**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

|  |  |  |
| --- | --- | --- |
| Студентка гр. 7381 |  | Алясова А.Н. |
| Преподаватель |  | Середа А.-В.И. |

Санкт-Петербург

2021

**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. 7381 |  | Кортев Ю.В. |
| Преподаватель |  | Середа А.-В.И. |

Санкт-Петербург

2021

**ЗАДАНИЕ**

**на курсовую работу**

|  |  |  |
| --- | --- | --- |
| Студентка Алясова А.Н. | | |
| Группа 7381 | | |
| Тема работы: Программная реализация и компьютерное исследование алгоритмов обработки экспериментальных данных | | |
| Исходные данные:  Из представленной генеральной совокупности формируется выборка заданного объема по одному из представленных в таблице признаков. 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: 06.04.2021 | | |
| Дата сдачи реферата: 13.04.2021 | | |
| Дата защиты реферата: 13.04.2021 | | |
| Студентка |  | Алясова А.Н. |
| Преподаватель |  | Середа А.-В.И. |

**ЗАДАНИЕ**

**на курсовую работу**

|  |  |  |
| --- | --- | --- |
| Студент Кортев Ю.В. | | |
| Группа 7381 | | |
| Тема работы: Программная реализация и компьютерное исследование алгоритмов обработки экспериментальных данных | | |
| Исходные данные:  Из представленной генеральной совокупности формируется выборка заданного объема по одному из представленных в таблице признаков. 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: 06.04.2021 | | |
| Дата сдачи реферата: 13.04.2021 | | |
| Дата защиты реферата: 13.04.2021 | | |
| Студент |  | Кортев Ю.В. |
| Преподаватель |  | Середа А.-В.И. |

**Аннотация**

В данной курсовой работе исследуется двухмерная выборка, состоящая из данных наблюдений относительно объемного веса при влажности и модуля упругости при сжатии вдоль волокон древесины резонансной ели. Исследование включает в себя выравнивание статистических рядов, нахождение точечных и интервальных статистических оценок, построение регрессионных кривых, проверку статистических гипотез о нормальном распределении выборки, о равенстве коэффициента корреляции нулю с помощью критерия Пирсона. Методы исследования включают в себя корреляционный анализ, регрессионный анализ методы кластеризации: k-средних и метод поиска сгущений. Построение регрессионных кривых осуществляется методом наименьших квадратов.

**Summary**

This term paper studies a two-dimensional sample consisting of observational data regarding the volume weight at 10% moisture content and the elastic modulus in compression along the fibers of resonant spruce wood. The study includes the alignment of statistical series, finding point and interval statistical estimates, the construction of regression curves, testing statistical hypotheses about the normal distribution of the sample, about the equality of the correlation coefficient to zero using Pearson's criterion. Research methods include correlation analysis, regression analysis clustering methods: k-means and the method of searching for clusters. The construction of regression curves is carried out by the method of least squares.

**содержание**

|  |  |  |
| --- | --- | --- |
|  | Введение | 8 |
| 1. | Выравнивание статистических рядов | 9 |
| 1.1. | Основные теоретические положения | 9 |
| 1.2.  1.3.  1.4.  1.5. | Формирование и первичная обработка выборки. Ранжированный и интервальный ряды  Нахождение точечных оценок параметров распределения  Нахождение интервальных оценок параметров распределения. Проверка статистической гипотезы о нормальном распределении  Выводы | 12  20  22  24 |
| 2. | Корреляционный и регрессионный анализ | 26 |
| 2.1. | Основные теоретические положения | 26 |
| 2.2.  2.3.  2.4. | Элементы корреляционного анализа. Проверка статистической гипотезы о равенстве коэффициента корреляции нулю  Элементы регрессионного анализа. Выборочные прямые. среднеквадратической регрессии. Корреляционные отношения.  Выводы | 31  43  48 |
| 3. | Кластерный анализ | 51 |
| 3.1. | Основные теоретические положения | 51 |
| 3.2.  3.3.  3.4. | Метод k-средних  Метод поиска сгущений  Выводы | 57  64  76 |
|  | Заключение | 77 |
|  | Список использованных источников | 78 |
|  | Приложение А. Программа для формирования и первичной обработки выборки, построения, ранжированного и интервального рядов  Приложение Б. Программа для нахождения точечных оценок параметров распределения  Приложение В. Программа для нахождения интервальных оценок параметров распределения и проверки статистической гипотезы о нормальном распределении  Приложение Г. Программа для нахождения элементов корреляционного анализа и проверки статистической гипотезы о равенстве коэффициента корреляции нулю  Приложение Д. Программа для нахождения элементов регрессионного анализа и построения выборочные прямых среднеквадратической регрессии, поиска корреляционного отношения  Приложение Е. Программа для метода k-cредних  Приложение Ж. Программа для метода поиска сгущений | 79  82  85  87  94  98  101 |

**введение**

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

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

Также в данной работе необходимо освоить и реализовать некоторые методы кластерного анализа, такие как, метод k-средних и метод поиска сгущений.

**1. выравнивание статистических рядов**

**1.1. Основные теоретические положения**

Статистический ряд – последовательность элементов выборки, расположенных в порядке их получения (наблюдения).

Ранжированный ряд – последовательность элементов выборки, расположенных в порядке возрастания их значений. Номер элемента ранжированного ряда в последовательности называется рангом.

Вариационный ряд – получается из ранжированного ряда в результате объединения одинаковых элементов. Элементы вариационного ряда называются вариантами.

Варианта – отдельные значения признака, по которому производится группировка.

Частота – число, показывающее, как часто встречается та или иная варианта. Сумма всех абсолютных частот равна общему числу наблюдений, относительных – единице.

Полигон частот – это один из способов графического представления плотности вероятности распределения выборки.

Гистограмма – это наглядное представление функции вероятности некоторой случайной величины, построенное по выборке. Гистограмма строится с помощью интервального ряда.

Эмпирической функцией распределения (функцией распределения выборки) называют функцию , определяющую для каждого значения относительную частоту события .

График представляет собой лестничный график, длина каждой ступеньки которого равна длине соответствующего интервала, а высота – отношению накопленной частоты до середины этого интервала к объёму выборки, т.е.:

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

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

Среднеквадратическим отклонением случайной величины Х (стандартом) называется квадратный корень из ее дисперсии:

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

где – центральный эмпирический момент третьего порядка,  *–* исправленнаявыборочная дисперсия.

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

Исправленная выборочная дисперсияопределяется по формуле:

где выборочная дисперсия.

Эксцессом называется численная характеристика случайной величины, которая определяется выражением:

Для нормального закона . Отсюда следует, что для нормального закона. Смысл термина «эксцесс» состоит в том, что он показывает, как быстро уменьшается плотность распределения вблизи её максимального значения.

Мода дискретной случайной величины – это наиболее вероятное значение этой случайной величины. Модой непрерывной случайной величины называется ее значение, при котором плотность вероятности максимальна.

где – начало модального интервала, – длина частичного интервала (шаг),  – частота предмодального интервала, – частота модального интервала, – частота послемодального интервала.

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

где  – начало медианного интервала,  – длина частичного интервала (шаг),  – объем совокупности,  – накопленная частота интервала, предшествующая медианному,  – частота медианного интервала.

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

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

,

где – статистическая оценка математического ожидания; – исправленная выборочная дисперсия; – объём выборки; – из таблицы.

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

Где – исправленная выборочная дисперсия; ­– из таблицы.

Критерий Пирсона, или критерий (Хи-квадрат), применяют для проверки гипотезы о соответствии эмпирического распределения предполагаемому теоретическому распределению .

Метод позволяет оценить статистическую значимость различий двух или нескольких относительных показателей (частот, долей).

Теоретические частоты вычисляются по формуле:

где

.

Следует привести теоретические частоты к функции Лапласа. Если , то примет следующий вид:

Для данной задачи . Преобразуя формулу получим:

,

где – функция ошибок.

Если - гипотеза принимается, иначе () – гипотезу отвергают.

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

Выборка состоит из данных наблюдений относительно объемного веса при влажности и модуля упругости при сжатии вдоль волокон древесины резонансной ели. Формирование репрезентативной выборки заданного объема из имеющейся генеральной совокупности экспериментальных данных представлены в табл. 1.2.1 и в табл. 1.2.2. Объём выборки: 117.

Таблица 1.2.1 - Генеральная совокупность экспериментальных данных

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 480 | 153,3 | 25 | 408 | 110,0 | 49 | 405 | 103,6 | 73 | 465 | 127,7 | 97 | 487 | 146,0 |
| 2 | 510 | 129,4 | 26 | 331 | 74,1 | 50 | 434 | 140,4 | 74 | 390 | 108,1 | 98 | 532 | 158,7 |
| 3 | 426 | 119,0 | 27 | 467 | 113,0 | 51 | 344 | 86,8 | 75 | 463 | 129,2 | 99 | 330 | 71,1 |
| 4 | 482 | 139,9 | 28 | 545 | 145,3 | 52 | 415 | 119,7 | 76 | 468 | 128,9 | 100 | 438 | 134,1 |
| 5 | 393 | 103,2 | 29 | 396 | 83,8 | 53 | 463 | 136,7 | 77 | 488 | 134,1 | 101 | 593 | 187,4 |
| 6 | 510 | 162,3 | 30 | 351 | 102,9 | 54 | 475 | 143,6 | 78 | 443 | 137,4 | 102 | 445 | 124,7 |
| 7 | 403 | 123,9 | 31 | 503 | 148,5 | 55 | 463 | 144,9 | 79 | 505 | 155,8 | 103 | 518 | 154,0 |
| 8 | 506 | 158,4 | 32 | 402 | 120,8 | 56 | 392 | 82,7 | 80 | 395 | 109,1 | 104 | 496 | 141,7 |
| 9 | 393 | 122,8 | 33 | 542 | 146,1 | 57 | 452 | 140,5 | 81 | 474 | 132,5 | 105 | 473 | 136,4 |
| 10 | 442 | 115,4 | 34 | 437 | 124,3 | 58 | 504 | 143,8 | 82 | 490 | 139,9 | 106 | 522 | 154,5 |
| 11 | 411 | 112,9 | 35 | 453 | 119,5 | 59 | 443 | 122,9 | 83 | 396 | 90,1 | 107 | 547 | 154,7 |
| 12 | 514 | 153,6 | 36 | 386 | 105,8 | 60 | 461 | 138,6 | 84 | 362 | 97,9 | 108 | 560 | 169,8 |
| 13 | 525 | 156,5 | 37 | 434 | 122,3 | 61 | 340 | 85,1 | 85 | 566 | 175,7 | 109 | 412 | 127,8 |
| 14 | 543 | 155,4 | 38 | 418 | 118,4 | 62 | 438 | 134,9 | 86 | 418 | 109,3 | 110 | 444 | 130,0 |
| 15 | 412 | 116,3 | 39 | 391 | 107,5 | 63 | 523 | 148,7 | 87 | 502 | 132,5 | 111 | 437 | 121,8 |
| 16 | 449 | 124,5 | 40 | 399 | 100,0 | 64 | 416 | 120,5 | 88 | 500 | 155,5 | 112 | 462 | 138,8 |
| 17 | 482 | 136,4 | 41 | 486 | 139,4 | 65 | 483 | 143,4 | 89 | 359 | 71,9 | 113 | 438 | 122,2 |
| 18 | 569 | 157,4 | 42 | 421 | 124,2 | 66 | 440 | 128,5 | 90 | 443 | 135,7 | 114 | 406 | 110,1 |
| 19 | 484 | 147,5 | 43 | 496 | 143,1 | 67 | 423 | 131,1 | 91 | 421 | 118,0 | 115 | 413 | 106,7 |
| 20 | 472 | 134,2 | 44 | 463 | 121,2 | 68 | 386 | 95,5 | 92 | 433 | 128,2 | 116 | 458 | 121,7 |
| 21 | 453 | 124,2 | 45 | 508 | 159,0 | 69 | 321 | 86,1 | 93 | 514 | 174,6 | 117 | 408 | 117,0 |
| 22 | 422 | 117,9 | 46 | 419 | 105,3 | 70 | 433 | 131,5 | 94 | 320 | 72,6 |  |  |  |
| 23 | 320 | 64,5 | 47 | 434 | 108,7 | 71 | 351 | 89,0 | 95 | 406 | 113,8 |  |  |  |
| 24 | 547 | 164,4 | 48 | 440 | 126,7 | 72 | 481 | 148,3 | 96 | 465 | 140,9 |  |  |  |

Таблица 1.2.2 - Репрезентативная выборка

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 480 | 18 | 569 | 35 | 453 | 52 | 415 | 69 | 321 | 86 | 418 | 103 | 518 |
| 2 | 510 | 19 | 484 | 36 | 386 | 53 | 463 | 70 | 433 | 87 | 502 | 104 | 496 |
| 3 | 426 | 20 | 472 | 37 | 434 | 54 | 475 | 71 | 351 | 88 | 500 | 105 | 473 |
| 4 | 482 | 21 | 453 | 38 | 418 | 55 | 463 | 72 | 481 | 89 | 359 | 106 | 522 |
| 5 | 393 | 22 | 422 | 39 | 391 | 56 | 392 | 73 | 465 | 90 | 443 | 107 | 547 |
| 6 | 510 | 23 | 320 | 40 | 399 | 57 | 452 | 74 | 390 | 91 | 421 | 108 | 560 |

Продолжение таблицы 1.2.2

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 7 | 403 | 24 | 547 | 41 | 486 | 58 | 504 | 75 | 463 | 92 | 433 | 109 | 412 |
| 8 | 506 | 25 | 408 | 42 | 421 | 59 | 443 | 76 | 468 | 93 | 514 | 110 | 444 |
| 9 | 393 | 26 | 331 | 43 | 496 | 60 | 461 | 77 | 488 | 94 | 320 | 111 | 437 |
| 10 | 442 | 27 | 467 | 44 | 563 | 61 | 340 | 78 | 443 | 95 | 406 | 112 | 462 |
| 11 | 411 | 28 | 545 | 45 | 508 | 62 | 438 | 79 | 505 | 96 | 465 | 113 | 438 |
| 12 | 514 | 29 | 396 | 46 | 419 | 63 | 523 | 80 | 395 | 97 | 487 | 114 | 406 |
| 13 | 525 | 30 | 351 | 47 | 434 | 64 | 416 | 81 | 474 | 98 | 532 | 115 | 413 |
| 14 | 543 | 31 | 503 | 48 | 440 | 65 | 483 | 82 | 490 | 99 | 330 | 116 | 458 |
| 15 | 412 | 32 | 402 | 49 | 405 | 66 | 440 | 83 | 396 | 100 | 438 | 117 | 408 |
| 16 | 449 | 33 | 542 | 50 | 434 | 67 | 423 | 84 | 362 | 101 | 593 |  |  |
| 17 | 482 | 34 | 437 | 51 | 344 | 68 | 386 | 85 | 566 | 102 | 445 |  |  |

Преобразование полученной выборки в ранжированный ряд представлено в табл. 1.2.3.

