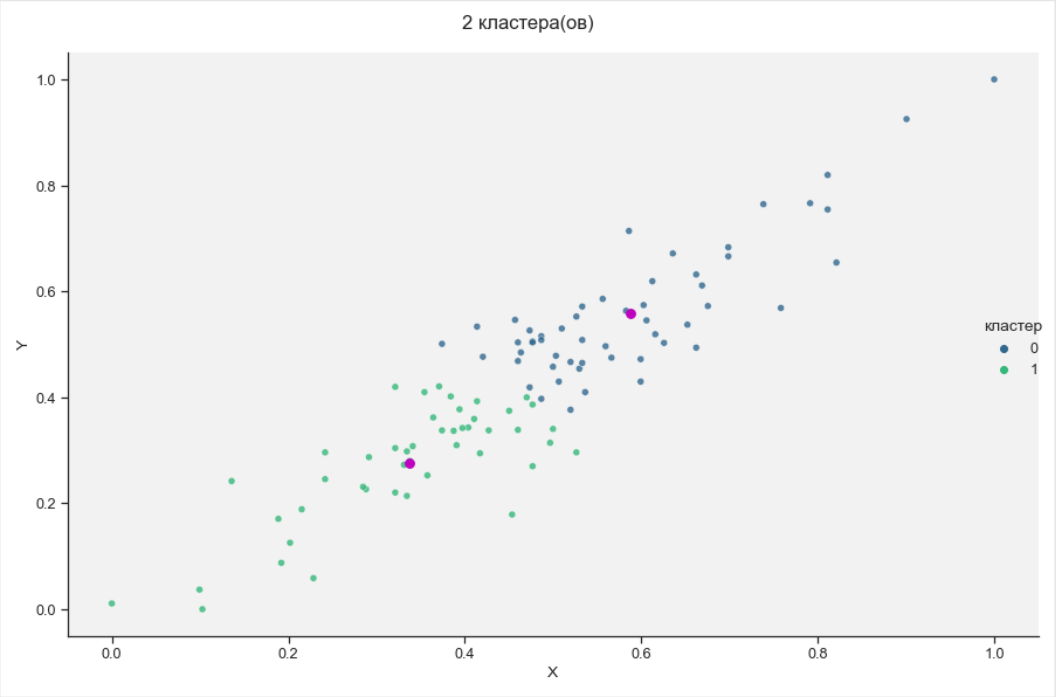

Рисунок 3.2.2 – 2 кластера (3 шага)

Таблица 24

Центр кластера		**Количество элементов**
0.5882	0.5593	54.0
0.3372	0.276	46.0

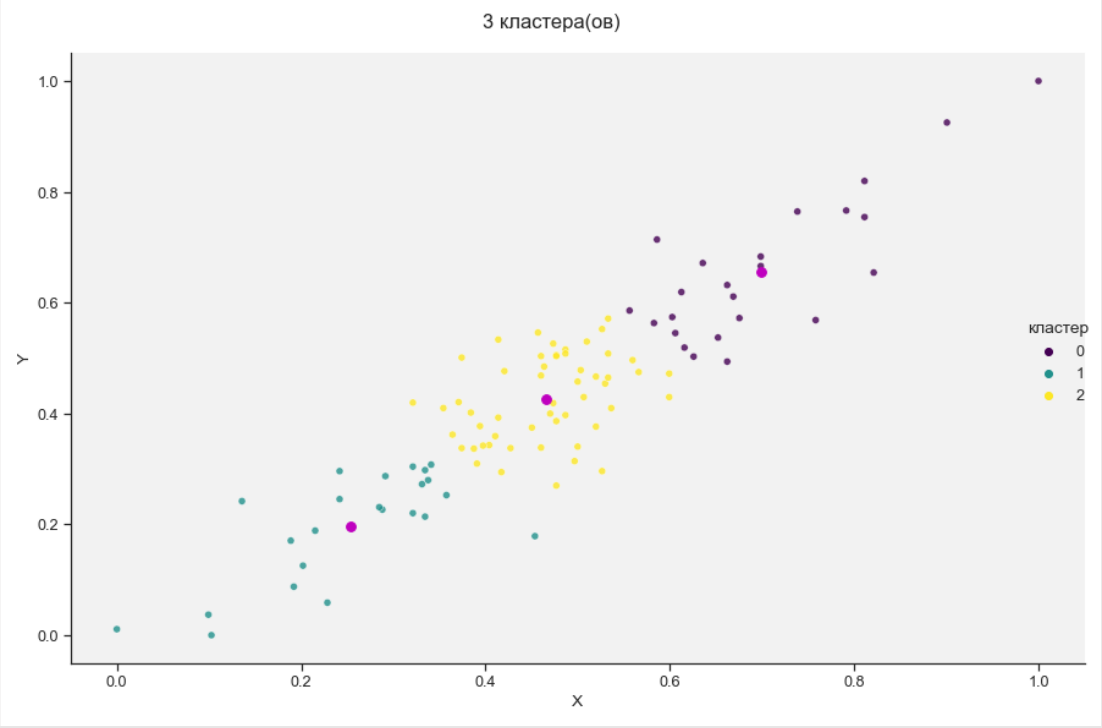

Рисунок 3.2.3 – 3 кластера (5 шагов)

Таблица 25

Центр кластера		Количество элементов
0.699	0.6559	24.0
0.2541	0.1971	23.0
0.4652	0.4268	53.0

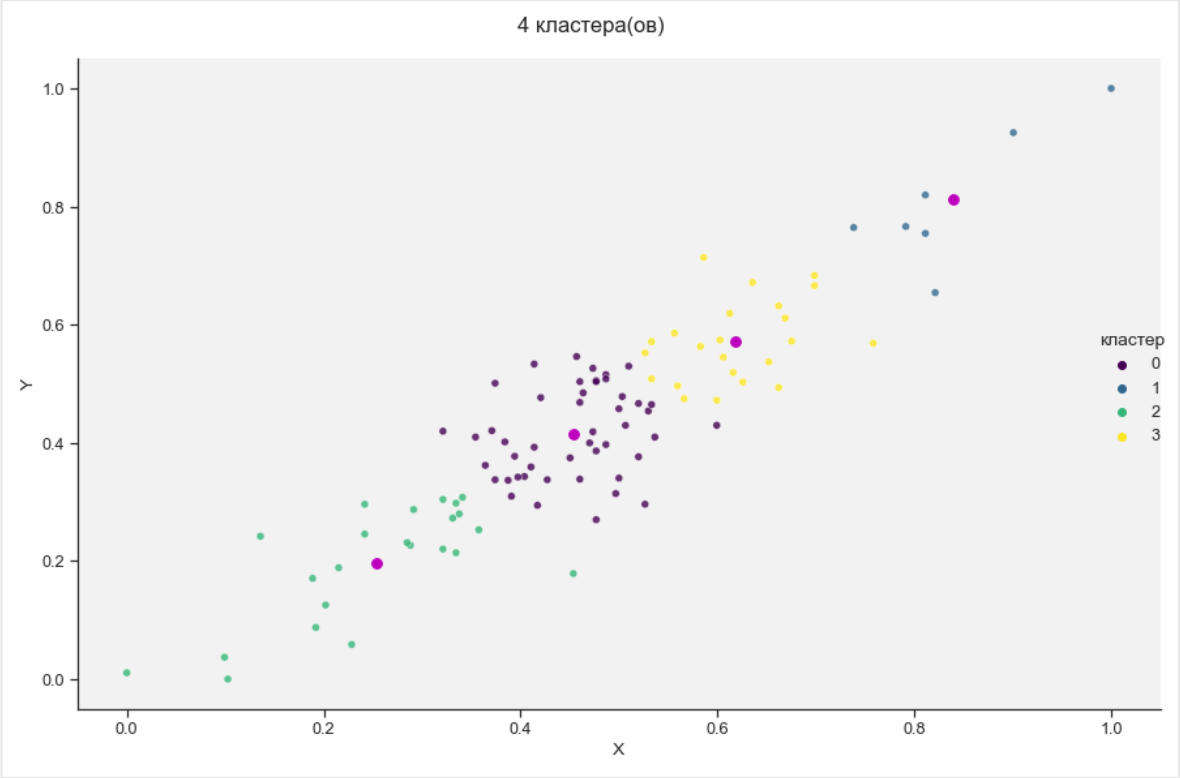

Рисунок 3.2.4 – 4 кластера (8 шагов)

Таблица 26

Центр кластера		Количество элементов
0.454	0.4159	47.0
0.8392	0.812	7.0
0.2541	0.1971	23.0
0.6182	0.571	23.0

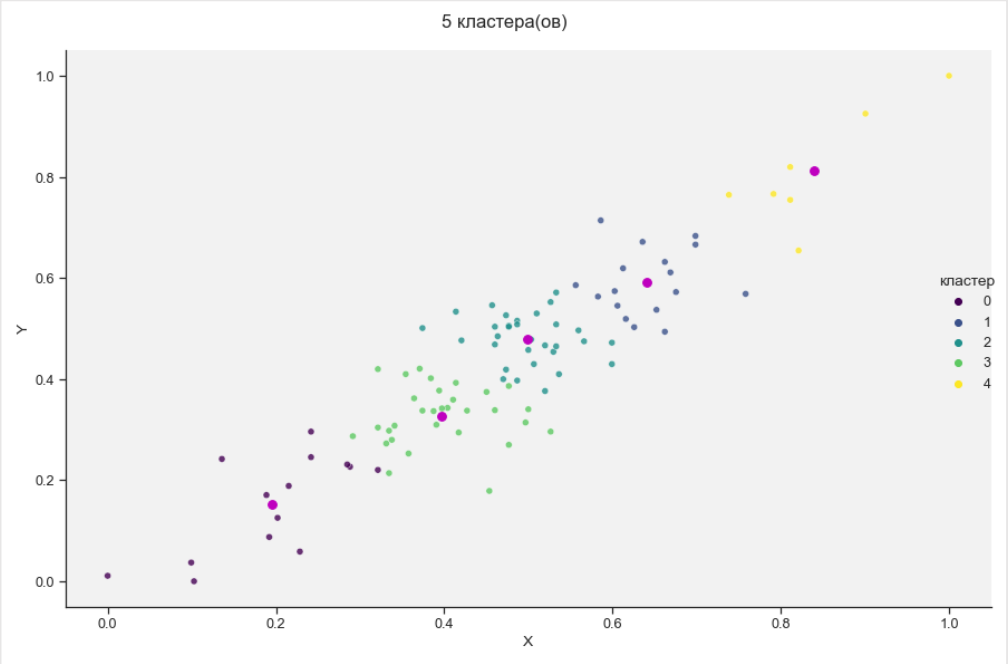

Рисунок 3.2.5 – 5 кластеров (10 шагов)

Таблица 27

Центр кластера		**Количество элементов**
0.1958	0.1528	14.0
0.6412	0.5916	17.0
0.4986	0.4793	31.0
0.3968	0.3276	31.0
0.8392	0.812	7.0