Таблица 1.2.3 – Ранжированный ряд

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 320 | 18 | 393 | 35 | 415 | 52 | 438 | 69 | 463 | 86 | 484 | 103 | 514 |
| 2 | 320 | 19 | 395 | 36 | 416 | 53 | 438 | 70 | 463 | 87 | 486 | 104 | 518 |
| 3 | 321 | 20 | 396 | 37 | 418 | 54 | 440 | 71 | 463 | 88 | 487 | 105 | 522 |
| 4 | 330 | 21 | 396 | 38 | 418 | 55 | 440 | 72 | 463 | 89 | 488 | 106 | 523 |
| 5 | 331 | 22 | 399 | 39 | 419 | 56 | 442 | 73 | 465 | 90 | 490 | 107 | 525 |
| 6 | 340 | 23 | 402 | 40 | 421 | 57 | 443 | 74 | 465 | 91 | 496 | 108 | 532 |
| 7 | 344 | 24 | 403 | 41 | 421 | 58 | 443 | 75 | 467 | 92 | 496 | 109 | 542 |
| 8 | 351 | 25 | 405 | 42 | 422 | 59 | 443 | 76 | 468 | 93 | 500 | 110 | 543 |
| 9 | 351 | 26 | 406 | 43 | 423 | 60 | 444 | 77 | 472 | 94 | 502 | 111 | 545 |
| 10 | 359 | 27 | 406 | 44 | 426 | 61 | 445 | 78 | 473 | 95 | 503 | 112 | 547 |
| 11 | 362 | 28 | 408 | 45 | 433 | 62 | 449 | 79 | 474 | 96 | 504 | 113 | 547 |
| 12 | 386 | 29 | 408 | 46 | 433 | 63 | 452 | 80 | 475 | 97 | 505 | 114 | 560 |
| 13 | 386 | 30 | 411 | 47 | 434 | 64 | 453 | 81 | 480 | 98 | 506 | 115 | 566 |
| 14 | 390 | 31 | 412 | 48 | 434 | 65 | 453 | 82 | 481 | 99 | 508 | 116 | 569 |
| 15 | 391 | 32 | 412 | 49 | 434 | 66 | 458 | 83 | 482 | 100 | 510 | 117 | 593 |
| 16 | 392 | 33 | 413 | 50 | 437 | 67 | 461 | 84 | 482 | 101 | 510 |  |  |
| 17 | 393 | 34 | 415 | 51 | 437 | 68 | 462 | 85 | 483 | 102 | 514 |  |  |

Из табл. 1.2.3 можно увидеть, что наименьшее значение в выборке , а наибольшее значение .

Преобразование полученной выборки в вариационный ряд с абсолютными частотами представлено в табл. 1.2.4.

Таблица 1.2.4 - Вариационный ряд с абсолютными частотами

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 320 | 2 | 19 | 403 | 1 | 37 | 438 | 3 | 55 | 474 | 1 | 73 | 508 | 1 |
| 2 | 321 | 1 | 20 | 405 | 1 | 38 | 440 | 2 | 56 | 475 | 1 | 74 | 510 | 2 |
| 3 | 330 | 1 | 21 | 406 | 2 | 39 | 442 | 1 | 57 | 480 | 1 | 75 | 514 | 2 |
| 4 | 331 | 1 | 22 | 408 | 2 | 40 | 443 | 3 | 58 | 481 | 1 | 76 | 518 | 1 |
| 5 | 340 | 1 | 23 | 411 | 1 | 41 | 444 | 1 | 59 | 482 | 2 | 77 | 522 | 1 |
| 6 | 344 | 1 | 24 | 412 | 2 | 42 | 445 | 1 | 60 | 483 | 1 | 78 | 523 | 1 |
| 7 | 351 | 2 | 25 | 413 | 1 | 43 | 449 | 1 | 61 | 484 | 1 | 79 | 525 | 1 |
| 8 | 359 | 1 | 26 | 415 | 1 | 44 | 452 | 1 | 62 | 486 | 1 | 80 | 532 | 1 |
| 9 | 362 | 1 | 27 | 416 | 1 | 45 | 453 | 2 | 63 | 487. | 1 | 81 | 542 | 1 |
| 10 | 386 | 2 | 28 | 418 | 2 | 46 | 458 | 1 | 64 | 488 | 1 | 82 | 543 | 1 |
| 11 | 390 | 1 | 29 | 419 | 1 | 47 | 461 | 1 | 65 | 490 | 1 | 83 | 545 | 1 |
| 12 | 391 | 1 | 30 | 421 | 2 | 48 | 462 | 1 | 66 | 496 | 2 | 84 | 547 | 2 |
| 13 | 392 | 1 | 31 | 422 | 1 | 49 | 463 | 4 | 67 | 500 | 1 | 85 | 560 | 1 |
| 14 | 393 | 2 | 32 | 423 | 1 | 50 | 465 | 2 | 68 | 502 | 1 | 86 | 566 | 1 |
| 15 | 395 | 1 | 33 | 426 | 1 | 51 | 467 | 1 | 69 | 503 | 1 | 87 | 569 | 1 |
| 16 | 396 | 2 | 34 | 433 | 2 | 52 | 468 | 1 | 70 | 504 | 1 | 88 | 593 | 1 |
| 17 | 399 | 1 | 35 | 434 | 3 | 53 | 472 | 1 | 71 | 505 | 1 |  |  |  |
| 18 | 402 | 1 | 36 | 437 | 2 | 54 | 473 | 1 | 72 | 506 | 1 |  |  |  |

Преобразование полученной выборки в вариационный ряд с относительными частотами представлено в табл. 1.2.5.

Таблица 1.2.5 - Вариационный ряд с относительными частотами

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 320 | 0.01709 | 19 | 403 | 0.00855 | 37 | 438 | 0.02564 | 55 | 474 | 0.00855 | 73 | 508 | 0.00855 |
| 2 | 321 | 0.00855 | 20 | 405 | 0.00855 | 38 | 440 | 0.01709 | 56 | 475 | 0.00855 | 74 | 510 | 0.01709 |
| 3 | 330 | 0.00855 | 21 | 406 | 0.01709 | 39 | 442 | 0.00855 | 57 | 480 | 0.00855 | 75 | 514 | 0.01709 |
| 4 | 331 | 0.00855 | 22 | 408 | 0.01709 | 40 | 443 | 0.02564 | 58 | 481 | 0.00855 | 76 | 518 | 0.00855 |

Продолжение таблицы 1.2.5

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 340 | 0.00855 | 23 | 411 | 0.00855 | 41 | 444 | 0.00855 | 59 | 482 | 0.01709 | 77 | 522 | 0.00855 |
| 6 | 344 | 0.00855 | 24 | 412 | 0.01709 | 42 | 445 | 0.00855 | 60 | 483 | 0.00855 | 78 | 523 | 0.00855 |
| 7 | 351 | 0.01709 | 25 | 413 | 0.00855 | 43 | 449 | 0.00855 | 61 | 484 | 0.00855 | 79 | 525 | 0.00855 |
| 8 | 359 | 0.00855 | 26 | 415 | 0.00855 | 44 | 452 | 0.00855 | 62 | 486 | 0.00855 | 80 | 532 | 0.00855 |
| 9 | 362 | 0.00855 | 27 | 416 | 0.00855 | 45 | 453 | 0.01709 | 63 | 487. | 0.00855 | 81 | 542 | 0.00855 |
| 10 | 386 | 0.01709 | 28 | 418 | 0.01709 | 46 | 458 | 0.00855 | 64 | 488 | 0.00855 | 82 | 543 | 0.00855 |
| 11 | 390 | 0.00855 | 29 | 419 | 0.00855 | 47 | 461 | 0.00855 | 65 | 490 | 0.00855 | 83 | 545 | 0.00855 |
| 12 | 391 | 0.00855 | 30 | 421 | 0.01709 | 48 | 462 | 0.00855 | 66 | 496 | 0.01709 | 84 | 547 | 0.01709 |
| 13 | 392 | 0.00855 | 31 | 422 | 0.00855 | 49 | 463 | 0.03419 | 67 | 500 | 0.00855 | 85 | 560 | 0.00855 |
| 14 | 393 | 0.01709 | 32 | 423 | 0.00855 | 50 | 465 | 0.01709 | 68 | 502 | 0.00855 | 86 | 566 | 0.00855 |
| 15 | 395 | 0.00855 | 33 | 426 | 0.00855 | 51 | 467 | 0.00855 | 69 | 503 | 0.00855 | 87 | 569 | 0.00855 |
| 16 | 396 | 0.01709 | 34 | 433 | 0.01709 | 52 | 468 | 0.00855 | 70 | 504 | 0.00855 | 88 | 593 | 0.00855 |
| 17 | 399 | 0.00855 | 35 | 434 | 0.02564 | 53 | 472 | 0.00855 | 71 | 505 | 0.00855 |  |  |  |
| 18 | 402 | 0.00855 | 36 | 437 | 0.01709 | 54 | 473 | 0.00855 | 72 | 506 | 0.00855 |  |  |  |

Для определения количества интервалов используем формулу Стерджесса: где – объем выборки.

Используя в качестве , получаем, что .

Чтобы определить шаг, с которым формировать интервалы, использована формула:

Соответственно, для , и получаем, что .

Полученный интервальный ряд приведен в табл. 1.2.6.

Таблица 1.2.6 – Интервальный ряд

|  |  |  |
| --- | --- | --- |
| **Интервал** | **Абсолютная частота** | **Относительная частота** |
| [320; 354) | 9 | 0,07692 |
| [354; 388) | 4 | 0,03419 |
| [388; 422) | 27 | 0,23077 |
| [422; 456) | 25 | 0,21368 |
| [456; 490) | 24 | 0,20513 |

Продолжение таблицы 1.2.6

|  |  |  |
| --- | --- | --- |
| [490; 524) | 17 | 0,14530 |
| [524; 558) | 7 | 0,05983 |
| [558; 592) | 3 | 0,02564 |
| [592; 593] | 1 | 0,00855 |

В сумме абсолютные частоты дают 117, что соответствует объему выборки, а относительные частоты суммируются к единице.

Полигон, построенный применительно к интервальному ряду для абсолютных частот представлен на рис. 1.2.1.

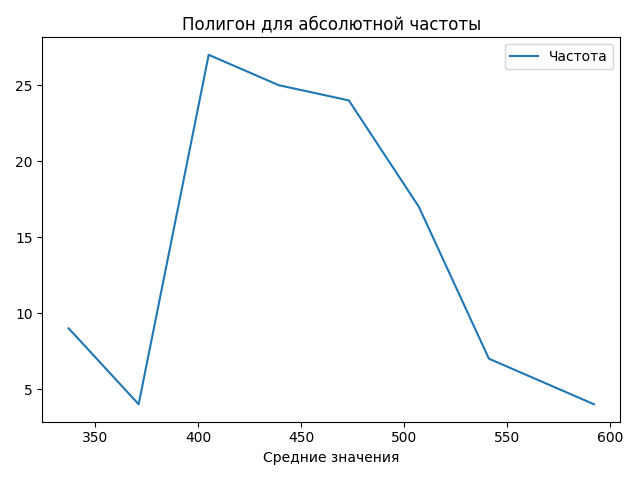


Рисунок 1.2.1 – Полигон для абсолютной частоты

Полигон, построенный применительно к интервальному ряду для относительных частот представлен на рис. 1.2.2.

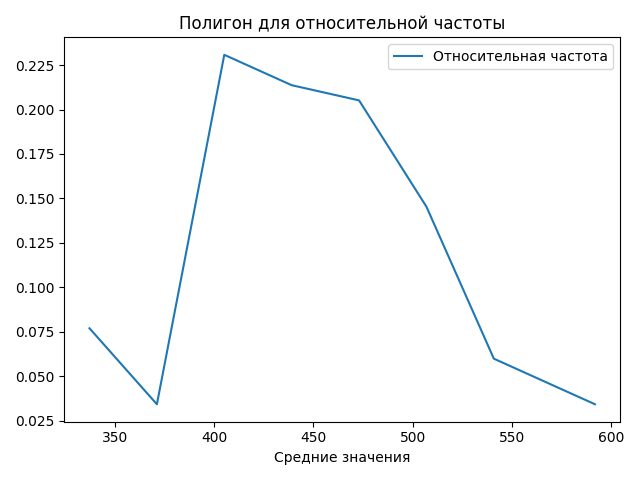


Рисунок 1.2.2 – Полигон для относительной частоты

Гистограмма, построенная применительно к интервальному ряду для абсолютных частот представлен на рис. 1.2.3.