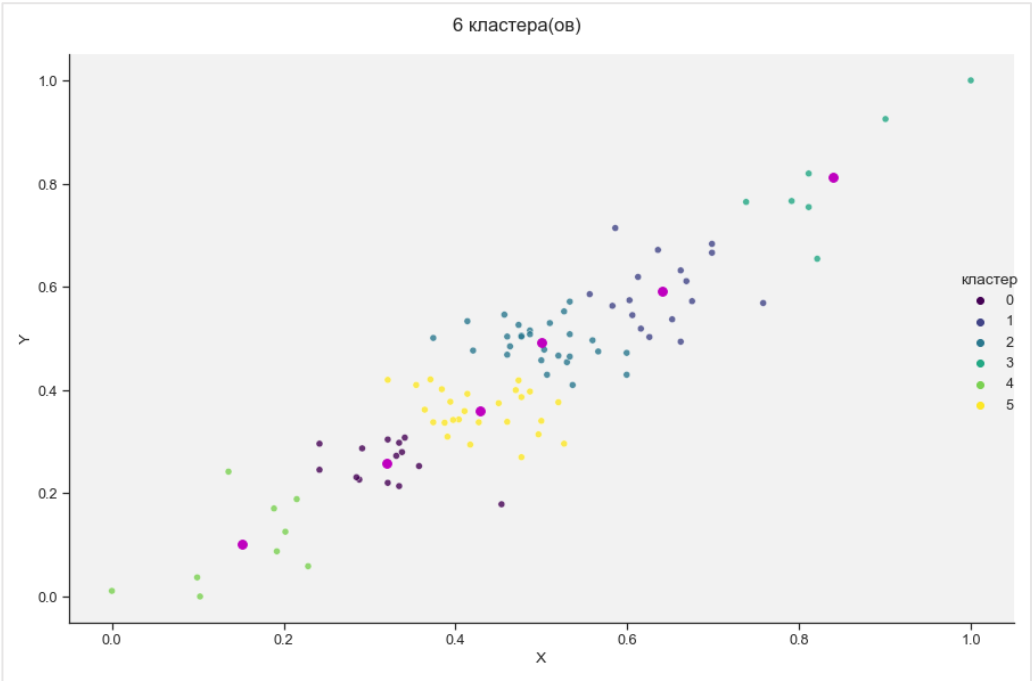

Рисунок 3.2.6 – 6 кластеров (11 шагов)

Таблица 28

Центр кластера		Количество элементов
0.32	0.2581	14.0
0.6412	0.5916	17.0
0.5002	0.4914	27.0
0.8392	0.812	7.0
0.1516	0.1023	9.0
0.4288	0.3598	26.0

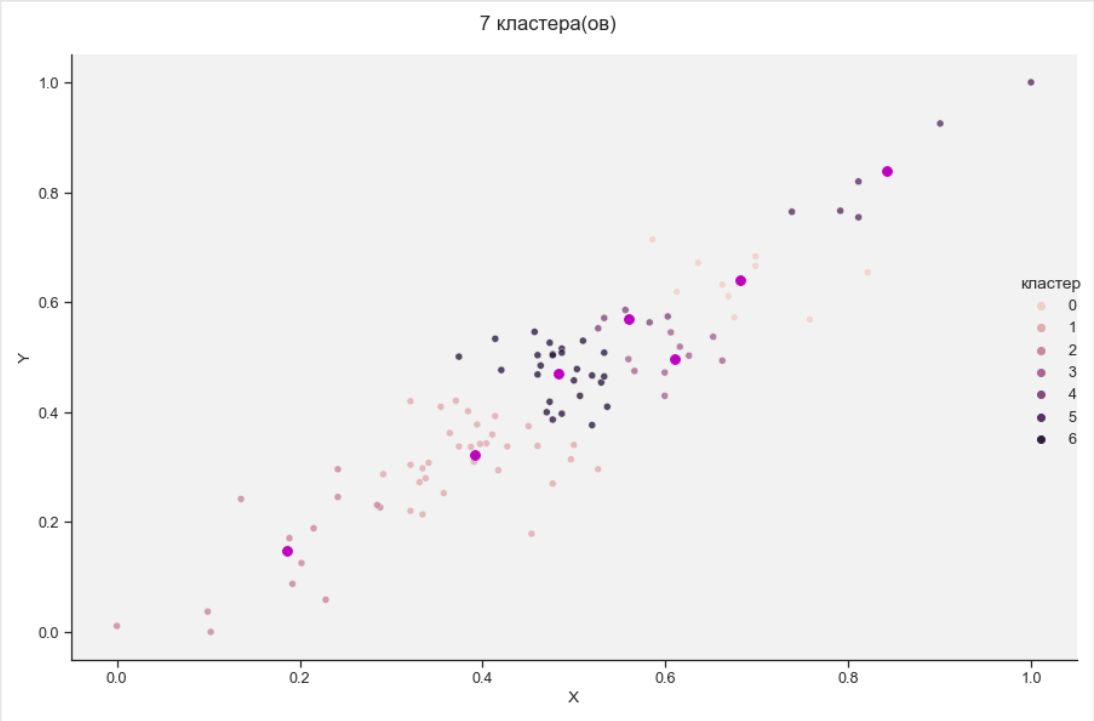

Рисунок 3.2.7 – 7 кластеров (16 шагов)

Таблица 29

Центр кластера		Количество элементов
0.6818	0.6392	10.0
0.3918	0.3223	31.0
0.1862	0.1477	13.0
0.6096	0.4967	9.0
0.5603	0.5693	5.0
0.8422	0.8383	6.0
0.4827	0.4711	26.0

- Оценка качества разбиения

Для каждого разбиения были вычислены функционалы качества F_1, F_2, F_3, которые были определены в начале. В таблице 30 приведены значения функционалов на первой и последней итерациях алгоритма для разного количества кластеров разбиения.

Таблица 30

Количество кластеров	**2**	**3**	**4**	**5**	**6**	**7**
F_1	2.92	1.948	2.045	1.791	1.134	1.27
F_1 **(конец)**	2.891	1.645	1.171	0.812	0.625	0.56
F_2	151.736	103.754	95.325	75.691	22.247	29.21
F_2 **(конец)**	146.915	54.414	34.918	16.851	10.607	7.662
F_3	0.058	0.061	0.07	0.066	0.065	0.063
F_3 **(конец)**	0.058	0.06	0.058	0.057	0.054	0.055

Из таблицы можно увидеть, что при увеличении числа кластеров, минимизируются все функционалы, а также, то насколько сильно они меняются в сравнении с первой итерацией.

Алгоритм был реализован в двух вариантах: в первом, который был представлен выше, центр пересчитывается только по завершении шага процедуры, второй же вариант предполагает изменение центра кластера после обработки каждого объекта. Сравнение алгоритмов представлено в таблице 31.

Таблица 31

Количество кластеров	**2**	**3**	**4**	**5**	**6**	**7**
Количество итераций (первый алгоритм)	3	5	8	10	11	16
Количество итераций (второй алгоритм)	2	2	3	4	3	3

Из таблицы видно, что при увеличении количества кластеров увеличивается число итераций, а также, что количество итераций второго

алгоритма меньше, чем первого, что связано с тем, что центр меняется после обработки каждого объекта.

3.3. Метод поиска сгущений

Реализован метод поиска сгущений. Полученные кластеры были отображены на рисунке, отмечены разными цветами, а также были отмечены их центроиды. Определены нижняя и верхняя границы радиуса сферы:

$$R_{min} = 0.00181;\ R_{max} = 1.40658$$

Значение R было выбрано $R = 0.25$, так как оно позволяет достичь стабильного разбиения на четыре кластера.

Формирование кластеров представлено на рис. 3.3.1 – 3.3.11. На рисунках текущий кластер выделен фиолетовым, оставшиеся элементы – оранжевым, центроид – красным.