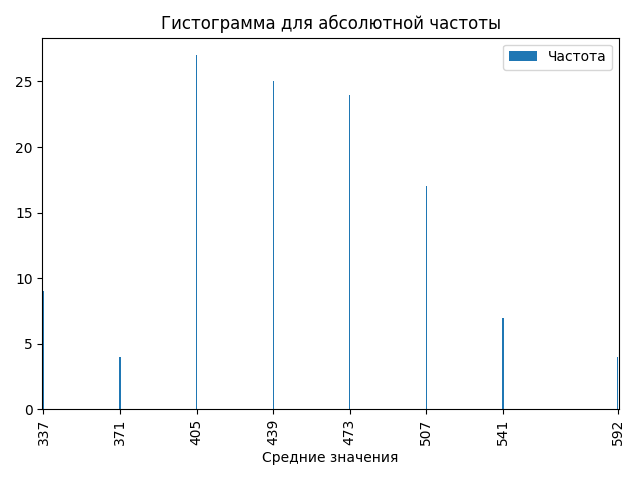


Рисунок 1.2.3 – Гистограмма для абсолютной частоты

Гистограмма, построенная применительно к интервальному ряду для относительных частот представлен на рис. 1.2.4.

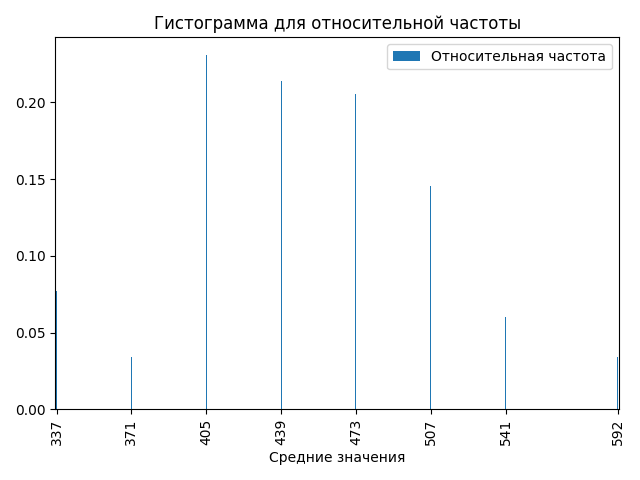


Рисунок 1.2.4 – Гистограмма для относительной частоты

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

Функция распределения:

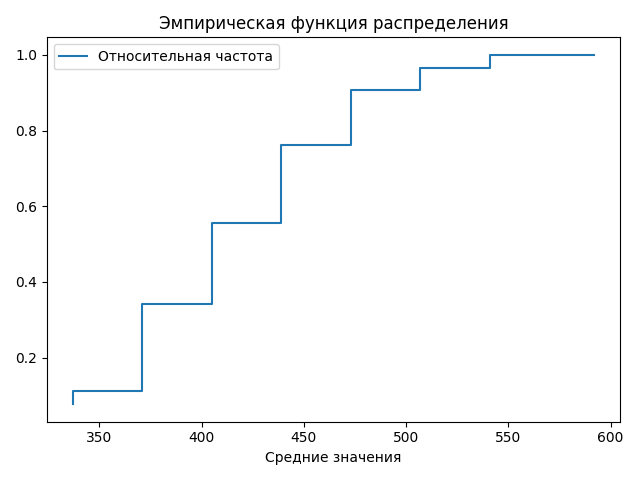


Рисунок 1.2.5 – График эмпирической функции распределения

**1.3. Нахождение точечных оценок параметров распределения.**

Найдем условные моменты по формуле: где где длина интервала, – ложный ноль.

Результаты вычислений представлены в табл. 1.3.1.

Таблица 1.3.1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| 337 | 9 | -4 | -36 | 144 | -576 | 2304 | 729 |
| 371 | 4 | -3 | -12 | 36 | -108 | 324 | 64 |
| 405 | 27 | -2 | -54 | 108 | -216 | 432 | 27 |
| 439 | 25 | -1 | -25 | 25 | -25 | 25 | 0 |
| 473 | 24 | 0 | 0 | 0 | 0 | 0 | 24 |
| 507 | 17 | 1 | 17 | 17 | 17 | 17 | 272 |
| 541 | 7 | 2 | 14 | 28 | 56 | 112 | 567 |
| 592 | 4 | 3,5 | 14 | 49 | 171,5 | 600,25 | 1640,25 |
| 3665 | 117 | -3,5 | -82 | 407 | -680,5 | 3814,25 | 3323,25 |
| Условные моменты: | | | -0,7009 | 3,4786 | -5,8162 | 32,6004 |  |

Проверим правильность вычислений:

Вычислим статистические оценки математического ожидания:

Вычислим статистические оценки дисперсии:

Отсюда следует, что среднеквадратическое отклонение:

Найдем исправленную выборочную дисперсию:

Для вычисления ассиметрии и эксцесса найдем центральные эмпирические моменты третьего и четвертого порядка:

Вычислим ассиметрию:

Вычислим эксцесс:

Далее найдем моду вариационного ряда по формуле:

Далее найдем медиану вариационного ряда по формуле:

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

Определим доверительный интервал для математического ожидания по формуле:

, где

– статистическая оценка математического ожидания;

– исправленная выборочная дисперсия;

– объём выборки;

– из таблицы (при уровне значимости и ).

Доверительный интервал для математического ожидания (; ).

Определим доверительный интервал для среднеквадратического отклонения по формуле:

– исправленная выборочная дисперсия;

– из таблицы (при уровне значимости и ).

Из полученных результатов можно сделать вывод, что полученный интервал покрывает величину с вероятностью 95%.

Проверим гипотезу о нормальном распределении исследуемой случайной величины с помощью критерия Пирсона Для этого вычислим теоретические вероятности и частоты попадания в каждый интервал. Результаты представлены в табл. 1.4.1 и в табл. 1.4.2.

Таблица 1.4.1

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 320 | -129,1709 | -2,1886 | -0,4856 |
| 354 | -95,1709 | -1,6125 | -0,4466 |
| 388 | -61,1709 | -1,0365 | -0,3500 |
| 422 | -27,1709 | -0,4604 | -0,1774 |
| 456 | 6,8291 | 0,1157 | 0,0461 |
| 490 | 40,8291 | 0,6918 | 0,2555 |
| 524 | 74, 8291 | 1,2679 | 0,3976 |
| 558 | 108, 8291 | 1,8440 | 0,4674 |
| 626 | 176, 8291 | 2,9961 | 0,4986 |

Таблица 1.4.2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| -1,9006 | 0,0655 | 0,0377 | 4,4177 | 0,0391 | 4,5758 |
| -1,3245 | 0,1659 | 0,0956 | 11,1850 | 0,0966 | 11,2989 |
| -0,7484 | 0,3015 | 0,1737 | 20,3213 | 0,1726 | 20,1977 |
| -0,1723 | 0,3931 | 0,2264 | 26,4932 | 0,2234 | 26,1419 |
| 0,4038 | 0,3677 | 0,2118 | 24,7847 | 0,2094 | 24,5008 |
| 0,9798 | 0,2468 | 0,1422 | 16,6381 | 0,1421 | 16,6272 |
| 1,5559 | 0,1189 | 0,0685 | 8,0148 | 0,0698 | 8,1697 |
| 2,4201 | 0,0213 | 0,0123 | 1,4382 | 0,0312 | 3,6536 |

Вычислим с использованием полученных частот по формуле: . Результаты представлены в табл. 1.4.3 и табл. 1.4.4.

Сравним полученные значения с табличным значением .

Из полученных результатов можно сделать вывод, что данные отвергаются гипотезой и не имеют нормального распределения, так как в обоих способах.

Таблица 1.4.3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 9 | 4,4177 | 4,5823 | 20,9976 | 4,7531 |
| 4 | 11,1850 | -7,1851 | 51,6254 | 4,6156 |
| 27 | 20,3213 | 6,6787 | 44,6047 | 2,1950 |
| 25 | 26,4932 | -1,4932 | 2,2295 | 0,0842 |
| 24 | 24,7847 | -0,7847 | 0,6158 | 0,0248 |
| 17 | 16,6381 | 0,3619 | 0,1310 | 0,0079 |
| 7 | 8,0148 | -1,0148 | 1,0298 | 0,1285 |
| 4 | 1,4382 | 2,5618 | 6,5627 | 4,5630 |
| **Сумма** | | | | 16,3720 |

Таблица 1.4.4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 9 | 4,5758 | 4,4241 | 19,5732 | 4,2775 |
| 4 | 11,2989 | -7,2989 | 53,2743 | 4,7150 |
| 27 | 20,1977 | 6,8023 | 46,2716 | 2,2909 |
| 25 | 26,1419 | -1,1419 | 1,3040 | 0,0499 |
| 24 | 24,5008 | -0,5008 | 0,2508 | 0,0102 |
| 17 | 16,6272 | 0,3728 | 0,1390 | 0,0084 |
| 7 | 8,1697 | -1,1697 | 1,3681 | 0,1675 |
| 4 | 3,6536 | 0,3464 | 0,1200 | 0,0329 |
| **Сумма** | | | | 11,5522 |

**1.5. Выводы.**

Была сформирована выборка данных и осуществлена её подготовка к статическому анализу. Выборка приведена к ранжированному, вариационному и интервальному видам. Используя полученный интервальный ряд построен полигон, гистограмма и эмпирическая функция распределения для абсолютных и относительных частот. Из полученного ранжированного ряда сразу видны минимальное и максимальное значение выборки. В данном случае были получены значения , . По полученному вариационному ряду виден наиболее частотный элемент выборки с частотой . По сформированному интервальному ряду можно увидеть, что большинство значений выборки сконцентрированы в интервале [388; 422). Более наглядно это представляют построенные гистограммы и полигоны частот. При этом их форма не зависит от того, какие частоты используются – абсолютные или относительные.

Также были получены практические навыки нахождения точечных статистических оценок параметров распределения. При вычислении условных моментов была сделана проверка, которая показала, что данные моменты были посчитаны верно. Так как полученное значение эксцесса , то можно сделать вывод, что плотность закона распределения случайной величины уменьшается медленно вблизи её моды. Из полученного значения коэффициента симметрии можно сделать вывод, что мода немного смещена влево относительно середины распределения, так как , но при этом находится достаточно близко к центру, так как значение близко к 0.

Были получены границы доверительных интервалов для математического ожидания и среднеквадратического отклонения случайной величины. Из полученных результатов можно сделать вывод, что интервал (; ) покрывает математическое ожидание и интервал покрывает величину с вероятностью 95%.

Также была проверена гипотеза о нормальном распределении исследуемой случайной величины с помощью критерия Пирсона . Из полученного результата можно сделать вывод, что гипотеза отвергается, т.к. , соответственно, исследуемая случайная величина не принадлежит нормальному закону распределения.

**2. корреляционный и регрессионный анализ**

**2.1. Основные теоретические положения.**

**Корреляционный анализ.**

Рассмотрим систему двух случайных величин . Эти случайные величины могут быть независимыми:

В противном случае между ними может быть:

1. Функциональная зависимость:
2. Статистическая зависимость:

Одним из видов (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 (регрессии):

Корреляционный момент: .

Коэффициент корреляции: .

Для коэффициента корреляции справедливо соотношение:

Случайные величины называют коррелированными, если их корреляционный момент или их коэффициент корреляции отличен от нуля. В противном случае эти величины некоррелированные.

Если случайные величины и коррелированы, то они зависимы. Обратное предположение в общем случае неверно:

1. и коррелированы и зависимы.
2. и некоррелированные и независимы

Коэффициент корреляции служит мерой тесноты линейной зависимости между случайными величинами и . При эта зависимость становится функциональной.

Значение – статистической оценки – коэффициента корреляции можно вычислить по формуле:

При в случае нормального распределния системы случайных величин для оценки значения можно использовать соотношение (не является доверительным интервалом):

Распределение при определённых условиях можно удовлетворительно аппроксимировать нормальным законом. Однако при увеличении интенсивности связи распределение становится всё более ассиметричным.

С помощью преобразования Фишера перейдём к случайной величине :

Распределение при неограниченном возрастании объёма выборки асимптотически нормальное со значением СКВО:

В результате доверительный интервал для генеральной совокупности с доверительной вероятностью определяется по следующей схеме:

1. По формуле (1) вычисляется выборочное значение .
2. По формуле (2) вычисляется значение .
3. Интервал для генерального значения представляется в виде:

где значение должно удовлетворять условию:

1. Для пересчёта интервала в доверительных интервал для коэффициента корреляции с тем же значением необходимо воспользоваться обратным преобразованием Фишера:

Пусть имеется выборка объёма значений двумерной нормально распределённой случайной величины и вычислено значение выборочного коэффициента корреляции . Поскольку является случайной величиной, то это ещё не значит, что – коэффициент корреляции для генеральной совокупности тоже отличен от нуля.

Возникает необходимость проверить гипотезу . Альтернативой будет гипотеза . Если основная гипотеза отвергается, то это означает, что выборочный коэффициент корреляции значимо отличается от нуля (значим). В противном случае – незначим.

В качестве критерия проверки статистической гипотезы о значимости выборочного коэффициента корреляции можно принять случайную величину:

При справедливости нулевой гипотезы случайная величина распределена по закону Стьюдента с степенями свободы. Критическая область для данного критерия двусторонняя.

Проверка гипотезы осуществляется по стандартной схеме:

1. По формуле (3) вычисляется значение .
2. По заданному уровню значимости и значению из таблицы определяется значение .
3. Если – нет оснований отвергать гипотезу .

Если – основная гипотеза с выборочными данными должна быть отвергнута.

Регрессионный анализ.

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

Регрессионный анализ – это статистический метод исследования влияния одной или нескольких независимых переменных на зависимую переменную .

Линейные функции выборочной среднеквадратической регрессии:

где - статистические оценки математических ожиданий выборок , соответственно, – статистическая оценка коэффициента корреляции, и – статистические оценки среднеквадратических отклонений для выборок , .