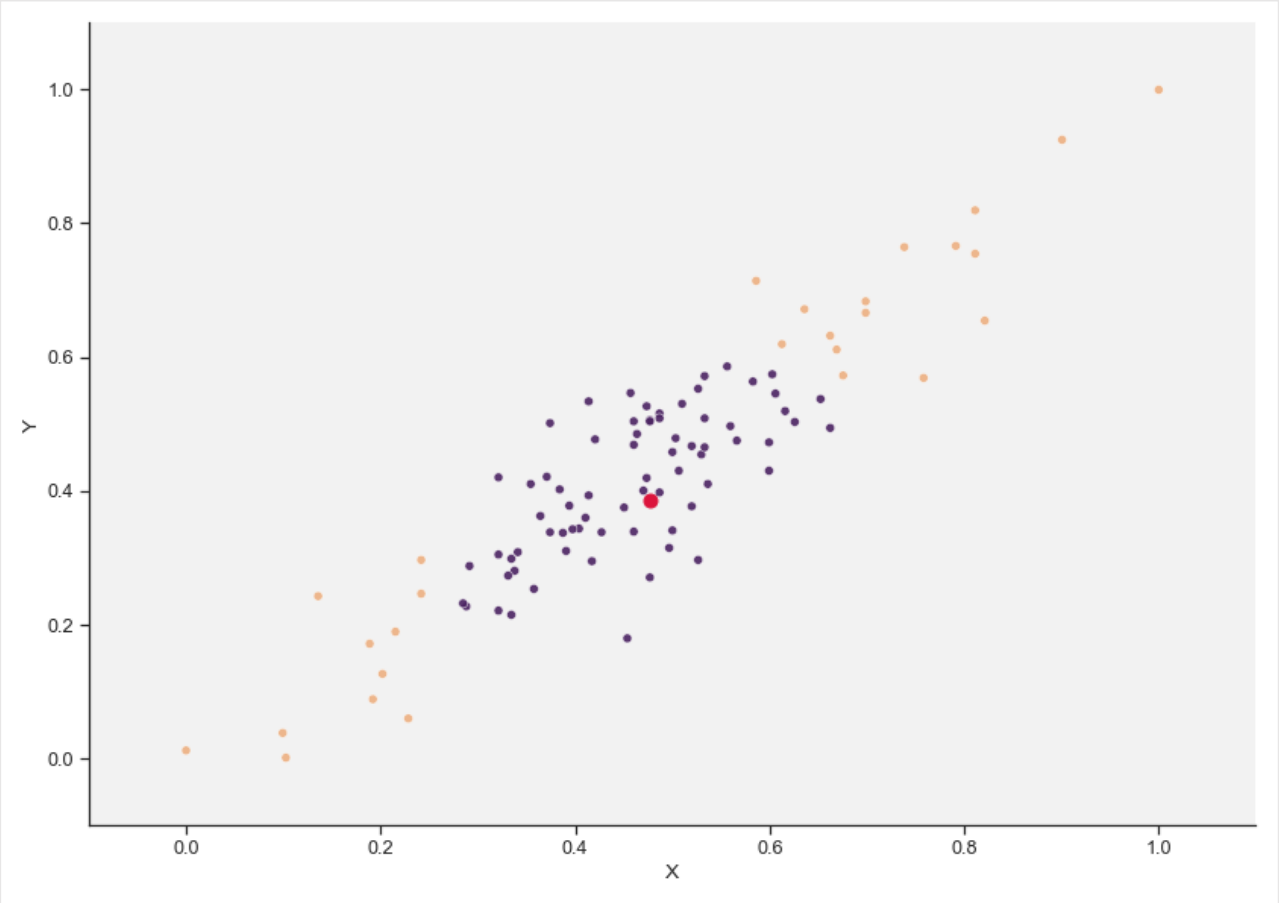

Рисунок 3.3.1 – Первый кластер, шаг 1

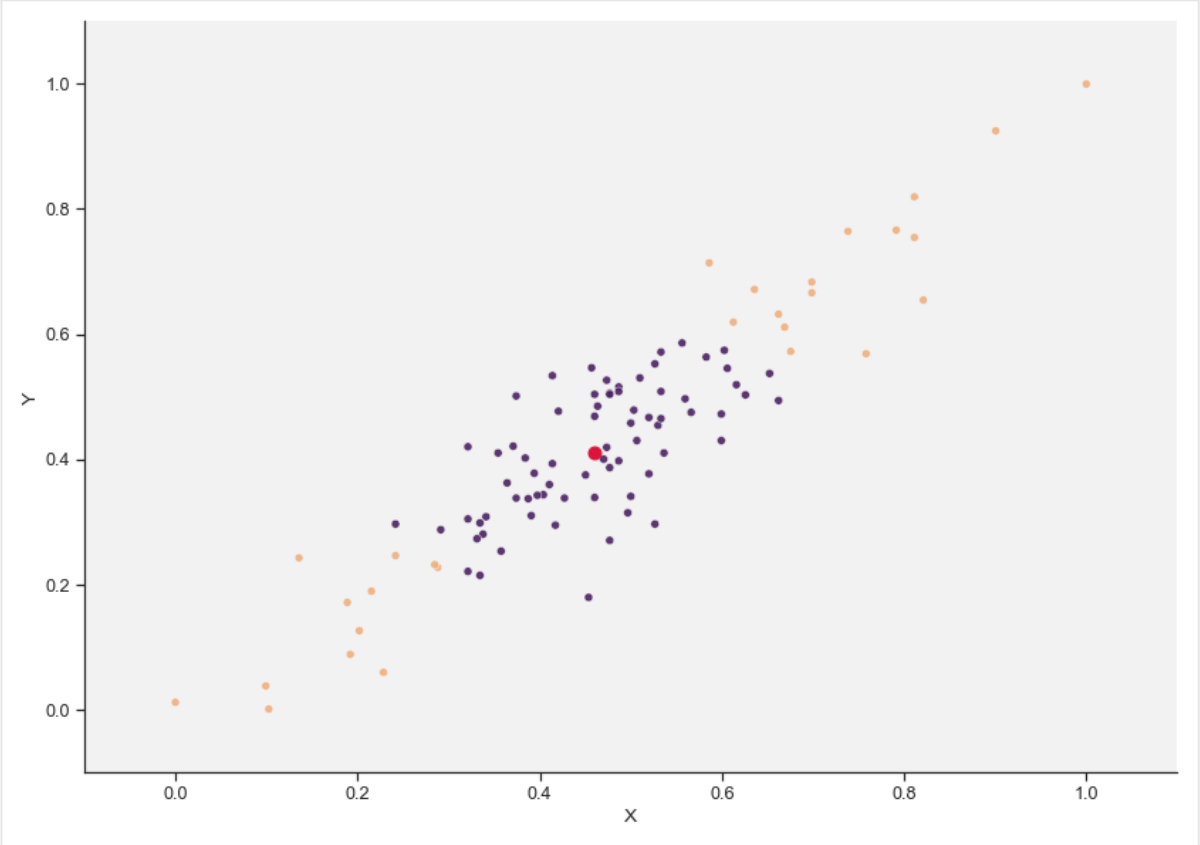

Рисунок 3.3.2 – Первый кластер, шаг 2

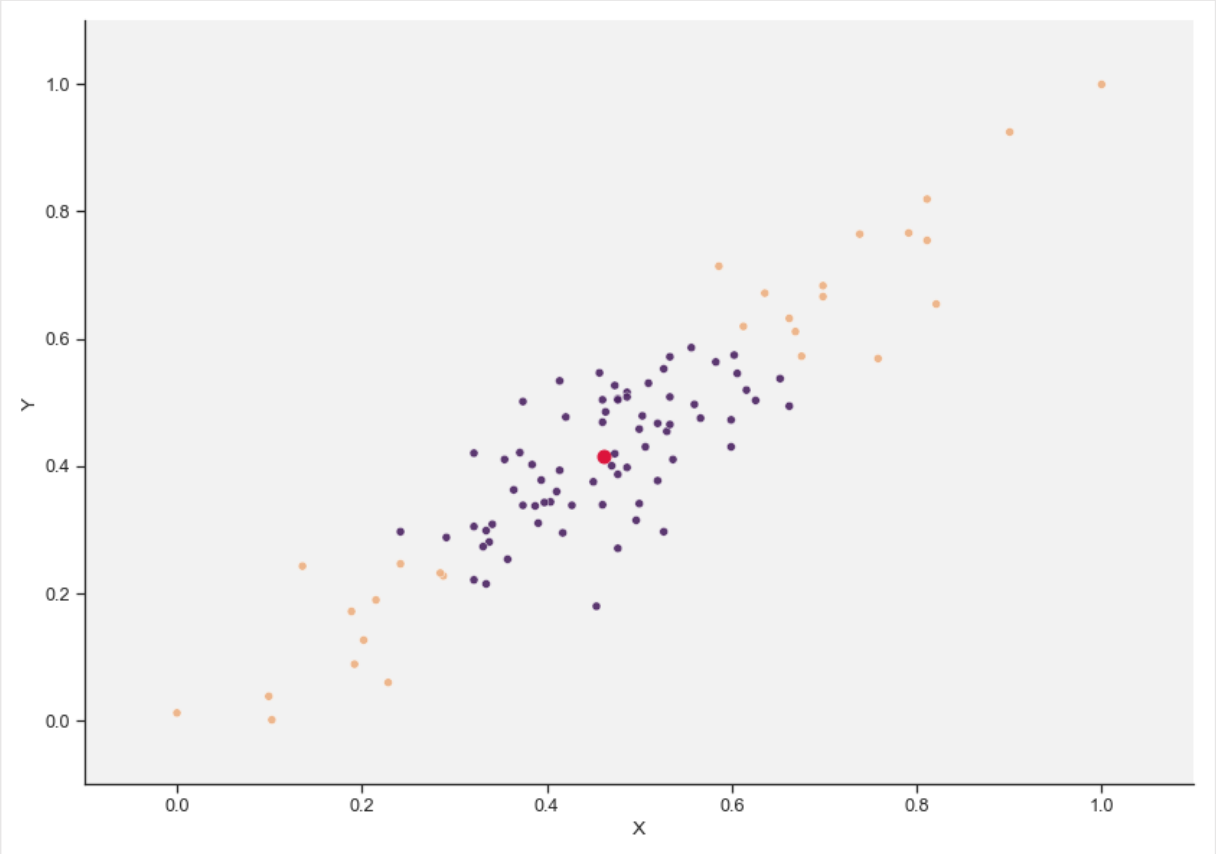

Рисунок 3.3.3 – Первый кластер, шаг 3

Таблица 32 – Первый кластер

Шаг	Центр		Количество элементов
1	0.47682	0.38628	73
2	0.45968	0.41116	72
3	0.46146	0.41462	72

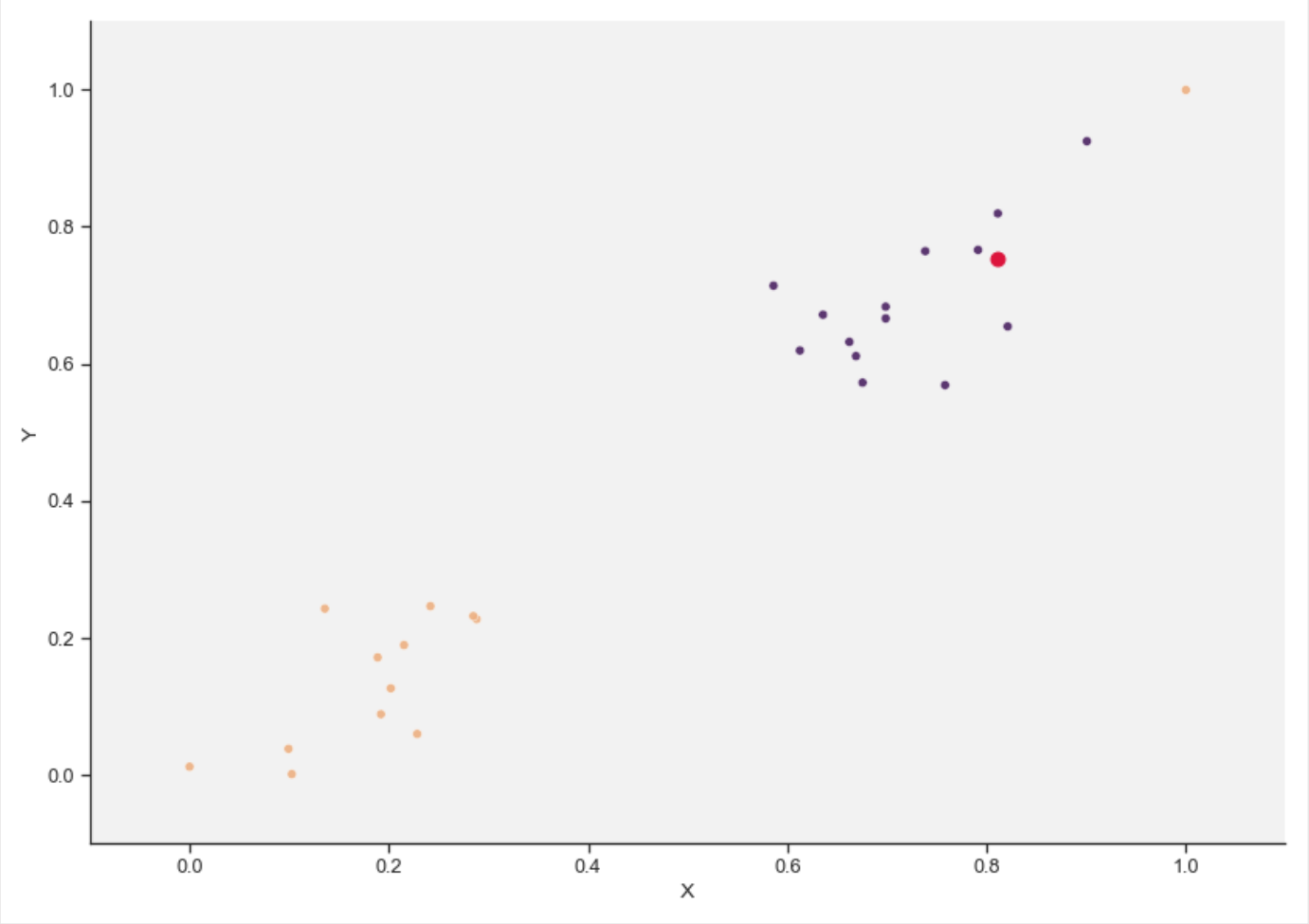

Рисунок 3.3.4 – Второй кластер, шаг 1

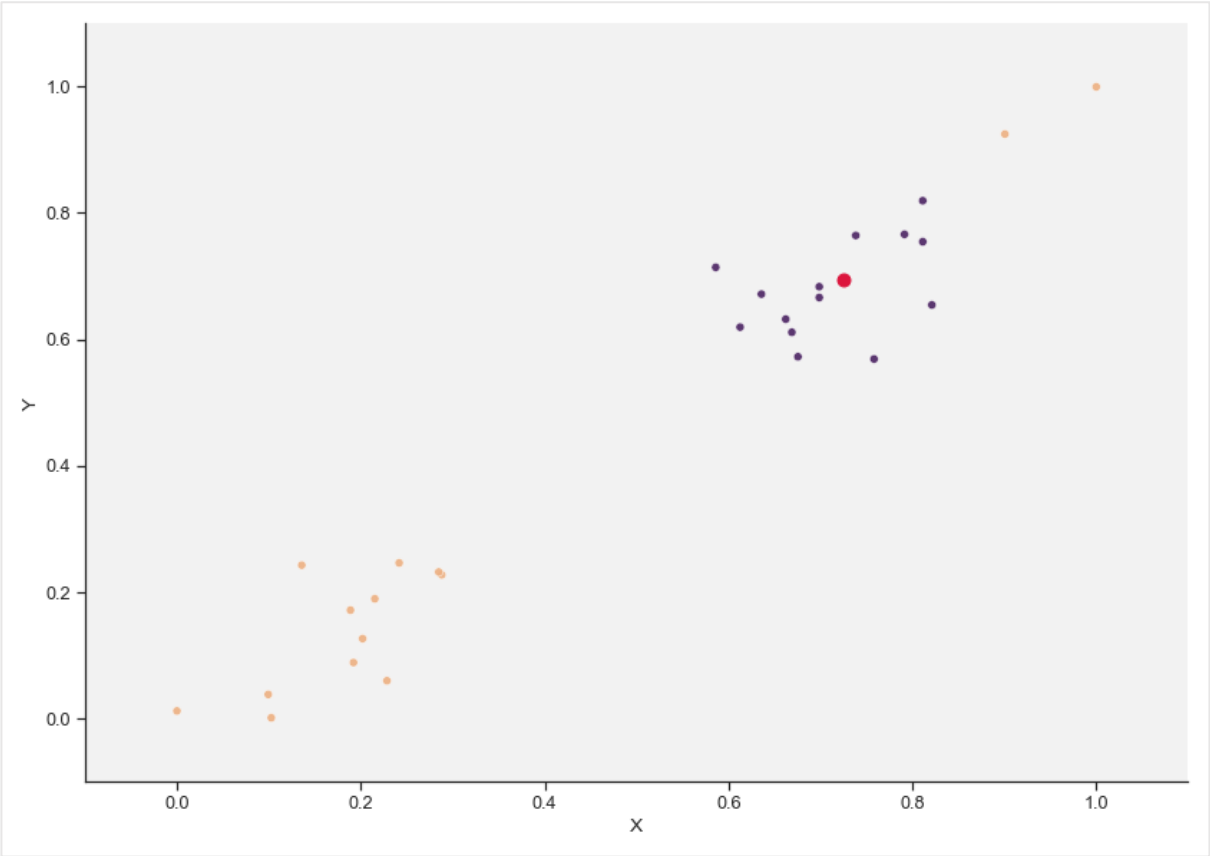

Рисунок 3.3.5 – Второй кластер, шаг 2

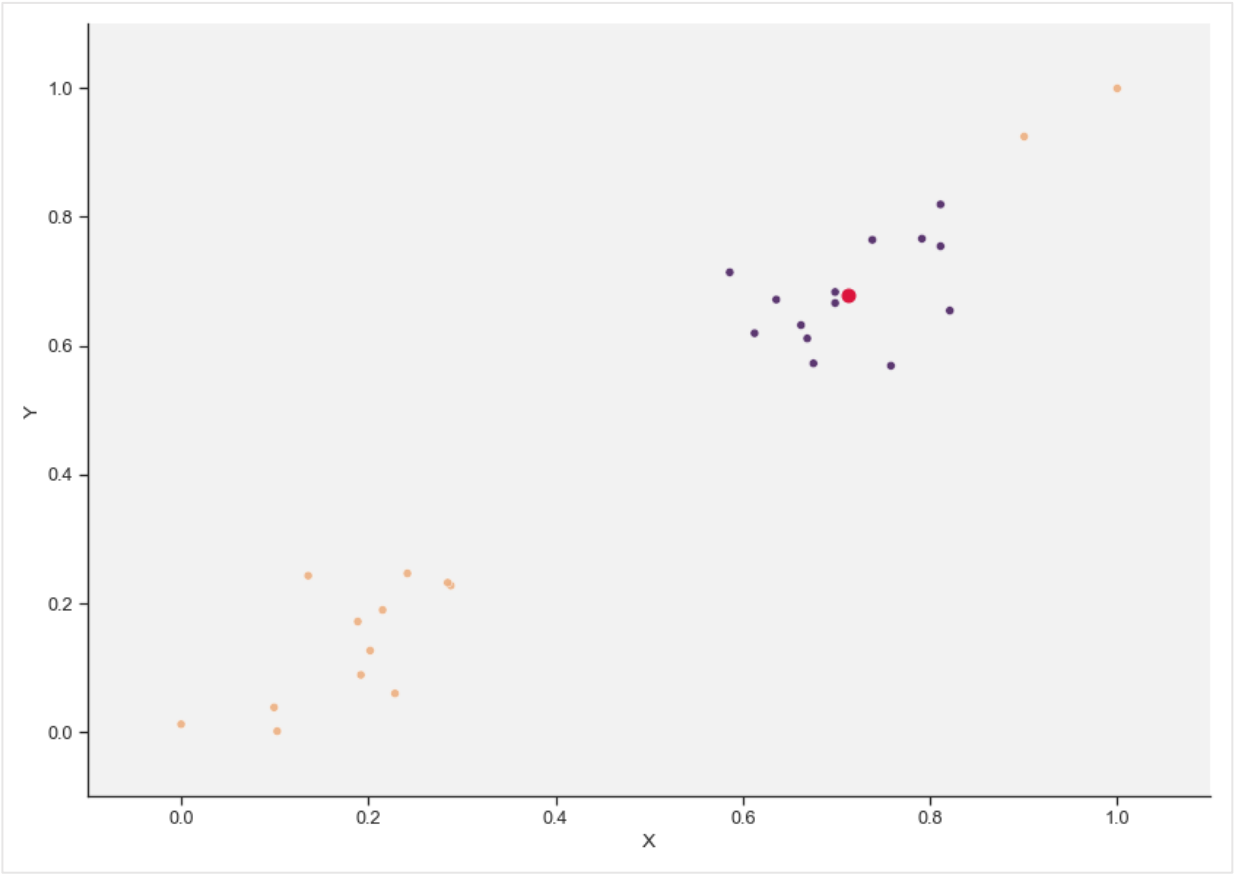

Рисунок 3.3.6 – Второй кластер, шаг 3

Таблица 33 – Второй кластер

Шаг	**Центр**		**Количество элементов**
1	0.81126	0.75451	15
2	0.72472	0.69477	14
3	0.71216	0.67831	14