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

Чтобы рассчитать выборочное корреляционное отношение нужно рассчитать внутригрупповую и межгрупповую дисперсии. Оценку общей дисперсии можно представить, как сумму:

Чтобы рассчитать выборочное корреляционное отношение к нужно рассчитать внутригрупповую и межгрупповую дисперсии.

Внутригрупповая дисперсия вычисляется по формуле:

где – объём выборки, – количество интервалов интервального ряда , – абсолютная частота для -ого интервала интервального ряда , – групповая дисперсия элементов выборки на -ом интервале интервального ряда .

Межгрупповая дисперсия вычисляется по формуле:

где – объём выборки, – количество интервалов интервального ряда , – абсолютная частота для i-ого интервала интервального ряда , 𝑖 – групповое математическое ожидание элементов выборки на -ом интервале интервального ряда , – статистическая оценка математического ожидания .

Выборочное корреляционное отношение к определяется в соответствии с выражением:

Где , – выборочные значения СКВО и соответственно. Аналогично определяется выборочное корреляционное отношение к .

Для расчёта выборочного корреляционного отношения к необходимо рассчитать те же величины по следующим формулам (меняем местами и ):

Выборочное уравнение регрессии на : .

Значения коэффициентов определим с помощью МНК, что приводит к необходимости решать систему линейных уравнений 3го порядка:

Решив данную систему, найдём коэффициенты квадратичной функции выборочной среднеквадратической регрессии.

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

Результаты формирования второй выборки заданного объема из имеющейся генеральной совокупности экспериментальных данных представлены в табл. 2.2.1.

Таблица 2.2.1 - Репрезентативная выборка

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 153,3 | 18 | 157,4 | 35 | 119,5 | 52 | 119,7 | 69 | 86,1 | 86 | 109,3 | 103 | 154,0 |
| 2 | 129,4 | 19 | 147,5 | 36 | 105,8 | 53 | 136,7 | 70 | 131,5 | 87 | 132,5 | 104 | 141,7 |
| 3 | 119,0 | 20 | 134,2 | 37 | 122,3 | 54 | 143,6 | 71 | 89,0 | 88 | 155,5 | 105 | 136,4 |
| 4 | 139,9 | 21 | 124,2 | 38 | 118,4 | 55 | 144,9 | 72 | 148,3 | 89 | 71,9 | 106 | 154,5 |
| 5 | 103,2 | 22 | 117,9 | 39 | 107,5 | 56 | 82,7 | 73 | 127,7 | 90 | 135,7 | 107 | 154,7 |
| 6 | 162,3 | 23 | 64,5 | 40 | 100,0 | 57 | 140,5 | 74 | 108,1 | 91 | 118,0 | 108 | 169,8 |
| 7 | 123,9 | 24 | 164,4 | 41 | 139,4 | 58 | 143,8 | 75 | 129,2 | 92 | 128,2 | 109 | 127,8 |
| 8 | 158,4 | 25 | 110,0 | 42 | 124,2 | 59 | 122,9 | 76 | 128,9 | 93 | 174,6 | 110 | 130,0 |
| 9 | 122,8 | 26 | 74,1 | 43 | 143,1 | 60 | 138,6 | 77 | 134,1 | 94 | 72,6 | 111 | 121,8 |
| 10 | 115,4 | 27 | 113,0 | 44 | 121,2 | 61 | 85,1 | 78 | 137,4 | 95 | 113,8 | 112 | 138,8 |
| 11 | 112,9 | 28 | 145,3 | 45 | 159,0 | 62 | 134,9 | 79 | 155,8 | 96 | 140,9 | 113 | 122,2 |
| 12 | 153,6 | 29 | 83,8 | 46 | 105,3 | 63 | 148,7 | 80 | 109,1 | 97 | 146,0 | 114 | 110,1 |
| 13 | 156,5 | 30 | 102,9 | 47 | 108,7 | 64 | 120,5 | 81 | 132,5 | 98 | 158,7 | 115 | 106,7 |
| 14 | 155,4 | 31 | 148,5 | 48 | 126,7 | 65 | 143,4 | 82 | 139,9 | 99 | 71,1 | 116 | 121,7 |
| 15 | 116,3 | 32 | 120,8 | 49 | 119,5 | 66 | 128,5 | 83 | 90,1 | 100 | 134,1 | 117 | 117,0 |
| 16 | 124,5 | 33 | 146,1 | 50 | 105,8 | 67 | 131,1 | 84 | 97,9 | 101 | 187,4 |  |  |
| 17 | 136,4 | 34 | 124,3 | 51 | 122,3 | 68 | 95,5 | 85 | 175,7 | 102 | 124,7 |  |  |

Преобразование полученной выборки в ранжированный ряд представлено в табл. 2.2.2.

Таблица 2.2.2 – Ранжированный ряд

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 64,5 | 18 | 103,6 | 35 | 117,9 | 52 | 124,2 | 69 | 134,1 | 86 | 141,7 | 103 | 154,7 |
| 2 | 71,1 | 19 | 105,3 | 36 | 118,0 | 53 | 124,3 | 70 | 134,1 | 87 | 143,1 | 104 | 155,4 |
| 3 | 71,9 | 20 | 105,8 | 37 | 118,4 | 54 | 124,5 | 71 | 134,2 | 88 | 143,4 | 105 | 155,5 |
| 4 | 72,6 | 21 | 106,7 | 38 | 119,0 | 55 | 124,7 | 72 | 134,9 | 89 | 143,6 | 106 | 155,8 |
| 5 | 74,1 | 22 | 107,5 | 39 | 119,5 | 56 | 126,7 | 73 | 135,7 | 90 | 143,8 | 107 | 156,5 |
| 6 | 82,7 | 23 | 108,1 | 40 | 119,7 | 57 | 127,7 | 74 | 136,4 | 91 | 144,9 | 108 | 157,4 |
| 7 | 83,8 | 24 | 108,7 | 41 | 120,5 | 58 | 127,8 | 75 | 136,4 | 92 | 145,3 | 109 | 158,4 |
| 8 | 85,1 | 25 | 109,1 | 42 | 120,8 | 59 | 128,2 | 76 | 136,7 | 93 | 146,0 | 110 | 158,7 |
| 9 | 86,1 | 26 | 109,3 | 43 | 121,2 | 60 | 128,5 | 77 | 137,4 | 94 | 146,1 | 111 | 159,0 |
| 10 | 86,8 | 27 | 110,0 | 44 | 121,7 | 61 | 128,9 | 78 | 138,6 | 95 | 147,5 | 112 | 162,3 |
| 11 | 89,0 | 28 | 110,1 | 45 | 121,8 | 62 | 129,2 | 79 | 138,8 | 96 | 148,3 | 113 | 164,4 |

Продолжение таблицы 2.2.2

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12 | 90,1 | 29 | 112,9 | 46 | 122,2 | 63 | 129,4 | 80 | 139,4 | 97 | 148,5 | 114 | 169,8 |
| 13 | 95,5 | 30 | 113,0 | 47 | 122,3 | 64 | 130,0 | 81 | 139,9 | 98 | 148,7 | 115 | 174,6 |
| 14 | 97,9 | 31 | 113,8 | 48 | 122,8 | 65 | 131,1 | 82 | 139,9 | 99 | 153,3 | 116 | 175,7 |
| 15 | 100,0 | 32 | 115,4 | 49 | 122,9 | 66 | 131,5 | 83 | 140,4 | 100 | 153,6 | 117 | 187,4 |
| 16 | 102,9 | 33 | 116,3 | 50 | 123,9 | 67 | 132,5 | 84 | 140,5 | 101 | 154,0 |  |  |
| 17 | 103,2 | 34 | 117,0 | 51 | 124,2 | 68 | 132,5 | 85 | 140,9 | 102 | 154,5 |  |  |

Из табл. 2.2.2 можно увидеть, что наименьшее значение в выборке , а наибольшее значение .

Преобразование полученной выборки в вариационный ряд с абсолютными частотами представлено в табл. 2.2.3.

Таблица 2.2.3 – Вариационный ряд с абсолютными частотами

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 64,5 | 1 | 24 | 108,7 | 1 | 47 | 122,3 | 1 | 70 | 135,7 | 1 | 93 | 148,7 | 1 |
| 2 | 71,1 | 1 | 25 | 109,1 | 1 | 48 | 122,8 | 1 | 71 | 136,4 | 2 | 94 | 153,3 | 1 |
| 3 | 71,9 | 1 | 26 | 109,3 | 1 | 49 | 122,9 | 1 | 72 | 136,7 | 1 | 95 | 153,6 | 1 |
| 4 | 72,6 | 1 | 27 | 110,0 | 1 | 50 | 123,9 | 1 | 73 | 137,4 | 1 | 96 | 154,0 | 1 |
| 5 | 74,1 | 1 | 28 | 110,1 | 1 | 51 | 124,2 | 2 | 74 | 138,6 | 1 | 97 | 154,5 | 1 |
| 6 | 82,7 | 1 | 29 | 112,9 | 1 | 52 | 124,3 | 1 | 75 | 138,8 | 1 | 98 | 154,7 | 1 |
| 7 | 83,8 | 1 | 30 | 113,0 | 1 | 53 | 124,5 | 1 | 76 | 139,4 | 1 | 99 | 155,4 | 1 |
| 8 | 85,1 | 1 | 31 | 113,8 | 1 | 54 | 124,7 | 1 | 77 | 139,9 | 2 | 100 | 155,5 | 1 |
| 9 | 86,1 | 1 | 32 | 115,4 | 1 | 55 | 126,7 | 1 | 78 | 140,4 | 1 | 101 | 155,8 | 1 |
| 10 | 86,8 | 1 | 33 | 116,3 | 1 | 56 | 127,7 | 1 | 79 | 140,5 | 1 | 102 | 156,5 | 1 |
| 11 | 89,0 | 1 | 34 | 117,0 | 1 | 57 | 127,8 | 1 | 80 | 140,9 | 1 | 103 | 157,4 | 1 |
| 12 | 90,1 | 1 | 35 | 117,9 | 1 | 58 | 128,2 | 1 | 81 | 141,7 | 1 | 104 | 158,4 | 1 |
| 13 | 95,5 | 1 | 36 | 118,0 | 1 | 59 | 128,5 | 1 | 82 | 143,1 | 1 | 105 | 158,7 | 1 |
| 14 | 97,9 | 1 | 37 | 118,4 | 1 | 60 | 128,9 | 1 | 83 | 143,4 | 1 | 106 | 159,0 | 1 |
| 15 | 100,0 | 1 | 38 | 119,0 | 1 | 61 | 129,2 | 1 | 84 | 143,6 | 1 | 107 | 162,3 | 1 |
| 16 | 102,9 | 1 | 39 | 119,5 | 1 | 62 | 129,4 | 1 | 85 | 143,8 | 1 | 108 | 164,4 | 1 |
| 17 | 103,2 | 1 | 40 | 119,7 | 1 | 63 | 130,0 | 1 | 86 | 144,9 | 1 | 109 | 169,8 | 1 |
| 18 | 103,6 | 1 | 41 | 120,5 | 1 | 64 | 131,1 | 1 | 87 | 145,3 | 1 | 110 | 174,6 | 1 |
| 19 | 105,3 | 1 | 42 | 120,8 | 1 | 65 | 131,5 | 1 | 88 | 146,0 | 1 | 111 | 175,7 | 1 |
| 20 | 105,8 | 1 | 43 | 121,2 | 1 | 66 | 132,5 | 2 | 89 | 146,1 | 1 | 112 | 187,4 | 1 |

Продолжение таблицы 2.2.3

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 21 | 106,7 | 1 | 44 | 121,7 | 1 | 67 | 134,1 | 2 | 90 | 147,5 | 1 |  |  |  |
| 22 | 107,5 | 1 | 45 | 121,8 | 1 | 68 | 134,2 | 1 | 91 | 148,3 | 1 |  |  |  |
| 23 | 108,1 | 1 | 46 | 122,2 | 1 | 69 | 134,9 | 1 | 92 | 148,5 | 1 |  |  |  |

Преобразование полученной выборки в вариационный ряд с относительными частотами представлено в табл. 2.2.4.