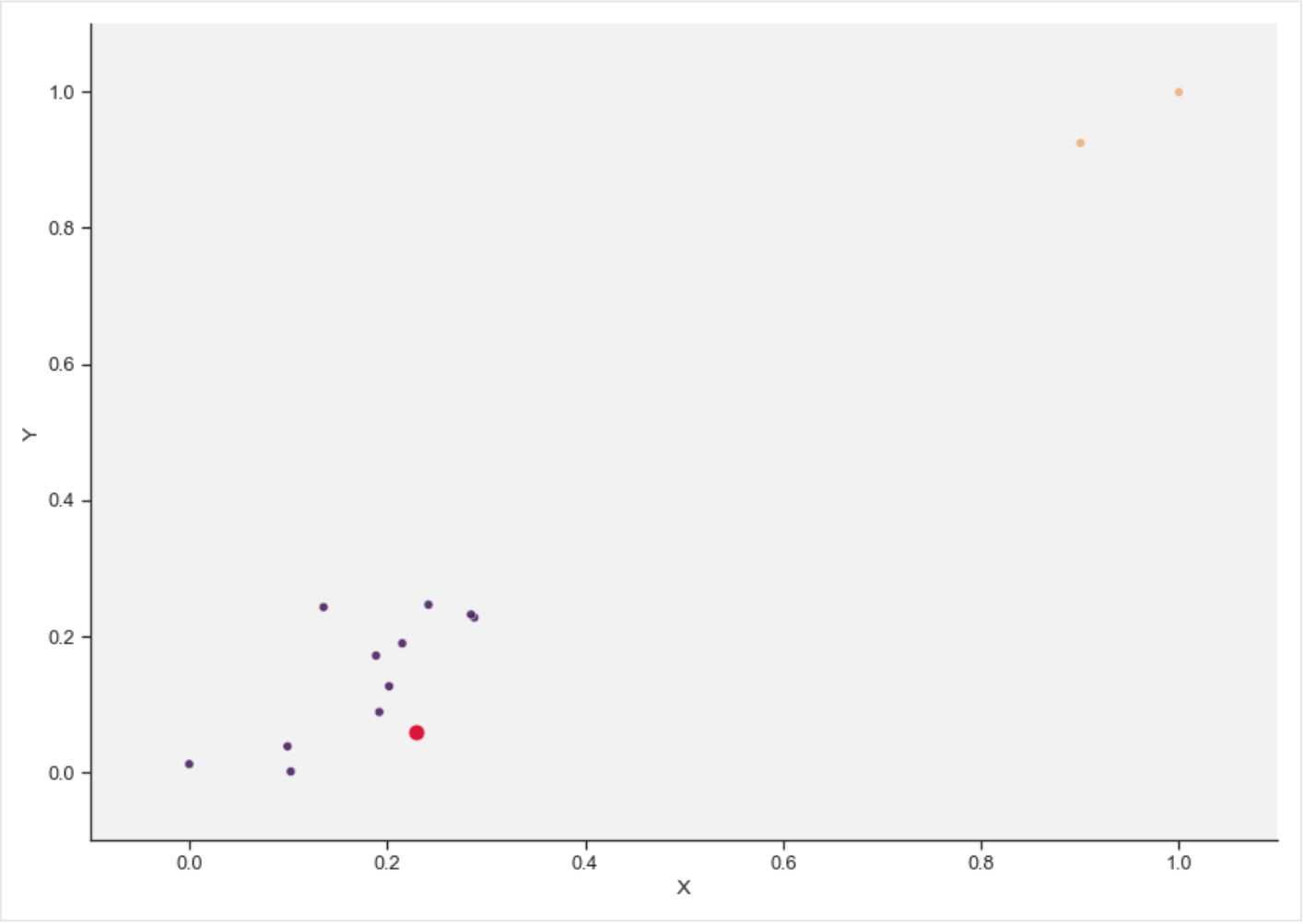

Рисунок 3.3.7 – Третий кластер, шаг 1

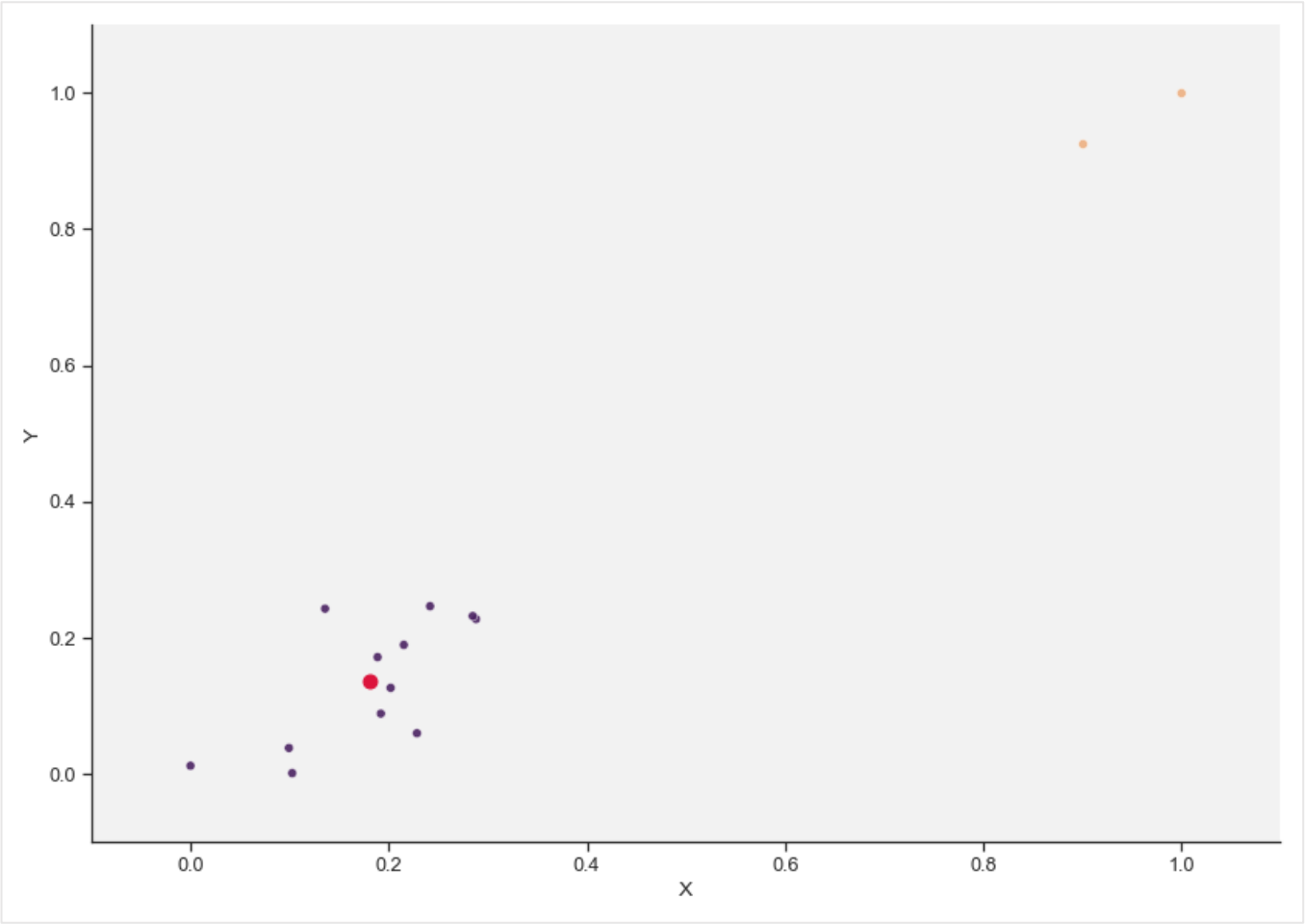

Рисунок 3.3.8 – Третий кластер, шаг 2

Таблица 34 – Третий кластер

Шаг	**Центр**		**Количество элементов**
1	0.22848	0.05867	12
2	0.18157	0.1353	12

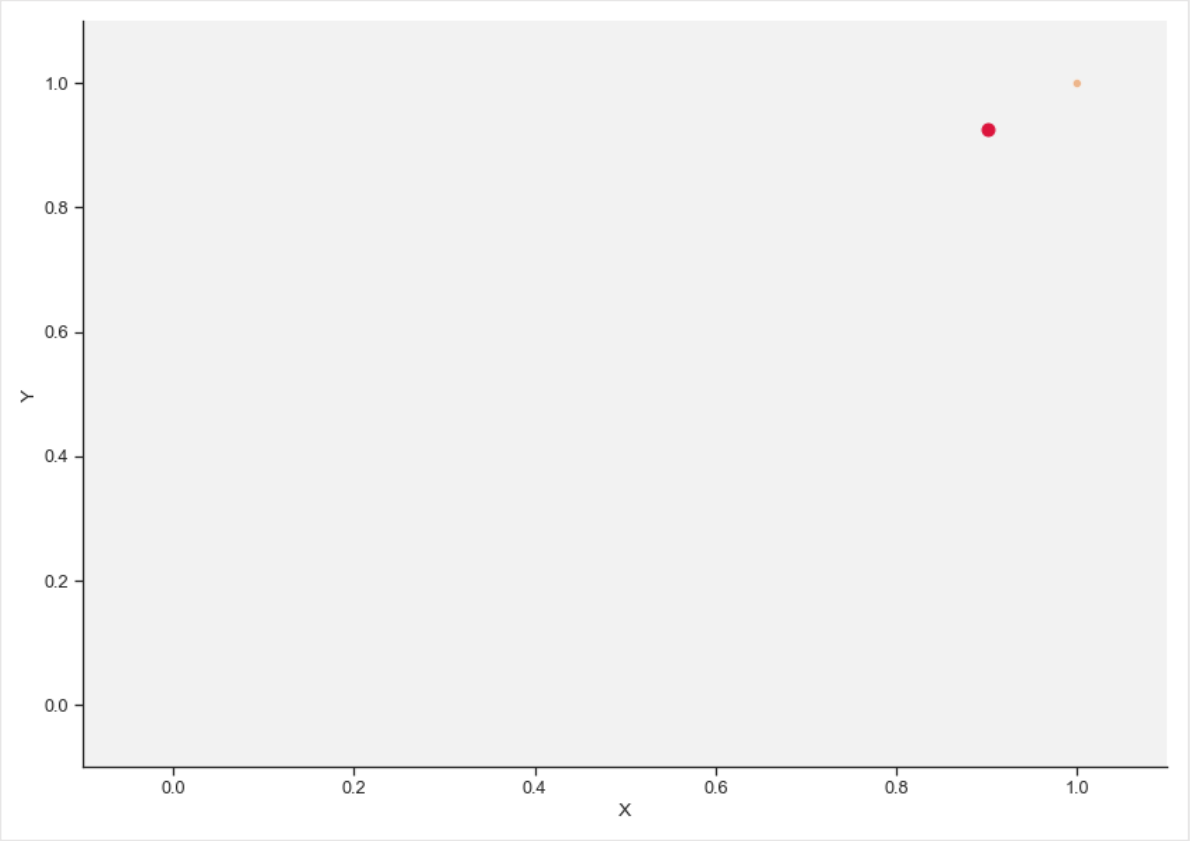

Рисунок 3.3.9 – Четвертый кластер, шаг 1

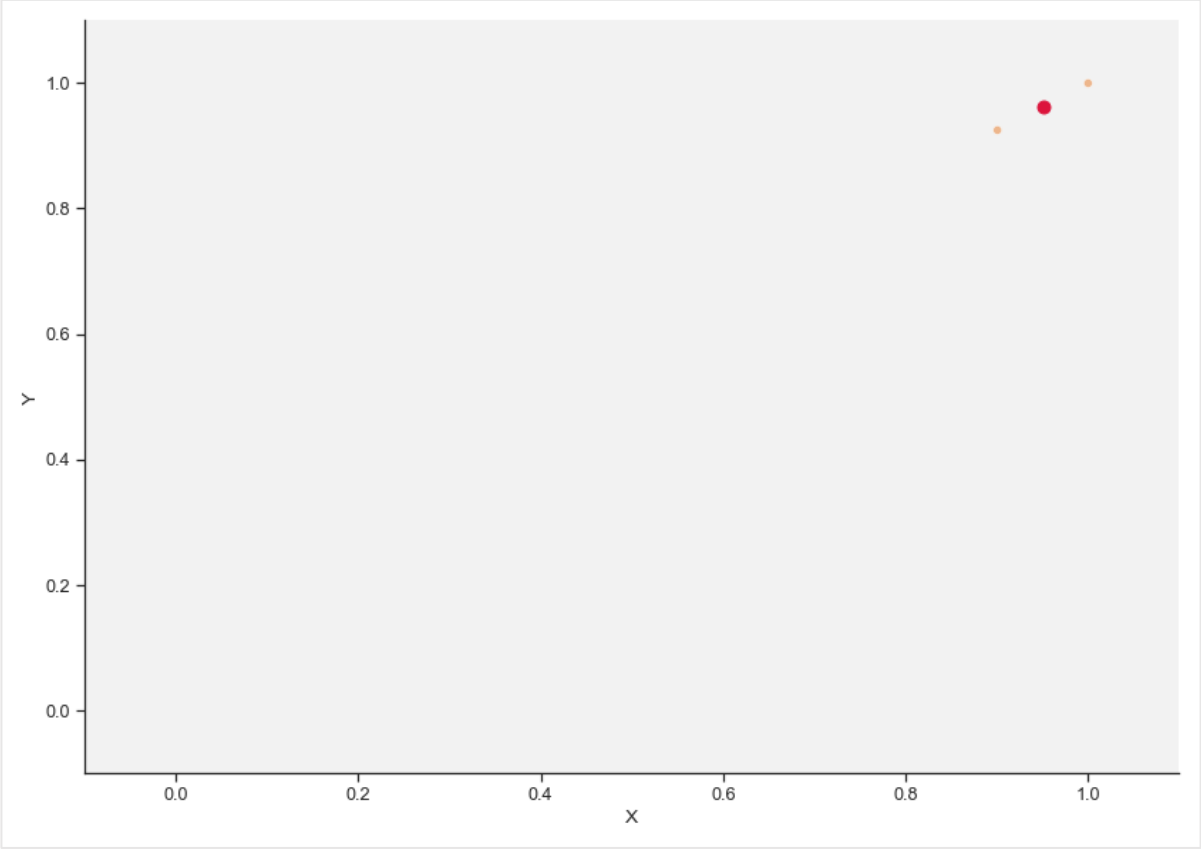

Рисунок 3.3.10 – Четвертый кластер, шаг 2

Таблица 35 – Четвертый кластер

Шаг	**Центр**		**Количество элементов**
1	0.90067	0.92509	2
2	0.95033	0.96255	2

Результат кластеризации представлен на рис. 3.3.11.

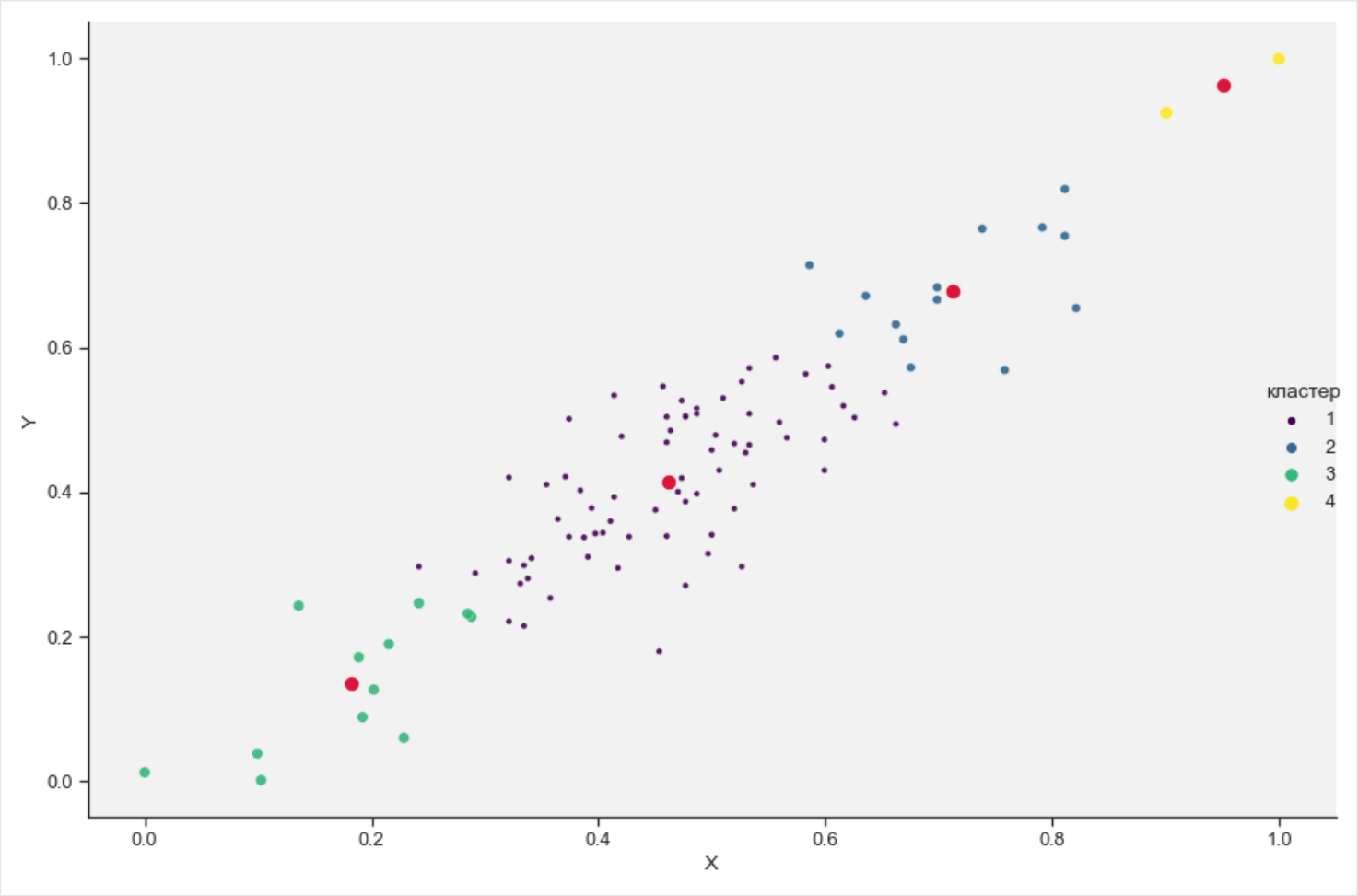

Рисунок 3.3.11 – Результат кластеризации

Несмотря на то, что имели место быть пересечения радиусов, спорные объекты не возникали. Это достигается данной реализацией алгоритма, при которой точки, попавшие в кластер, убираются для последующих итераций. А центр первого кластера выбирается как имеющий наибольшее число соседей.

- Чувствительность к погрешностям

Проведена проверка чувствительности метода к погрешностям. Радиус $R = 0.25$ был изменен на небольшое число, после чего сделано сравнение функционалов качества. Результаты представлены в таблице 36.

Таблица 36

Радиус	$\boldsymbol{F_1}$	$\boldsymbol{F_2}$	$\boldsymbol{F_3}$
$\mathbf{0.23}$	1.43236	78.98116	0.04956
$\mathbf{0.24}$	1.48355	79.92584	0.05225
$\boldsymbol{R = 0.25}$	1.67371	100.52548	0.05433
$\mathbf{0.26}$	1.7756	112.27951	0.05507
$\mathbf{0.27}$	1.86215	124.57263	0.05457

Из таблицы видно, что при изменении радиуса на небольшое значение функционалы качества изменяются на существенное значение, поэтому можно сделать вывод, что метод чувствителен к погрешностям.

- Сравнение методов

Сравним методы кластеризации с помощью значений функционалов и графиков конечного разбиения при количестве кластеров равном 4. Значения функционалов представлены в таблице 37. Графики представлены на рис. 3.3.12, 3.3.13.

Таблица 37

Метод	$\boldsymbol{F_1}$	$\boldsymbol{F_2}$	$\boldsymbol{F_3}$
k-средних	1.171	34.918	0.058
Поиск сгущений	1.67371	100.52548	0.05833

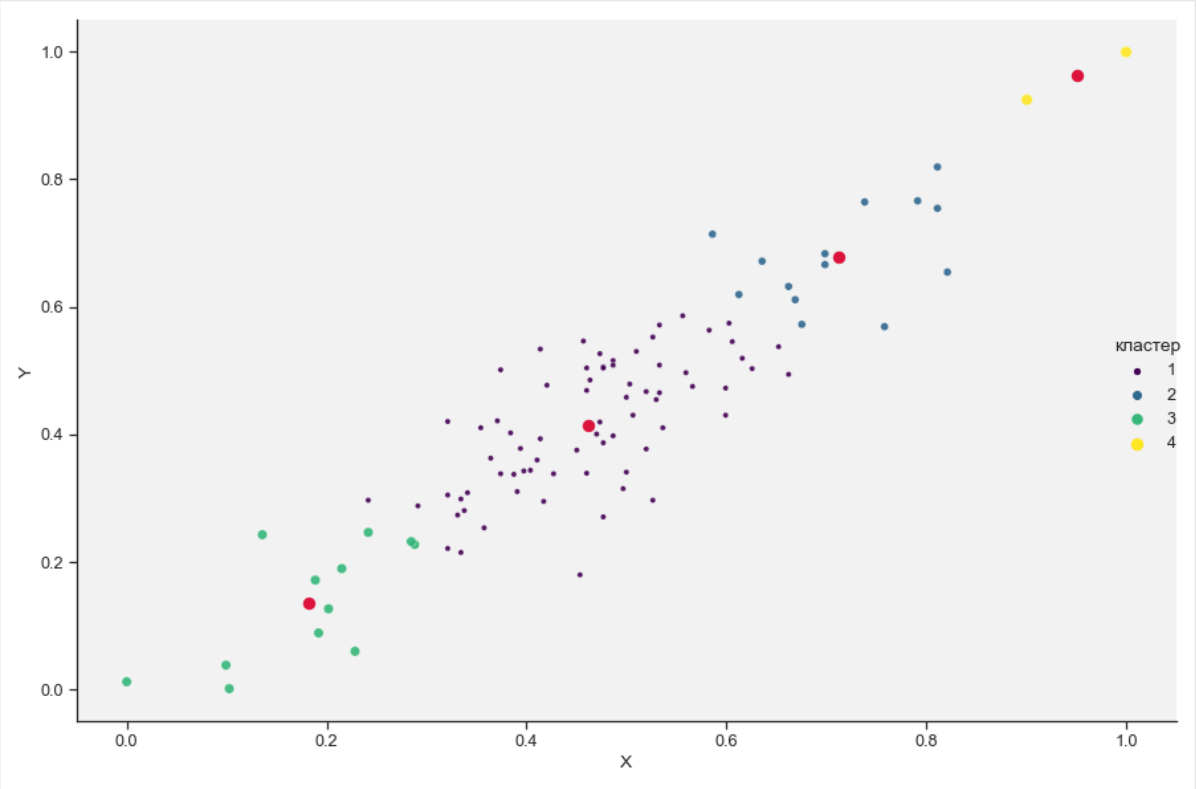

Рисунок 3.3.12 – Метод поиска сгущений

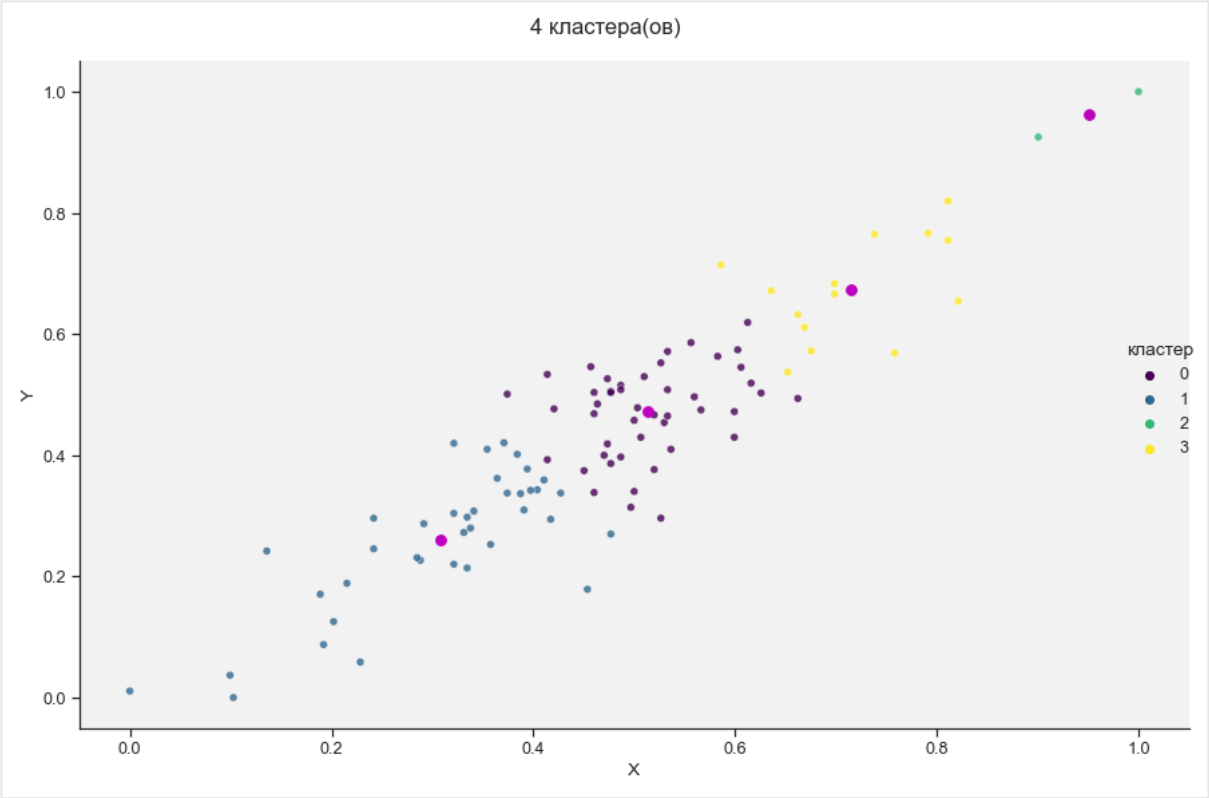

Рисунок 3.3.13 – Метод k-средних

Исходя из таблицы можно увидеть, что метод k-средних показал себя намного лучше, чем метод поиска сгущений, так как значения функционалов качества значительно меньше.

На рисунках же можно увидеть, что оба метода примерно одинаково определили два правых верхних кластера, но есть различия в двух других. Кластеры метода k-средних более сбалансированы и расстояние между точками в нем меньше по сравнению с методом поиска сгущений.

3.4. Выводы.

Освоены основные понятия кластерного анализа и метода k-means. Первоначальная двумерная выборка была нормализована методом минимакс, после которого минимальное и максимальное масштабируемые значения равны 0 и 1 соответственно. Нормализованная выборка была отображена на рисунке.

С помощью алгоритма k-средних было произведено разбиение выборки на 2, 3, …, 7 кластеров. Для каждого разбиения были выведены центры кластеров и количество элементов в кластерах. Было оценено качество разбиения с помощью функционалов качества. В сравнительной таблице можно заметить, что при увеличении числа кластеров, минимизируются все функционалы качества, а также, то насколько сильно они меняются при сравнении первой и последней итераций.