Таблица 2.2.4 - Вариационный ряд с относительными частотами

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 64,5 | 0,008547 | 29 | 112,9 | 0,008547 | 57 | 127,8 | 0,008547 | 85 | 143,8 | 0,008547 |
| 2 | 71,1 | 0,008547 | 30 | 113,0 | 0,008547 | 58 | 128,2 | 0,008547 | 86 | 144,9 | 0,008547 |
| 3 | 71,9 | 0,008547 | 31 | 113,8 | 0,008547 | 59 | 128,5 | 0,008547 | 87 | 145,3 | 0,008547 |
| 4 | 72,6 | 0,008547 | 32 | 115,4 | 0,008547 | 60 | 128,9 | 0,008547 | 88 | 146,0 | 0,008547 |
| 5 | 74,1 | 0,008547 | 33 | 116,3 | 0,008547 | 61 | 129,2 | 0,008547 | 89 | 146,1 | 0,008547 |
| 6 | 82,7 | 0,008547 | 34 | 117,0 | 0,008547 | 62 | 129,4 | 0,008547 | 90 | 147,5 | 0,008547 |
| 7 | 83,8 | 0,008547 | 35 | 117,9 | 0,008547 | 63 | 130,0 | 0,008547 | 91 | 148,3 | 0,008547 |
| 8 | 85,1 | 0,008547 | 36 | 118,0 | 0,008547 | 64 | 131,1 | 0,008547 | 92 | 148,5 | 0,008547 |
| 9 | 86,1 | 0,008547 | 37 | 118,4 | 0,008547 | 65 | 131,5 | 0,008547 | 93 | 148,7 | 0,008547 |
| 10 | 86,8 | 0,008547 | 38 | 119,0 | 0,008547 | 66 | 132,5 | 0,017094 | 94 | 153,3 | 0,008547 |
| 11 | 89,0 | 0,008547 | 39 | 119,5 | 0,008547 | 67 | 134,1 | 0,017094 | 95 | 153,6 | 0,008547 |
| 12 | 90,1 | 0,008547 | 40 | 119,7 | 0,008547 | 68 | 134,2 | 0,008547 | 96 | 154,0 | 0,008547 |
| 13 | 95,5 | 0,008547 | 41 | 120,5 | 0,008547 | 69 | 134,9 | 0,008547 | 97 | 154,5 | 0,008547 |
| 14 | 97,9 | 0,008547 | 42 | 120,8 | 0,008547 | 70 | 135,7 | 0,008547 | 98 | 154,7 | 0,008547 |
| 15 | 100,0 | 0,008547 | 43 | 121,2 | 0,008547 | 71 | 136,4 | 0,017094 | 99 | 155,4 | 0,008547 |
| 16 | 102,9 | 0,008547 | 44 | 121,7 | 0,008547 | 72 | 136,7 | 0,008547 | 100 | 155,5 | 0,008547 |
| 17 | 103,2 | 0,008547 | 45 | 121,8 | 0,008547 | 73 | 137,4 | 0,008547 | 101 | 155,8 | 0,008547 |
| 18 | 103,6 | 0,008547 | 46 | 122,2 | 0,008547 | 74 | 138,6 | 0,008547 | 102 | 156,5 | 0,008547 |
| 19 | 105,3 | 0,008547 | 47 | 122,3 | 0,008547 | 75 | 138,8 | 0,008547 | 103 | 157,4 | 0,008547 |
| 20 | 105,8 | 0,008547 | 48 | 122,8 | 0,008547 | 76 | 139,4 | 0,008547 | 104 | 158,4 | 0,008547 |
| 21 | 106,7 | 0,008547 | 49 | 122,9 | 0,008547 | 77 | 139,9 | 0,017094 | 105 | 158,7 | 0,008547 |
| 22 | 107,5 | 0,008547 | 50 | 123,9 | 0,008547 | 78 | 140,4 | 0,008547 | 106 | 159,0 | 0,008547 |
| 23 | 108,1 | 0,008547 | 51 | 124,2 | 0,017094 | 79 | 140,5 | 0,008547 | 107 | 162,3 | 0,008547 |
| 24 | 108,7 | 0,008547 | 52 | 124,3 | 0,008547 | 80 | 140,9 | 0,008547 | 108 | 164,4 | 0,008547 |
| 25 | 109,1 | 0,008547 | 53 | 124,5 | 0,008547 | 81 | 141,7 | 0,008547 | 109 | 169,8 | 0,008547 |

Продолжение таблицы 2.2.4

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 26 | 109,3 | 0,008547 | 54 | 124,7 | 0,008547 | 82 | 143,1 | 0,008547 | 110 | 174,6 | 0,008547 |
| 27 | 110,0 | 0,008547 | 55 | 126,7 | 0,008547 | 83 | 143,4 | 0,008547 | 111 | 175,7 | 0,008547 |
| 28 | 110,1 | 0,008547 | 56 | 127,7 | 0,008547 | 84 | 143,6 | 0,008547 | 112 | 187,4 | 0,008547 |

Для определения количества интервалов используем формулу Стерджесса: где – объем выборки.

Используя в качестве , получаем, что .

Чтобы определить шаг, с которым формировать интервалы, использована формула:

Соответственно, для , и получаем, что .

Полученный интервальный ряд приведен в табл. 2.2.5.

Таблица 2.2.5 – Интервальный ряд

|  |  |  |
| --- | --- | --- |
| **Интервал** | **Абсолютная частота** | **Относительная частота** |
| [64,5 ; 79,5) | 5 | 0,042735 |
| [79,5 ; 94,5) | 7 | 0,059829 |
| [94,5 ; 109,5) | 14 | 0,119658 |
| [109,5 ; 124,5) | 27 | 0,230769 |
| [124,5 ; 139,5) | 27 | 0,230769 |
| [139,5 ; 154,5) | 21 | 0,179487 |
| [154,5 ; 169,5) | 12 | 0,102564 |
| [169,5 ; 184,5) | 3 | 0,025641 |
| [184,5 ; 187,4] | 1 | 0,008547 |

В сумме абсолютные частоты дают 117, что соответствует объему выборки, а относительные частоты суммируются к единице.

Полигон, построенный применительно к интервальному ряду для абсолютных частот представлен на рис. 2.2.1.

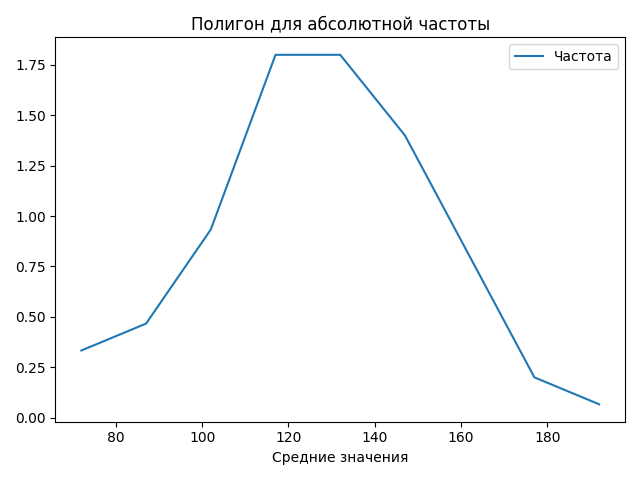
**

Рисунок 2.2.1 – Полигон для абсолютной частоты

Полигон, построенный применительно к интервальному ряду для относительных частот представлен на рис. 2.2.2.

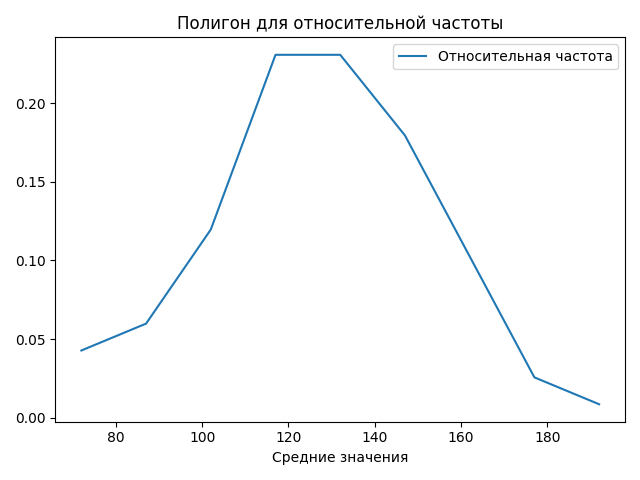


Рисунок 2.2.2 – Полигон для относительной частоты

Гистограмма, построенная применительно к интервальному ряду для абсолютных частот представлен на рис. 2.2.3.

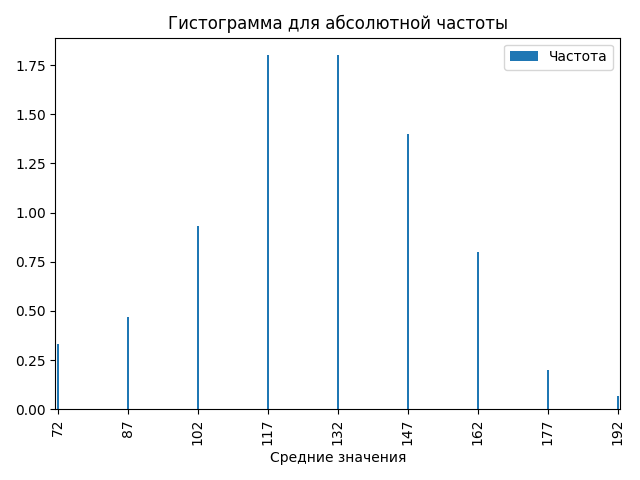


Рисунок 2.2.3 – Гистограмма для абсолютной частоты

Гистограмма, построенная применительно к интервальному ряду для относительных частот представлен на рис. 2.2.4.

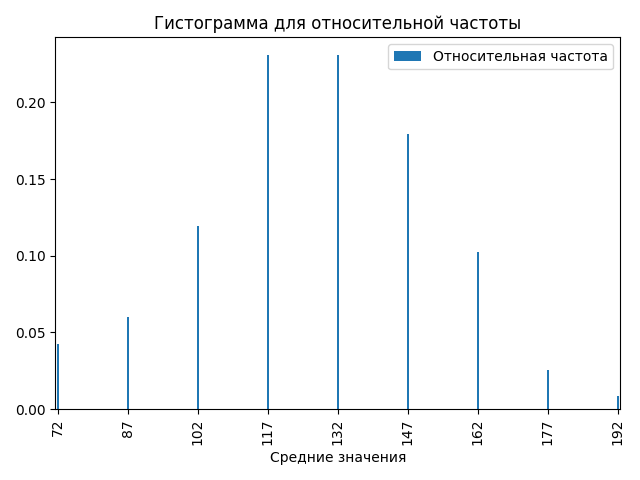
**

Рисунок 2.2.4 – Гистограмма для относительной частоты

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

Функция распределения:

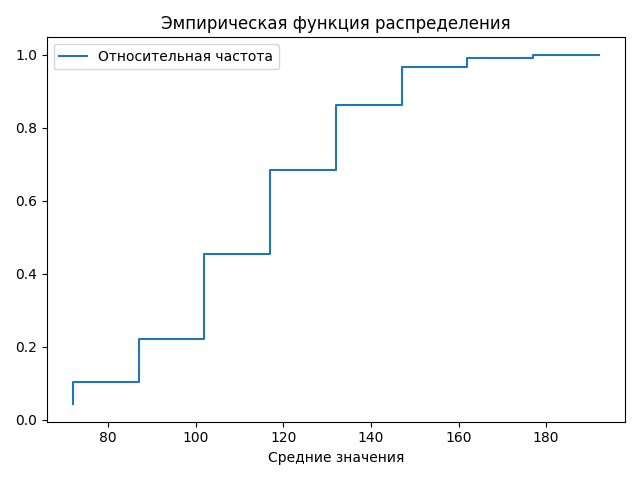


Рисунок 2.2.5 – График эмпирической функции распределения

Найдем условные моменты по формуле:

где длина интервала, – ложный ноль.

Результаты вычислений представлены в табл. 2.2.6.

Таблица 2.2.6

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| 72 | 5 | -4 | -20 | 80 | -320 | 1280 | 405 |
| 87 | 7 | -3 | -21 | 63 | -189 | 567 | 112 |
| 102 | 14 | -2 | -28 | 56 | -112 | 224 | 14 |
| 117 | 27 | -1 | -27 | 27 | -27 | 27 | 0 |
| 132 | 27 | 0 | 0 | 0 | 0 | 0 | 27 |
| 147 | 21 | 1 | 21 | 21 | 21 | 21 | 336 |
| 162 | 12 | 2 | 24 | 48 | 96 | 192 | 972 |
| 177 | 3 | 3 | 9 | 27 | 81 | 243 | 768 |
| 192 | 1 | 4 | 4 | 16 | 64 | 256 | 625 |
|  |  |  |  |  |  |  |  |
| Условные моменты: | | |  |  |  |  |  |

Проверим правильность вычислений:

Вычислим статистические оценки математического ожидания:

Вычислим статистические оценки дисперсии:

Отсюда следует, что среднеквадратическое отклонение:

Найдем исправленную выборочную дисперсию:

Для вычисления ассиметрии и эксцесса найдем центральные эмпирические моменты третьего и четвертого порядка:

Вычислим ассиметрию:

Вычислим эксцесс:

Далее найдем моду вариационного ряда по формуле:

Далее найдем медиану вариационного ряда по формуле:

Построим двумерный интервальный вариационный ряд (табл. 2.2.7).

Таблица 2.2.7 - Двумерный интервальный вариационный ряд

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | [64,5; 79,5) | [79,5; 94,5) | [94,5; 109,5) | [109,5; 124,5) | [124,5; 139,5) | [139,5; 154,5) | [154,5; 169,5) | [169,5; 184,5) | [184,5; 187,4] | N |
| [320,354) | 4 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| [354,388) | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| [388,422) | 0 | 3 | 9 | 14 | 1 | 0 | 0 | 0 | 0 | 27 |
| [422,456) | 0 | 0 | 1 | 10 | 12 | 2 | 0 | 0 | 0 | 25 |
| [456,490) | 0 | 0 | 0 | 3 | 12 | 9 | 0 | 0 | 0 | 24 |
| [490,524) | 0 | 0 | 0 | 0 | 2 | 8 | 6 | 1 | 0 | 17 |
| [524,558) | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 0 | 0 | 7 |
| [558,592) | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 3 |
| [592,593] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| N | 5 | 7 | 14 | 27 | 27 | 21 | 12 | 3 | 1 | 117 |

Основываясь на двумерном интервальном вариационном ряде построим корреляционную таблицу. Результат представлен в табл. 2.2.8.

Таблица 2.2.8 - Корреляционная таблица

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-4** | **-3** | **-2** | **-1** | **0** | **1** | **2** | **3** | **4** |  |
| **-4** | 4 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| **-3** | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| **-2** | 0 | 3 | 9 | 14 | 1 | 0 | 0 | 0 | 0 | 27 |
| **-1** | 0 | 0 | 1 | 10 | 12 | 2 | 0 | 0 | 0 | 25 |
| **0** | 0 | 0 | 0 | 3 | 12 | 9 | 0 | 0 | 0 | 24 |
| **1** | 0 | 0 | 0 | 0 | 2 | 8 | 6 | 1 | 0 | 17 |
| **2** | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 0 | 0 | 7 |
| **3** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 3 |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
|  | 5 | 7 | 14 | 27 | 27 | 21 | 12 | 3 | 1 | 117 |

Исходя из результатов корреляционной таблицы вычислим статистическую оценку корреляционного момента по формуле где - условные средние для условных вариант, – оценки стандартных отклонений условных вариант. Для вычисления построим вспомогательную таблицу. Результаты представлены в табл. 2.2.9.