Алгоритм k-средних был реализован в двух вариантах. В первом, центр пересчитывается только по завершении шага процедуры, второй же вариант предполагает изменение центра кластера после обработки каждого объекта. Из сравнительной таблицы можно заметить, что при увеличении количества кластеров увеличивается число итераций, а также, что количество итераций второго варианта алгоритма меньше, чем первого, это связано с тем, что центр меняется после обработки каждого элемента.

Освоены основные понятия метода поиска сгущений. Сначала была произведена нормализация множества точек с помощью метода минимакс, так что минимальное значение равно нулю, а максимальное единице. Были найдены границы радиуса сферы:

$$R_{min} = 0.00181; \qquad R_{max} = 1.40658$$

С помощью реализованного метода поиска сгущений для $R = 0.25$ выборка была разбита на четыре кластера. Все шаги алгоритма были отображены, текущий кластер был выделен цветом.

Несмотря на то, что имели место быть пересечения радиусов, спорные объекты не возникали. Это достигается реализацией алгоритма, при которой точки, попавшие в кластер, убираются для последующих итераций. А центр первого кластера выбирается как имеющий наибольшее число соседей.

Проведена проверка чувствительности метода к погрешностям. Из сравнительной таблицы видно, что при изменении радиуса на небольшое значение функционалы качества изменяются на существенное значение, поэтому можно сделать вывод, что метод чувствителен к погрешностям.

Проведено сравнение методов кластеризации. Исходя из сравнительной таблицы можно увидеть, что метод k-средних показал себя намного лучше, чем метод поиска сгущений, так как значения функционалов качества значительно меньше. На рисунках же можно увидеть, что кластеры метода k-средних более сбалансированы и расстояние между точками в нем меньше по сравнению с методом поиска сгущений, что тоже лучше.

ЗАКЛЮЧЕНИЕ

В ходе выполнения курсовой работы были выполнены все первоначальные задачи: построена выборка 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, СКВО, асимметрии, эксцесса, медианы и моды, построены доверительные интервалы для математического ожидания и СКВО, проверена гипотеза о нормальном законе с помощью критерия хи-квадрат. Построена корреляционная таблица, найдена оценка коэффициента корреляции, проверена гипотеза о равенстве коэффициента корреляции нулю, построены уравнения выборочных прямых среднеквадратической регрессии, найдены оценки корреляционных отношений. 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. Белоногов А.М., Попов Ю.И., Посредник О.В. Статистическая обработка результатов физического эксперимента [Комплект] : учеб. пособие: - СПб. : Изд-во СП6ГЭТУ "ЛЭТИ", 2009.

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

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

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

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

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

9. Метод поиска сгущений // csaa.ru URL: http://csaa.ru/metodika-klasternogo-analiza/ (дата обращения: 07.04.2022).

10. Метод наименьших квадратов// machinelearning.ru URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BD%D0%B0%D0%B8%D0%BC%D0%B5%D0%BD%D1%8C%D1%88%D0%B8%D1%85_%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%BE%D0%B2 (дата обращения: 07.04.2022).

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
df.to_csv('data/data1.csv', index=False)
n = len(df)
df2 = df.drop('E', axis=1)
df2.to_csv('data/data2.csv', index=False)
df2.head()
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']
X.mode()
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)
k = 1+3.31*np.log10(n)
k = int(np.floor(k))
min(X), max(X)
h = (max(X)-min(X))/k
h = int(np.ceil(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.iloc[76, 3] = data_interval.iloc[75, 3]
data_interval['inter'].value_counts().sort_index()
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['rf'] = np.round(f_inter['rf'], 2)
```

```
f_inter.to_csv('data/data7.csv', index=False)
sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style('ticks', {"axes.facecolor": ".94"})
ax = sns.relplot(data=f_inter, x='avg_inter', y='af', kind='line',
                 height=8.27, aspect=11.7/8.27, linewidth=3)
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'])
plt.savefig('pics/3.png')
ax = sns.displot(data=df, x='nu', bins=ivs, kind='hist',
                 height=8.27, aspect=11.7/8.27, linewidth=3)
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'], yticks=f_inter['af'])
plt.savefig('pics/4.png')
f_inter['sum_rf'] = f_inter['rf'].cumsum()
f_inter
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='w')
for i in range(6):
    plt.hlines(f_inter['sum_rf'][i], f_inter['avg_inter'][i],
f_inter['avg_inter'][i+1], color='r')
plt.hlines(1, 604, 624, color='r')
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='r', linestyle='-')
plt.vlines(343, 0, 0.04, color='r', linestyle='-')
for i in range(6):
    plt.annotate('', xy=(f_inter['avg_inter'][i]-1,
f_inter['sum_rf'][i]),
                 xytext=(f_inter['avg_inter'][i+1],
f_inter['sum_rf'][i]),
                 arrowprops=dict(arrowstyle="->", color='r', lin-
ewidth=3))
plt.annotate('', xy=(604, 1),
                 xytext=(624, 1),
                 arrowprops=dict(arrowstyle="->", color='r', lin-
ewidth=3))
ax.set_axis_labels('Середины интервалов', '')
ax.set(xticks=f_inter['avg_inter'])
plt.savefig('pics/5.png')
ax = sns.relplot(data=f_inter, x='avg_inter', y='rf', kind='line',
                 height=8.27, aspect=11.7/8.27, linewidth=3)
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'])
plt.savefig('pics/6.png')
ax = sns.displot(data=df, x='nu', bins=ivs, kind='hist', linewidth=3,
                 height=8.27, aspect=11.7/8.27, stat='density')
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'], yticks=round((f_inter['rf']/h), 4))
plt.savefig('pics/7.png')
```

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

original =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/lab1/data/data2.csv')

var_row =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/lab1/data/data4.csv')

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

n = 100

h = 44

int_row =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/data/interval.csv')

int_row['cum_sum'] = np.round(np.cumsum(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

usl_mom_f = usl_mom.append(np.round(usl_mom.sum(), 3), ignore_index=True)

usl_mom_f.to_csv('data/usl_mom.csv', index=False)

moms = usl_mom_f.iloc[7, [3,4,5,6]]

checker = moms[3]+4*moms[2]+6*moms[1]+4*moms[0]+1

'True' if checker == usl_mom_f.loc[7, ['nu4+']][0] else 'False'
```

```
checker

M1 = moms[0]*h+475

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)

(M1, m2, m3, m4)

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)

int_mean

int_var

s

std_s

std_var

As = m3/(pow(s, 3))

Ex = (m4/(pow(s, 4))) - 3

As, Ex

raw_mode = 453+h*(8/26)

raw_median = 453+(((0.5*n)-40)/35)*h

raw_mode

raw_median

int_mean

sns.set_theme(palette='crest', font_scale=1.15)

sns.set_style('ticks', {"axes.facecolor": ".94"})

ax = sns.displot(data=original, x='nu', bins=np.array([321, 365, 409,
453, 497, 541, 585, 623]),
                 kind='hist', height=8.27, aspect=11.7/8.27,
stat='density')
```

```
plt.vlines(raw_mode, 0, int_row.loc[3, 'rf']/h, colors='b', linestyles='-
-', label='$мода$')

plt.vlines(raw_median, 0, int_row.loc[3, 'rf']/h, colors='r', lin-
estyles='--', label='$медиана$')

# plt.vlines(int_mean, 0, int_row.loc[3, 'rf']/h, colors='k', lin-
estyles='--', label='$x_в$')

ax.set_axis_labels('Середины интервалов', 'Частоты')

ax.set(xticks=int_row['avg_inter'])

plt.legend()

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

original =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/lab1/data2/data2.csv')

var_row =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/lab1/data2/data4.csv')

var_row.to_csv('data/var_row2.csv', index=False)

n = 100

h = 16

int_row =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/data/interval2.csv')

int_row['cum_sum'] = np.round(np.cumsum(int_row['rf']), 3)

int_row.to_csv('data/int_row2.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

usl_mom_f = usl_mom.append(np.round(usl_mom.sum(), 3), ignore_index=True)

usl_mom_f.to_csv('data/usl_mom2.csv', index=False)

usl_mom_f
```

```
moms = usl_mom_f.iloc[7, [3,4,5,6]]

checker = moms[3]+4*moms[2]+6*moms[1]+4*moms[0]+1

'True' if checker == usl_mom_f.loc[7, ['nu4+']][0] else 'False'

checker

M1 = moms[0]*h+140

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)

M1, m2, m3, m4

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)

int_mean

int_var

s

std_s

std_var

As = m3/(pow(s, 3))

Ex = (m4/(pow(s, 4))) - 3

As, Ex

original.mean()

raw_mode = 132.9+h*(1/25)

raw_median = 116.9+(((0.5*n)-20)/32)*h

raw_mode

raw_median

int_mean

sns.set_theme(palette='crest', font_scale=1.15)

sns.set_style('ticks', {"axes.facecolor": ".94"})
```

```
ax = sns.displot(data=original, x='E', bins=np.array([84.9, 100.9, 116.9, 132.9, 148.9, 164.9, 180.9, 195.7]),
                 kind='hist', height=8.27, aspect=11.7/8.27, stat='density')

plt.vlines(raw_mode, 0, int_row.loc[3, 'rf']/h, colors='b', linestyles='--', label='$мода$')

plt.vlines(raw_median, 0, int_row.loc[2, 'rf']/h, colors='r', linestyles='--', label='$медиана$')

# plt.vlines(int_mean, 0, int_row.loc[2, 'rf']/h, colors='k', linestyles='--', label='$x_в$')

ax.set_axis_labels('Середины интервалов', 'Частоты')

ax.set(xticks=int_row['avg_inter'])

plt.legend()

plt.savefig('pics/2.png')
```

ПРИЛОЖЕНИЕ В

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

```
import numpy as np

import pandas as pd

import scipy

from IPython.core.interactiveshell import InteractiveShell

InteractiveShell.ast_node_interactivity = "all"

int_row =
pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/data/interval.csv')

N = int_row['af'].sum()

h = 44

N

int_row

xv = (np.dot(int_row['avg_inter'], int_row['af'])/N).round(3)

dv = (np.dot((int_row['avg_inter']-xv)**2, int_row['af'])/N).round(3)

s = np.sqrt(dv*(N/(N-1))).round(3)

xv, dv, (dv*(N/(N-1))).round(3), s

gamma = 0.99

tg = 2.627

di_a = np.round((xv-tg*s/np.sqrt(N), xv+tg*s/np.sqrt(N)), 2)

xv

di_a

q = 0.198

di_s = np.round((s*(1-q), s*(1+q)), 3)

s

di_s

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] = 585, 623
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
df['F(zi)'] = np.array([-5000,-4671,-3485,-871,2190,4177,4858])/10000
df['F(zi+1)'] = np.array([-4671,-3485,-871,2190,4177,4858,5000])/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
df_nabl = pd.DataFrame()
df_nabl['ni'], df_nabl['ni*'] = df['ni'], df['ni*']
df_nabl['-'] = np.round(df_nabl['ni']-df_nabl['ni*'], 4)
df_nabl['-2'] = np.round(df_nabl['-']**2, 4)
df_nabl['-2/'] = np.round(df_nabl['-2']/df_nabl['ni*'], 4)
df_nabl.to_csv('data/data2.csv', index=False)
hi_nabl = df_nabl['-2/'].sum().round(4)
df_nabl
alpha = 0.05
k = len(df)-3
(k, alpha)
hi_crit = 9.5
(hi_nabl, hi_crit)
'True' if hi_nabl <= hi_crit else 'False'
```

ПРИЛОЖЕНИЕ Г

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

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('data.csv')
X = df['nu']
Y = df['E']

h1, h2 = 44, 16
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))))

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)

df_int.iloc[77, 2:6] = df_int.iloc[50, 2:6]
# 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()

N = 100
xv = 465.26
sx = 54.57
yv = 132.09
sy = 19.97

df_kor =
pd.DataFrame(columns=['yi','x1','x2','x3','x4','x5','x6','x7','Xi','yX'])
df_kor['yi'] =
[np.NaN,92.9,108.9,124.9,140.9,156.9,172.9,188.3,np.NaN,np.NaN]
df_kor['x1'] = [343,3,1,0,0,0,0,0,np.NaN,np.NaN]
```

```
df_kor['x2'] = [387,3,5,1,0,0,0,0,np.NaN,np.NaN]
df_kor['x3'] = [431,0,6,18,3,0,0,0,np.NaN,np.NaN]
df_kor['x4'] = [475,0,2,12,20,1,0,0,0,np.NaN]
df_kor['x5'] = [519,0,0,1,9,7,0,0,np.NaN,np.NaN]
df_kor['x6'] = [563,0,0,0,1,1,4,0,np.NaN,np.NaN]
df_kor['x7'] = [604,0,0,0,0,0,0,2,np.NaN,np.NaN]

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()
df_curr2
df_kor

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

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

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

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

gamma = 0.99
F = gamma/2
l = 2.58
z1 = (z-l*sz).round(4)
z2 = (z+l*sz).round(4)
(z1,z2)
```

```
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)

K = 7
Tn = ((r*np.sqrt(N-2))/np.sqrt(1-r**2)).round(3)
Tn
tk = 1.986

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

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

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('data/main_data.csv')
X = df['nu']
Y = df['E']
int_rowX = pd.read_csv('data/interval.csv')
int_rowY = pd.read_csv('data/interval2.csv')
kor = pd.read_csv('data/kor.csv')

sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style("ticks", {"axes.facecolor": ".95"})
ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27, as-
pect=11.7/8.27)
ax.set_axis_labels('X', 'Y')
plt.savefig('pics/1.png')

N = 100
xv, yv = 465.26, 132.09
sx, sy = 54.57, 19.97
r = 0.853

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

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

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

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

ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27,
                 aspect=11.7/8.27, s=40, label='Выборка')
plt.plot(regr_xy(df['E']), df['E'], label='x(y)', zorder=0, c='r')
plt.plot(df['nu'], regr_yx(df['nu']), label='y(x)', zorder=1, c='m')
ax.set_axis_labels('X', 'Y')
plt.legend()
plt.savefig('pics/2.png')

ost_var_xy
ost_var_yx

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] = 100

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)]

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)]

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)
```

```
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)

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)

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)

D_obsh_xy = D_vngr_xy + D_mezh_xy
D_obsh_xy.round(4)

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

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

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)

D_obsh_yx = D_vngr_yx + D_mezh_yx
D_obsh_yx.round(4)

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

```
r
kor

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

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]] = 0
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.round(4), b.round(4), c.round(4)

ax = sns.relplot(data=df, x='nu', y=parab_regr(df['nu']), kind='line',
linewidth=3,
```

```
                height=8.27, aspect=11.7/8.27, label='y(x) параб.', col-
or='m')
plt.scatter(df['nu'], df['E'], s=40, label='Выборка')
ax.set_axis_labels('X', 'Y')
plt.legend()
plt.savefig('pics/3.png')

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

df_step_x['ln(x)'] = np.log(df_step_x['x'])
df_step_x['ln2(x)'] = (np.log(df_step_x['x']))**2
df_step_x['ln(x)y'] = df_step_x['ln(x)']*df_step_x['y']

df_step_xf = df_step_x.append(df_step_x.sum(), ignore_index=True)
df_step_xf.iloc[-1,[0]] = np.NaN
df_step_xf.round(3).to_csv('data/logxy.csv', index=False)
df_step_xf

b2         =         ((df_step_xf.loc[7,'n']*df_step_xf.loc[7,'ln(x)y'])-
(df_step_xf.loc[7,'y']*df_step_xf.loc[7,'ln(x)']))/(((df_step_xf.loc[7,'n
']*df_step_xf.loc[7,'ln2(x)'])-(df_step_xf.loc[7,'ln(x)'])**2))
a2                        =                        (df_step_xf.loc[7,'y']-
(df_step_xf.loc[7,'ln(x)']*b2))/df_step_xf.loc[7,'n']
a2.round(4), b2.round(4)
log_regr = lambda x: a2+b2*np.log(x)

X_new_ln = np.hstack((np.ones((N,1)),np.expand_dims(np.log(X),1)))
beta_curr_hat                                                              =
np.matmul(np.matmul(np.linalg.inv(np.matmul(X_new_ln.T,X_new_ln)),X_new_l
n.T),Y)
plt.scatter(X,Y)
plt.xlabel("radius")
```

```
plt.ylabel("perimeter")
plt.plot(X,beta_curr_hat[0] + np.log(X) * beta_curr_hat[1],"-r")
plt.show()
a2 = beta_curr_hat[0]
b2 = beta_curr_hat[1]
log_regr = lambda x: a2+b2*np.log(x)
a2, b2

ax = sns.relplot(data=df, x='nu', y=log_regr(df['nu']), kind='line', lin-
ewidth=3,
                 height=8.27, aspect=11.7/8.27, label='y(x) лог.', col-
or='m')
plt.scatter(df['nu'], df['E'], s=40, label='Выборка')
ax.set_axis_labels('X', 'Y')
plt.legend()
plt.savefig('pics/4.png')

dfst = df.copy()
dfst['1'] = parab_regr(dfst['nu'])
dfst['2'] = log_regr(dfst['nu'])
dfstm = dfst.melt(id_vars='nu', value_vars=['1','2'])

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()
```

ПРИЛОЖЕНИЕ Е

ПРОГРАММА ДЛЯ МЕТОДА К-СРЕДНИХ

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from scipy.spatial import distance

df0 = pd.read_csv('data/main_data.csv')
X = df0['nu']
Y = df0['E']

X_norm = MinMaxScaler().fit_transform(df0)
df = pd.DataFrame(data=X_norm, columns=['nu','E'])
df.round(3).to_csv('data/df_norm.csv', index=False)

sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style("ticks", {"axes.facecolor": ".95"})
ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27,
aspect=11.7/8.27)
ax.set_axis_labels('X', 'Y')
plt.tight_layout()
plt.savefig('pics/1.png')

N = len(df)
sup_k = np.floor(np.sqrt(N/2))
sup_k

def sc_plots(data, means, Ncl, step):
    if Ncl > 6:
```

```
        ax = sns.relplot(data=data, x='nu', y='E', hue='кластер',
kind='scatter', alpha=0.8,
                         height=8.27, aspect=11.7/8.27)
        plt.scatter(means[:,0],means[:,1], c='m', s=60)
    else:
        ax = sns.relplot(data=data, x='nu', y='E', hue='кластер',
kind='scatter', palette='viridis', alpha=0.8,
                         height=8.27, aspect=11.7/8.27)
        plt.scatter(means[:,0],means[:,1], c='m', s=60)


    print(f'кластеры = {Ncl}; шаги = {step}')
    ax.set_axis_labels('X', 'Y')
    ax.fig.suptitle(f'{Ncl} кластера(ов)')
    plt.tight_layout()
    plt.savefig(f'pics/2_{Ncl}.png')
    plt.close()
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



def Fs(data):
    curr_data = data.copy()
    cts = curr_data.groupby('кластер').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['кластер']==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('кластер').var().values.sum()

    return F1,F2,F3

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['кластер'] = -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],'кластер'] = 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,'кластер'] = curr_clust # соотносим кластер
            if chng_ctr: # пересчет центра при новой точке
                centers[curr_clust][:2] =
data[data['кластер']==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['кластер']==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['кластер'].astype('int')
            sc_plots(data, centers, n_clusters, j+1)
            break
        F1,F2,F3 = cur_F1,cur_F2,cur_F3
        data['кластер'] = -1


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

data.groupby('кластер')['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


for i in range(2,8):
F, ctrs, sil = custKM(df, n_clusters=i, chng_ctr=1)
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)
```

ПРИЛОЖЕНИЕ Ж

ПРОГРАММА ДЛЯ МЕТОДА ПОИСКА СГУЩЕНИЙ

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from scipy.spatial import distance
import functools

df0 = pd.read_csv('data/main_data.csv')
X = df0['nu']
Y = df0['E']

X_norm = MinMaxScaler().fit_transform(df0)
df = pd.DataFrame(data=X_norm, columns=['nu','E'])
df.round(3).to_csv('data/df_norm.csv', index=False)

sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style("ticks", {"axes.facecolor": ".95"})
ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27, as-
pect=11.7/8.27)
ax.set_axis_labels('X', 'Y')
plt.tight_layout()
plt.savefig('pics/1.png')

def sc_plots(data, center, R, step, itera):
    ax = sns.relplot(data=data, x='nu', y='E', hue='кластер',
kind='scatter', palette='flare',
                     alpha=0.9, height=8.27, aspect=11.7/8.27,
legend=False)
    for j in [center]:
        plt.scatter(j[0],j[1], c='crimson', s=80)
```

```
#     df_CN = df_CN.append({'x':center.values[0].round(4),
#                           'y':center.values[1].round(4),
#                           'N':data[data["кластер"]!=-1]["nu"].count()})
    print(center.values[:2])
    print(data[data['кластер']!=-1]['nu'].count())

    circle = np.array([], dtype=np.float64)
    for i in data[data['кластер']!=-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*50000, alpha=0.1)

    ax.set_axis_labels('X', 'Y')
    ax.set(xlim=[-0.1,1.1], ylim=[-0.1,1.1])
    plt.tight_layout()
    plt.savefig(f'pics/{itera}_{step}.png')
    plt.show()

def Fs(data):
    curr_data = data.copy()
    cts = curr_data.groupby('кластер').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['кластер']==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('кластер').var().values.sum(where=~np.isnan(curr_data.g
roupby('кластер').var().values), initial=0)

    return F1,F2,F3

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['кластер'] = -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['кластер']==itera].mean()
        data['кластер'] = -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['кластер']==-1], data, np.array(center.values[:2])

coords = df.iloc[:,:2].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(5), rmax.round(5)

upd_df = df.copy()
```

```
it = 1
radius = 0.37
df['кластер'] = -1
ctrs = np.array([], dtype=np.float64)

while len(upd_df):
    upd_df, main, ctr = custFE(upd_df, radius, it, 0)
    ctrs = np.append(ctrs, [ctr])
    it += 1
    df.loc[main[main['кластер']!=-1].index, :] =
main.loc[main[main['кластер']!=-1].index, :]
df.to_csv('data/result.csv', index=False)

F1, F2, F3 = Fs(df)
F1.round(5), F2.round(5), F3.round(5)

ax = sns.relplot(data=df, x='nu', y='E', hue='кластер', kind='scatter',
pal-ette='viridis', alpha=0.9,
                 size='кластер', height=8.27, aspect=11.7/8.27)
ctrs = ctrs.reshape((-1,2))
for i in ctrs:
    plt.scatter(i[0], i[1], c='crimson', s=70)
ax.set_axis_labels('X', 'Y')
plt.tight_layout()
plt.savefig('pics/result.png')
plt.show()
```