Таблица 2.2.9 – Вспомогательная таблица для вычисления статистической оценки коэффициента корреляции

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **-4** | **-3** | **-2** | **-1** | **0** | **1** | **2** | **3** | **4** | **N** |
| **-4** | 64 | 48 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 120 |
| **-3** | 12 | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 30 |
| **-2** | 0 | 18 | 36 | 28 | 0 | 0 | 0 | 0 | 0 | 82 |
| **-1** | 0 | 0 | 2 | 10 | 0 | -2 | 0 | 0 | 0 | 10 |
| **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 0 | 0 | 0 | 8 | 12 | 3 | 0 | 23 |
| **2** | 0 | 0 | 0 | 0 | 0 | 4 | 20 | 0 | 0 | 24 |
| **3** | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 18 | 0 | 24 |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 16 |
| **N** | 76 | 66 | 64 | 38 | 0 | 10 | 38 | 21 | 16 | 329 |

Вычислим коэффициент корреляции. Для этого оценим значение для случая нормального распределения по формуле:

Получили коэффициент корреляции отличный от нуля, а значит случайные величины в выборке коррелированы и зависимы.

Построим доверительный интервал для коэффициента корреляции при уровне значимости .Для этого перейдем к случайной величине :

Вычислим СКВО для распределения:

Доверительный интервал для с доверительной вероятностью будет определяться:

где должно удовлетворять условию:

Тогда для

Тогда для

Для построения доверительного интервала для коэффициента корреляции сделаем обратное преобразование Фишера:

Для

Для

При увеличении уровня надежности получили более широкий доверительный интервал.

Проверим гипотезу о равенстве нулю коэффициента корреляции. Вычислим по формуле:

Для уровня значимости и было найдено .

Исходя из того, что , гипотеза о равенстве нулю коэффициента корреляции отвергается.

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

Отобразим двумерную выборку на графике и для заданной выборки построим уравнения средней квадратичной регрессии на и на соответственно. Построим полученные прямые на множестве выборки.

Линейная функция среднеквадратической регрессии 𝑦(𝑥) для заданной выборки:

;

.

Линейная функция среднеквадратической регрессии для заданной выборки:

Двумерная выборка и графики линейной функции выборочной среднеквадратической регрессии и представлены на рис. 2.3.1.

Найдем оценки остаточной дисперсии для полученных выборочных уравнений регрессии:

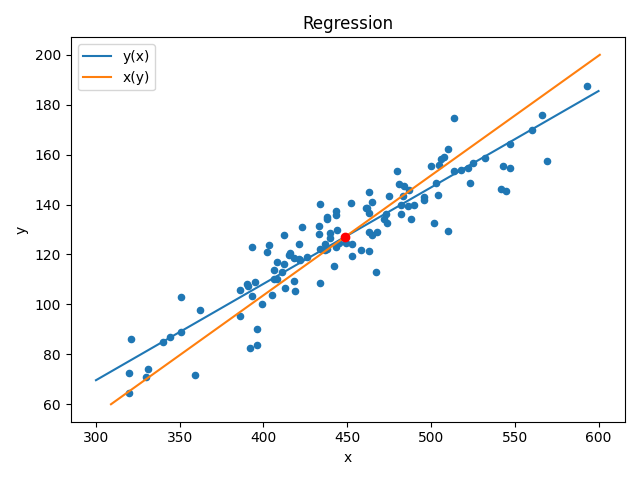
****

Рисунок 2.3.1 - Графики линейной функции выборочной среднеквадратической регрессии и

Составим корреляционную таблицу для нахождения выборочного корреляционного отношения. Убедимся, что неравенства и выполняются.

Таблица 2.3.1 - Корреляционная таблица

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***72*** | ***87*** | ***102*** | ***117*** | ***132*** | ***147*** | ***162*** | ***177*** | ***192*** |  |  |  |
| ***337*** | 0 | 4 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 9 | 82 | 100 |
| ***371*** | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 94,5 | 168,750 |
| ***405*** | 0 | 3 | 9 | 14 | 1 | 0 | 0 | 0 | 0 | 27 | 109,222 | 122,840 |
| ***439*** | 0 | 0 | 1 | 10 | 12 | 2 | 0 | 0 | 0 | 25 | 126 | 108 |
| ***473*** | 0 | 0 | 0 | 3 | 12 | 9 | 0 | 0 | 0 | 24 | 135,75 | 98,438 |
| ***507*** | 0 | 0 | 0 | 0 | 2 | 8 | 6 | 1 | 0 | 17 | 152,294 | 130,796 |
| ***541*** | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 0 | 0 | 7 | 157,714 | 45,918 |
| ***575*** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 3 | 172 | 50 |
| ***609*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 192 | 0 |
|  | 5 | 7 | 14 | 27 | 27 | 21 | 12 | 3 | 1 | 117 |  |  |
|  | 343,8 | 366,143 | 395,286 | 425,148 | 457,889 | 489,190 | 526,833 | 552,333 | 609 |  |  |  |
|  | 46,24 | 23,592 | 94,367 | 191,874 | 228,346 | 317,179 | 200,694 | 513,778 | 0 |  |  |  |

Для того, чтобы посчитать выборочное корреляционное отношение, рассчитаем внутригрупповую, межгрупповую, общую дисперсии.

Расчёт выборочного корреляционного отношения к :

Расчёт выборочного корреляционного отношения к :

Проверим выполнение неравенств и :

Неравенства и выполняются.

Для заданной выборки построим корреляционную кривую параболического вида .

Для определения коэффициентов корреляционной кривой параболического вида была решена следующая система уравнений:

Система была решена с помощью написанной программы на языке Python (код представлен в ПРИЛОЖЕНИИ Г).В результате работы программы были получены следующие значения коэффициентов:

Полученное уравнение примет вид:

График квадратичной функции выборочной среднеквадратической регрессии представлен на рис. 2.3.2.

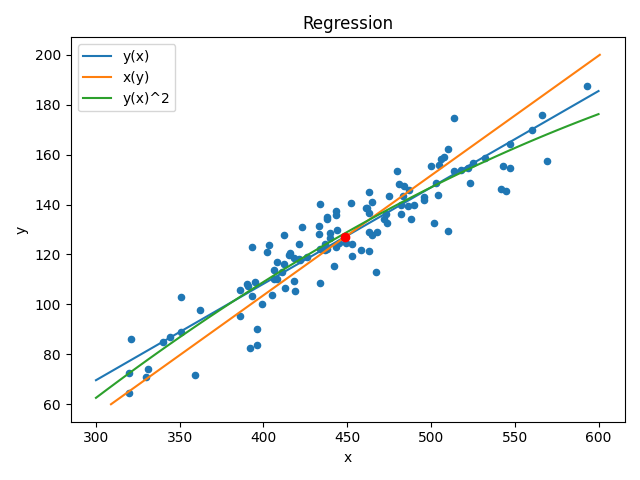


Рисунок 2.3.2 - Корреляционная кривая параболического вида

Построим корреляционную кривую экспоненциальной функции .

Запишем выборочное уравнение в виде .

Найдем коэффициенты и с помощью написанной программы на языке Python (код представлен в ПРИЛОЖЕНИИ Г), получим следующие коэффициенты:

Корреляционная кривая экспоненциального вида имеет следующий вид:

.

График полученной кривой на множестве выборки представлен на рис. 2.3.3.

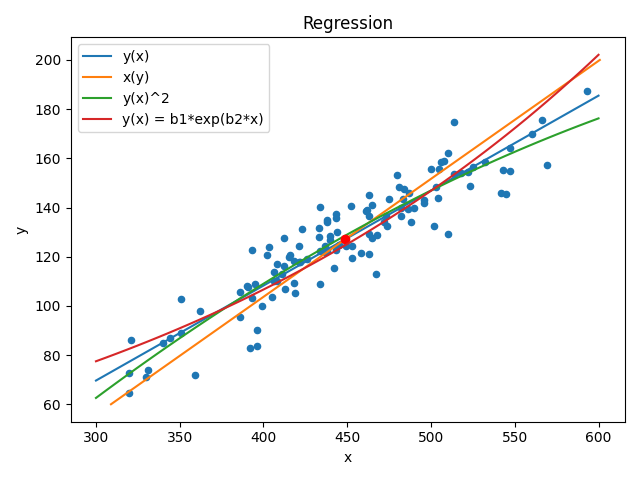


Рисунок 2.3.3 - Корреляционная кривая экспоненциального вида

**2.4. Выводы.**

Были освоены основные понятия, связанные с корреляционной зависимостью между случайными величинами, статистическими гипотезами и проверкой их «справедливости».

Была сформирована вторая выборка данных и осуществлена её подготовка к статическому анализу. Выборка приведена к ранжированному, вариационному и интервальному видам. Используя полученный интервальный ряд построен полигон, гистограмма и эмпирическая функция распределения для абсолютных и относительных частот. Из полученного ранжированного ряда сразу видны минимальное и максимальное значение выборки. В данном случае были получены значения , . По полученному вариационному ряду видно что наибольшая частота . По сформированному интервальному ряду можно увидеть, что большинство значений выборки сконцентрированы в интервалах [109,5; 124,5) и [124,5; 139,5). Более наглядно это представляют построенные гистограммы и полигоны частот. При этом их форма не зависит от того, какие частоты используются – абсолютные или относительные.

Также были получены практические навыки нахождения точечных статистических оценок параметров распределения. При вычислении условных моментов была сделана проверка, которая показала, что данные моменты были посчитаны верно. Так как полученное значение эксцесса , то можно сделать вывод, что плотность закона распределения случайной величины уменьшается медленно вблизи её моды. Из полученного значения коэффициента симметрии можно сделать вывод, что мода немного смещена вправо относительно середины распределения, так как , но при этом находится достаточно близко к центру, так как значение близко к 0.

Был построен двумерный интервальный вариационный ряд, по нему была построена корреляционная таблица, с помощью которой была вычислена статистическая оценка корреляционного момента . Коэффициент корреляции и отличный от нуля, а значит случайные величины в выборке коррелированы и зависимы.

Был построен доверительный интервал для коэффициента корреляции при уровне значимости :

При увеличении уровня надежности получили более широкий доверительный интервал.

Также была проверена гипотеза о равенстве коэффициента корреляции нулю при заданном уровне значимости . Из полученного результата можно сделать вывод, что гипотеза отвергается, т.к. , соответственно, коэффициент корреляции не равен нулю.

Также были получены уравнения прямых среднеквадратической регрессии: и . Найдены корреляционные соотношения и . Эти значения близки к 1, что говорит о том, что между и есть сильная статистическая зависимость. В свою очередь было найдено и построено уравнение выборочных кривых для параболической среднеквадратической регрессии и была построена корреляционной кривая экспоненциального вида .

**3. Кластерный анализ**

**3.1. Основные теоретические положения**

Кластерный анализ.

Кластерный анализ – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.

К характеристикам кластера относятся в частности: центр, радиус; среднеквадратическое отклонение; размер кластера.

Центр кластера – это среднее геометрическое место точек, принадлежащих кластеру, в пространстве данных.

Радиус кластера – максимальное расстояние точек, принадлежащих кластеру, от центра кластера.

Кластеры могут быть перекрывающимися. В этом случае невозможно при помощи используемых процедур однозначно отнести объект к одному из двух или более кластеров. Такие объекты называют спорными.

Спорный объект - это объект, который по мере сходства может быть отнесен к более, чем одному кластеру.

Размер кластера может быть определен либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

Большое значение в кластерном анализе имеет выбор масштаба. Пусть, например, значения переменной *х* превышают 100, а переменной *у* - в интервале от 0 до 1.

Тогда, при расчете расстояния между точками переменная *х*, будет практически полностью доминировать над переменной *у*. В результате практически невозможно корректно рассчитать расстояния между точками.

Расстоянием (метрикой) между объектами *a* и *b* пространстве параметров называется такая величина *dab*, которая удовлетворяет аксиомам:



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

непрерывна; 

Существует возможность простого перехода от расстояния к мерам близости:

.

Метод k-средних.

Алгоритм -means – это наиболее популярный метод кластеризации, который разделяет определенный набор данных на заданное пользователем число кластеров . Алгоритм прост для реализации и запуска, относительно быстрый, легко адаптируется и распространен на практике. Это исторически один из самых важных алгоритмов интеллектуального анализа данных.

Суть алгоритма заключается в том, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

где  *–* это число кластеров,  *–* полученные кластеры, и центры масс.

Центроиды выбираются в тех местах, где визуально скопление точек выше. Алгоритм разбивает множество элементов векторного пространства на заранее известное число кластеров . Основная идея заключается в том, что на каждой итерации перевычисляется 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. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение  не увеличивается, поэтому зацикливание невозможно.

Возможны две разновидности метода *k* -средних.

Первая предполагает пересчет центра кластера после каждого изменения его состава, как рассмотрено выше, а вторая —лишь после завершения цикла.

В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется. Перед началом работы метода целесообразно нормировать характеристики объектов:  .

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

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

В качестве таких функционалов могут быть использованы:

1. Сумма квадратов расстояний до центров кластеров



2. Сумма внутрикластерных расстояний между объектами



3. Сумма внутрикластерных дисперсий

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

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

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

Метод поиска сгущений.

Метод поиска сгущений является еще одним итеративным методом кластерного анализа.

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

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

В алгоритме поиска сгущений сначала выбирается начальный центр первого кластера. Выбор такого объекта может быть произвольным, а может основываться на предварительном анализе точек и их окрестностей. В рассматриваемом случае, центры выбираются вручную.

Как правило, на первом шаге центром сферы служит объект (точка), в ближайшей (заданной) окрестности которого расположено наибольшее число соседей. На основе заданного радиуса сферы (R) определяется совокупность точек внутри этой сферы, и для них вычисляются координаты центра (вектор средних для попавших в сферу значений признаков).

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 расстояние между -м и -м объектами, то в качестве нижней границы значения радиуса сферы можно выбрать минимальное из таких расстояний, а в качестве верхней границы - максимальное:

Тогда, если начинать работу алгоритма с

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

Следует отметить следующие существенные при реализации метода поиска сгущений моменты:

1. В случае разномасштабности квалификационных признаков необходимо проведение их нормировки перед началом работы метода;
2. Возможны два варианта реализации метода. Один из них не предполагает изменения заданного значения радиуса сферы до завершения кластеризации, а другой — предполагает изменение этого радиуса в процессе кластеризации при начале построения очередной сферы;
3. В отличие от метода k-средних метод поиска сгущений не требует задания количества кластеров, на которые предполагается разбить исходное множество объектов;
4. Качество полученного в результате применения метода итогового разбиения на кластеры оценивается, как и в методе k-средних, с помощью введенных на предыдущей лекции критериев качества разбиения , ,.
5. Получение в результате кластеризации пересекающихся кластеров (наличие спорных объектов) в принципе является неудовлетворительным результатом. На практике в этом случае необходимо скорректировать процесс, либо выбрать другой метод кластеризации.

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



2. Сумма внутрикластерных расстояний между объектами



3. Сумма внутрикластерных дисперсий

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

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

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

**3.2. Метод k-средних**

Нормализуем множество точек и отобразим полученное множество.

Отображение исходной выборки представлено на рис. 3.2.1.

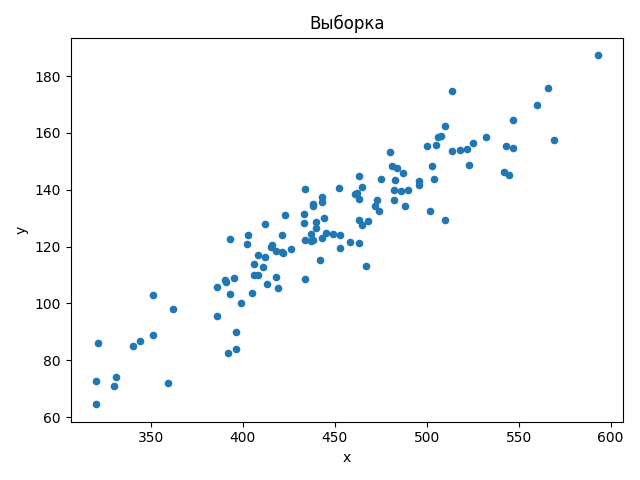


Рисунок 3.2.1 – Исходная выборка

Нормализация координат точек определяется по формуле:

Отображение нормализованной выборки представлено на рис. 3.2.2.

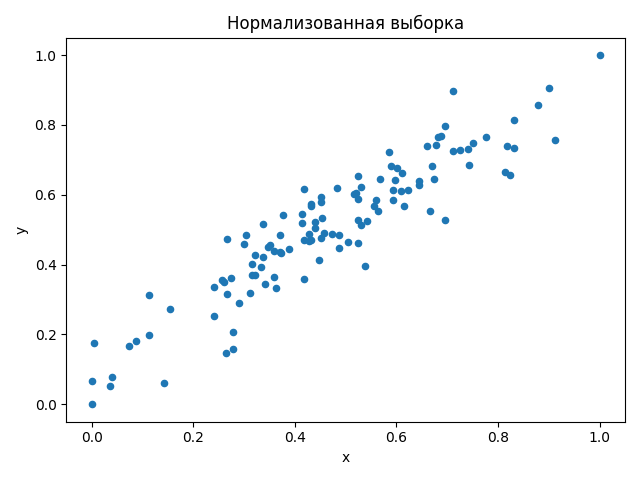


Рисунок 3.2.2 ­– Нормализованная выборка

Определим верхнюю оценку количества кластеров по формуле: , где – число точек:

Реализуем алгоритм k-means и отобразим полученные кластеры, выделим каждый кластер разным цветом, отметим центроиды.

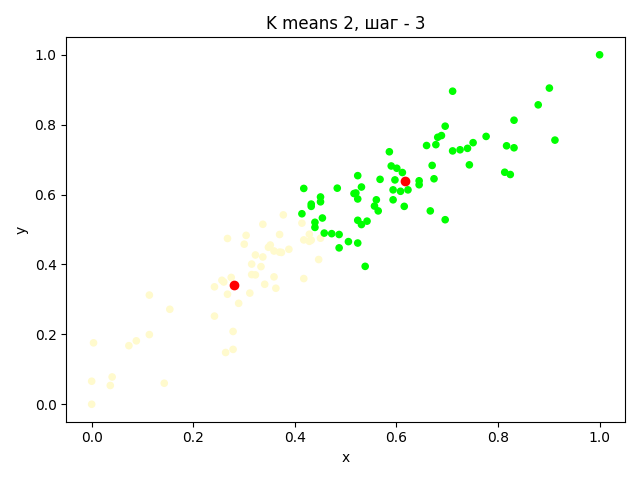


Рисунок 3.2.3 – Кластеризация алгоритмом k-means (2 кластера)

Таблица 3.2.1

|  |  |  |
| --- | --- | --- |
| Номер кластера | Центр кластера | Количество элементов в кластере |
| 1 | (0.28075845722904547; 0.34033727404712905) | 51 |
| 2 | (0.6171606171606171; 0.6384717804571344) | 66 |

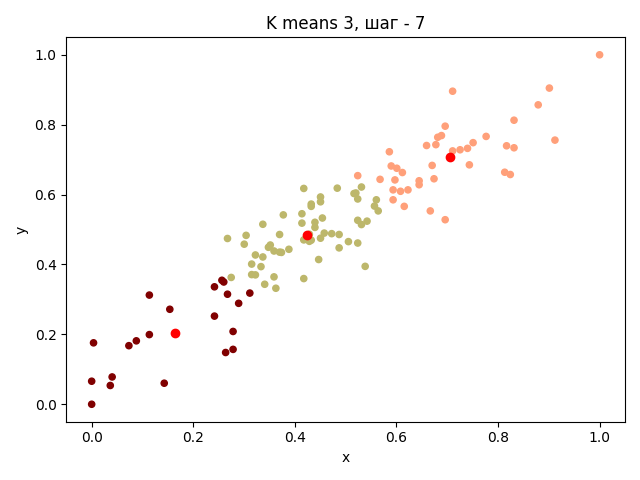


Рисунок 3.2.4 – Кластеризация алгоритмом k-means (3 кластера)

Таблица 3.2.2

|  |  |  |
| --- | --- | --- |
| Номер кластера | Центр кластера | Количество элементов в кластере |
| 1 | (0.7045177045177047; 0.7068494293880787) | 39 |
| 2 | (0.16448630734345018; 0.204502305397342) | 21 |
| 3 | (0.42317331791016; 0.48481863731745967) | 57 |

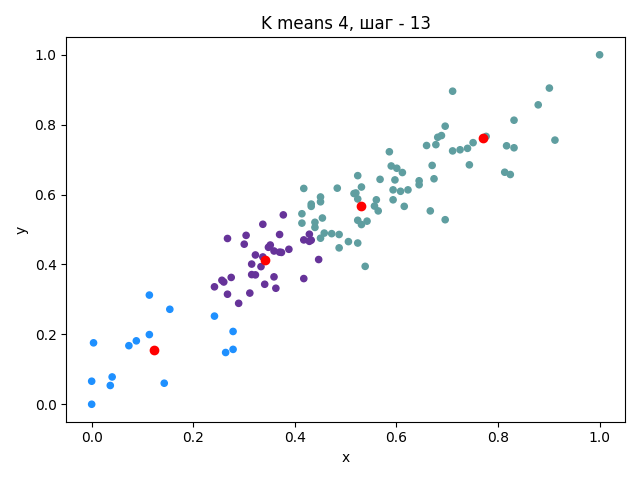


Рисунок 3.2.5 – Кластеризация алгоритмом k-means (4 кластера)

Таблица 3.2.3

|  |  |  |
| --- | --- | --- |
| Номер кластера | Центр кластера | Количество элементов в кластере |
| 1 | (0.7701863354037267; 0.7629391162840059) | 23 |
| 2 | (0.12185592185592185; 0.15546514781665308) | 15 |
| 3 | (0.5307285307285308; 0.5685561884097278) | 45 |
| 4 | (0.34195216548157736; 0.41269803283396345) | 34 |

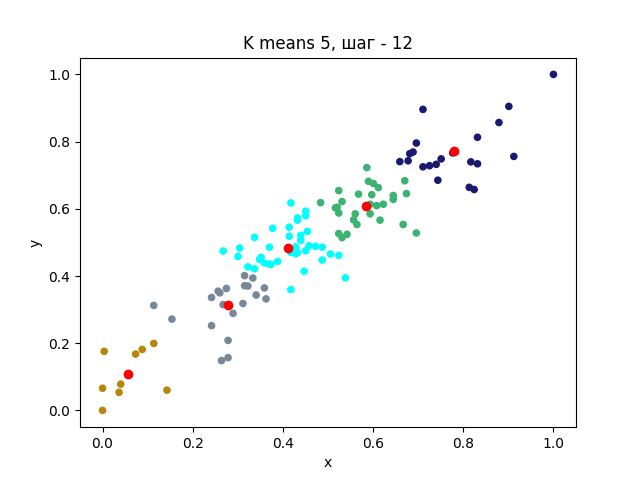


Рисунок 3.2.6 – Кластеризация алгоритмом k-means (5 кластеров)

Таблица 3.2.4

|  |  |  |
| --- | --- | --- |
| Номер кластера | Центр кластера | Количество элементов в кластере |
| 1 | (0.779522065236351; 0.7723274826610872) | 21 |
| 2 | (0.5848174813692055; 0.6087371285878621) | 29 |
| 3 | (0.05535205535205535; 0.10912214085525718) | 9 |
| 4 | (0.4117023327549644; 0.48396214294891016) | 38 |
| 5 | (0.2789377289377289; 0.31257119609438566) | 20 |

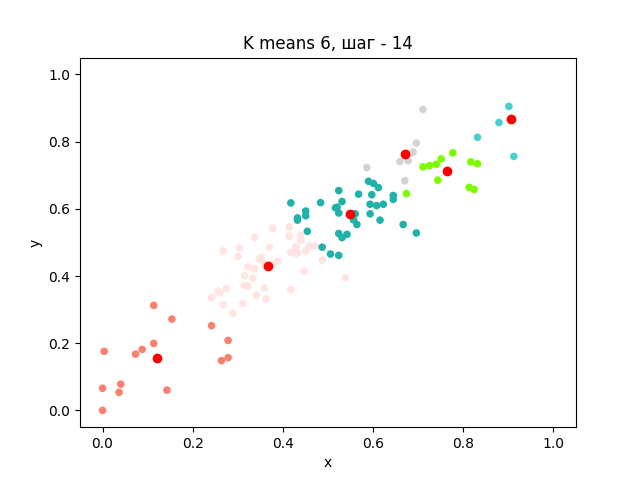


Рисунок 3.2.7 – Кластеризация алгоритмом k-means (6 кластеров)

Таблица 3.2.5

|  |  |  |
| --- | --- | --- |
| Номер кластера | Центр кластера | Количество элементов в кластере |
| 1 | (0.7642357642357642; 0.7114431540794436) | 11 |
| 2 | (0.12185592185592185; 0.15546514781665308) | 15 |
| 3 | (0.5480900052328623; 0.5850517261420435) | 35 |
| 4 | (0.9047619047619048; 0.8660699755899104) | 5 |
| 5 | (0.6712454212454213; 0.7642392188771359) | 8 |
| 6 | (0.36604480790527305; 0.42831191931424684) | 43 |

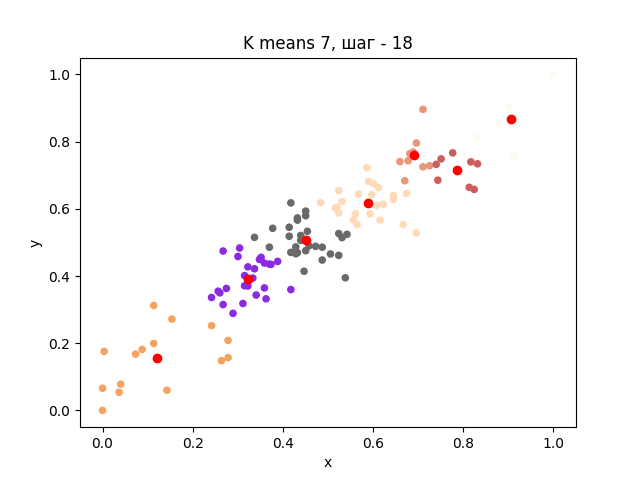


Рисунок 3.2.8 – Кластеризация алгоритмом k-means (7 кластеров)

Таблица 3.2.6

|  |  |  |
| --- | --- | --- |
| Номер кластера | Центр кластера | Количество элементов в кластере |
| 1 | (0.9047619047619048; 0.8660699755899104) | 5 |
| 2 | (0.6910866910866911; 0.7605098996474099) | 9 |
| 3 | (0.4518125552608312; 0.5060183496534889) | 29 |
| 4 | (0.3226260918568611; 0.3916254616010514) | 26 |
| 5 | (0.5876923076923076; 0.6162082994304312) | 25 |
| 6 | (0.7870879120879122; 0.7159275834011392) | 8 |
| 7 | (0.12185592185592185; 0.15546514781665308) | 15 |

Для проведения оценки качества разбиения для различных разбиений используются функционалы качества:

1. Сумма квадратов расстояний до центров кластеров



2. Сумма внутрикластерных расстояний между объектами



3. Сумма внутрикластерных дисперсий

Здесь - дисперсия -й переменной в -м кластере.

Для различных значений рассчитаем функционалы качества и результаты занесём в табл. 3.2.7.

Таблица 3.2.7

|  |  |  |  |
| --- | --- | --- | --- |
| **Количество кластеров** |  |  |  |
| 2 |  |  |  |
| 3 |  |  |  |
| 4 |  |  |  |
| 5 |  |  |  |
| 6 |  |  |  |
| 7 |  |  |  |

По полученным данным можно сделать вывод о том, что с увеличением числа кластеров, минимизируются значения перечисленных функционалов качества.

**3.3. Метод поиска сгущений**

Отображение исходной выборки представлено на рис. 3.3.1.

Нормализация координат точек определяется по формуле:

Отображение нормализованной выборки представлено на рис. 3.3.2.

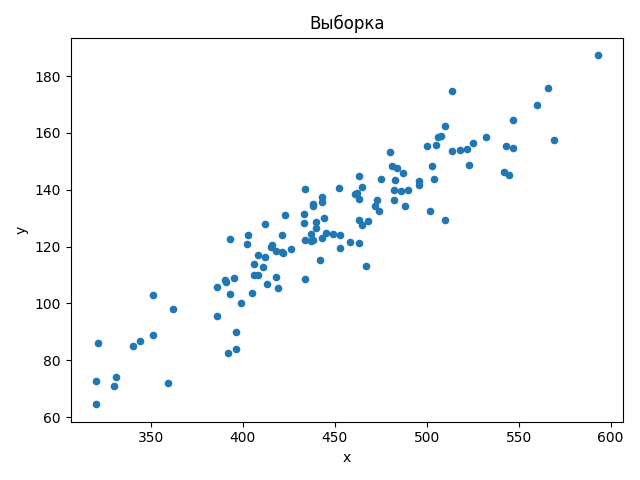


Рисунок 3.3.1 – Исходная выборка



Рисунок 3.3.2 – Нормализованная выборка

Реализуем алгоритм поиска сгущений. Отобразим полученные кластеры, выделим каждый кластер разным цветом, отметим центроиды.

Определим нижнюю и верхнюю границы радиуса сферы:

Выберем из промежутка [] радиус .

Запустим алгоритм:

Формирование 1го кластера представлено на рис. 3.3.3.

|  |  |
| --- | --- |
|  |  |
|  |  |

Рисунок 3.3.3

Формирование 2го кластера представлено на рис. 3.3.4.

|  |  |
| --- | --- |
|  |  |
|  |  |

Рисунок 3.3.4

Формирование 3го кластера представлено на рис. 3.3.5 и на рис. 3.3.6.

|  |  |
| --- | --- |
|  |  |

Рисунок 3.3.5

|  |  |
| --- | --- |
|  |  |
|  |  |

Рисунок 3.3.6

Формирование 4го кластера представлено на рис. 3.3.7.

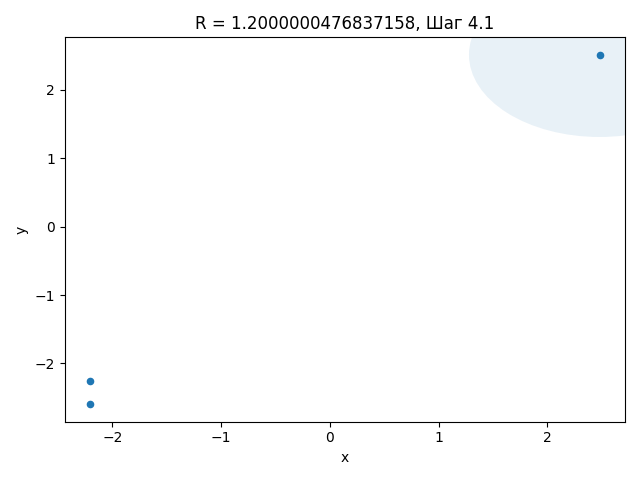


Рисунок 3.3.7

Формирование 5го кластера представлено на рис. 3.3.8.

|  |  |
| --- | --- |
|  |  |

Рисунок 3.3.8

Результат кластеризации представлен на рис. 3.3.9.

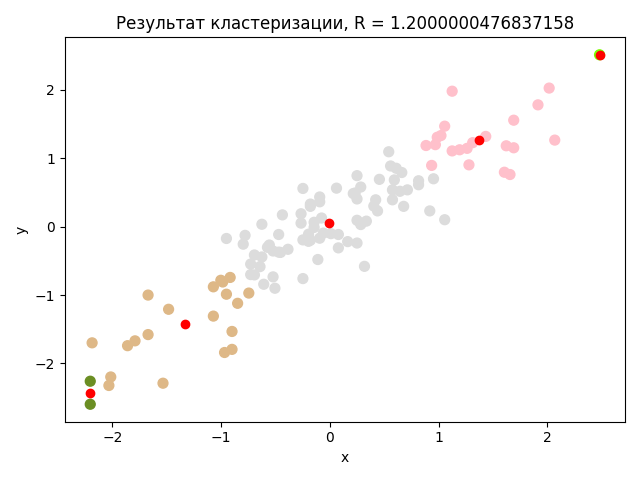


Рисунок 3.3.9

Таблица 3.3.2

|  |  |  |
| --- | --- | --- |
| Номер  кластера | Центр кластера | Количество элементов в кластере |
| 1 | (-0,004243736116897832 ; 0,05696773691479348) | 73 |
| 2 | (1,3728603917013382 ; 1,269402589075278) | 21 |

Продолжение таблицы 3.3.2

|  |  |  |
| --- | --- | --- |
| 3 | (-1,3296908010805915 ; -1,423519995425545) | 20 |
| 4 | (2,4783418922683094 ; 2,5082120327432573) | 1 |
| 5 | (-2,2024006799255216 ; -2,4269556447966085) | 2 |

Проверим чувствительность метода к погрешностям.Для этого сформируем кластеры с погрешностями. Формирование 1го кластера с погрешностями представлено на рис. 3.3.10.

|  |  |
| --- | --- |
|  |  |
|  |  |

Рисунок 3.3.10

Формирование 2го кластера с погрешностями представлено на рис. 3.3.11.

|  |  |
| --- | --- |
|  |  |

Рисунок 3.3.11

Формирование 2го кластера с погрешностями представлено на рис. 3.3.12.

|  |  |
| --- | --- |
|  |  |
|  |  |

Рисунок 3.3.12-1

|  |  |
| --- | --- |
|  |  |

Рисунок 3.3.12-2

Формирование 2го кластера с погрешностями представлено на рис. 3.3.13.

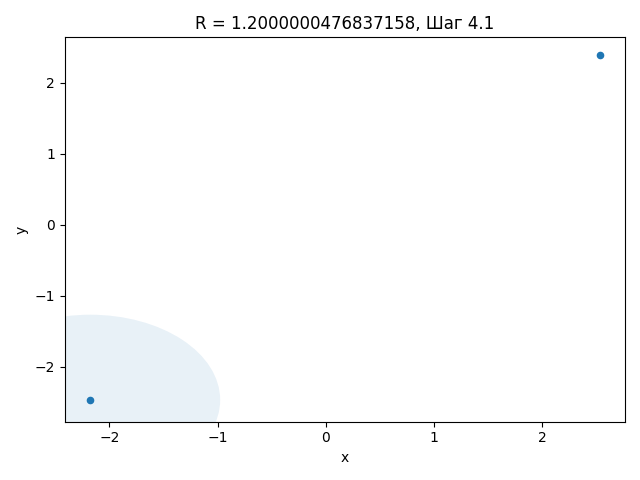
****

Рисунок 3.3.13

Формирование 5го кластера с погрешностями представлено на рис. 3.3.14.

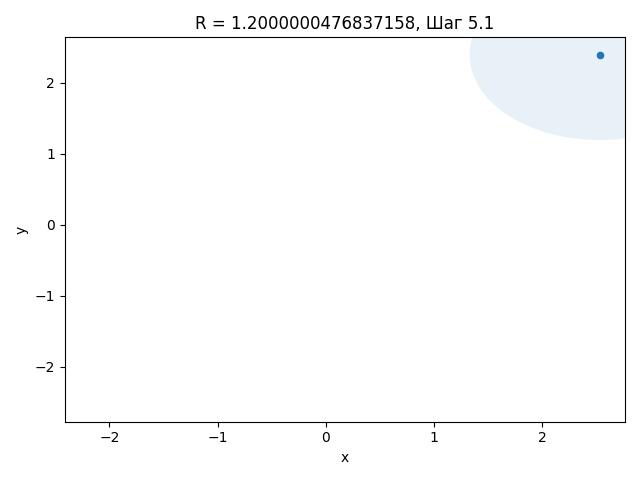
****

Рисунок 3.3.14

Результат кластеризации с погрешностями представлен на рис. 3.3.15.

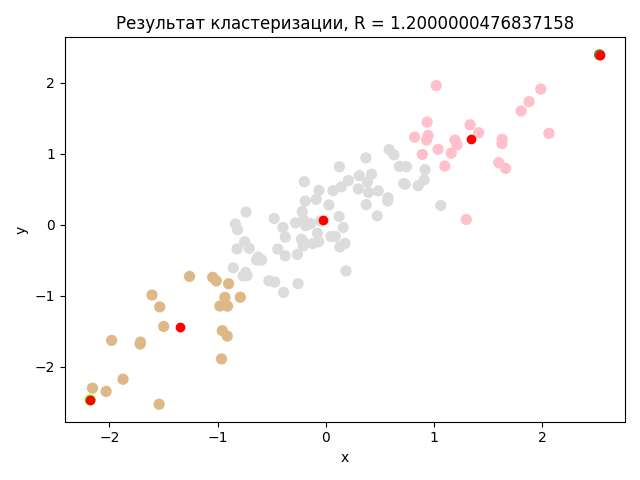
****

Рисунок 3.3.15

Таблица 3.3.3

|  |  |  |
| --- | --- | --- |
| Номер  кластера | Центр кластера | Количество элементов в кластере |
| 1 | (-0,0294042169854218 ; 0,06868813095241261) | 72 |

Продолжение таблицы 3.3.3

|  |  |  |
| --- | --- | --- |
| 2 | (1,342186825428838 ; 1,210104850517267) | 22 |
| 3 | (-1,34686405930226 ; -1,4401333952484419) | 21 |
| 4 | (-2,1754394392808516 ; -2,465932878978102) | 1 |
| 5 | (2,528303696654842 ; 2,393057608215714) | 1 |

Для определения чувствительности метода поиска сгущений к погрешностям посчитаем функционалы качества:

1. Сумма квадратов расстояний до центров кластеров



2. Сумма внутрикластерных расстояний между объектами



3. Сумма внутрикластерных дисперсий

Здесь - дисперсия -й переменной в -м кластере.

Для метода поиска сгущений рассчитаем функционалы качества:

Для метода поиска сгущений c учетом погрешностей рассчитаем функционалы качества:

На основании этого можем сделать вывод, что метод несильно, но чувствителен к погрешностям, так как значения функционалов качества с учетом погрешностей возросли.

Сравнить метод поиска сгущений с методом k-средних. Для этого возьмем значения функционалов качества при количестве кластеров из подраздела 3.2.

Таблица 3.3.4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Метод** | **Количество кластеров** |  |  |  |
| k-средних | 5 |  |  |  |
| поиска сгущений | 5 |  |  |  |

Визуальное представление работы двух методов представлено на рис. 3.3.16.

|  |  |
| --- | --- |
| Метод поиска сгущений | Метод k-средних |
|  |  |

Рисунок 3.3.16

Рассмотрим результаты работы двух методов на рис. 3.3.16. Визуальная разница довольно большая. В данном случае алгоритм k-средних показал себя лучше так, как максимальное расстояние между точками в его кластерах меньше, чем у метода поиска сгущений. Сравним значения функционалов качества для данных разбиений (табл. 4). Аналогичный вывод можно сделать для сравнения по функционалам качества: метод k-средних лучше показал себя для исходных данных.

**3.4. Выводы.**

Были освоены основные понятия кластерного анализа.

В методе k-средних верхняя оценка количества кластеров была посчитана по формуле: и равна 7. С помощью алгоритма k-means исходная выборка была разбита на различное количество кластеров: 2, 3, 4, 5, 6, 7. С увеличением числа кластеров, уменьшаются значения функционалов качества, используемых в работе. Было также замечено, что чем больше кластеров, тем больше шагов необходимо проделать алгоритму, чтобы на последнем шаге имели минимальное значение. Из этого можно сделать выводы, что разбиение каждый раз улучшалось и в итоге получилось оптимальным.

С помощью алгоритма поиска сгущений исходная выборка была разбита на 5 кластеров. Метод поиска сгущений несильно, но чувствителен к погрешностям. При сравнении данного метода с методом k-средних метод поиска сгущений оказался хуже для исходных данных. По результатам работы, можно заметить, что алгоритм поиска сгущений требует значительное количество итераций, также заранее неизвестно количество получаемых кластеров, оно зависит от выбранного радиуса.

**заключение**

Были выполнены все поставленные цели: построены выборки из генеральной совокупности заданного объёма, построены статистические, ранжированные, вариационные и интервальные ряды, графически построены полигоны частот, гистограммы, эмпирические функции распределения двухмерной выборки, найдены выборочные оценки: среднего, дисперсии, исправленной дисперсии, СКВО, асимметрии, эксцесса, медианы и моды, построены доверительные интервалы для математического ожидания и СКО, проверена гипотеза о нормальном законе по критерию Пирсона, построена корреляционная таблица, найдена оценка коэффициента корреляции, проверена гипотеза о равенстве коэффициента корреляции нулю, построены уравнения выборочных прямых среднеквадратической регрессии, найдены оценки корреляционных отношений, нормализовано множество точек, реализован алгоритм k-средних, отображены полученные кластеры, реализован метод поиска сгущений, проверена чувствительность метода поиска сгущений к погрешностям, произведено сравнение методов, произведена оценка качества кластеризации.

**список использованных источников**

1. Белоногов А.М., Попов Ю.И., Посредник О.В. Статиcтическая обработка результатов физического эксперимента [Комплект] : учеб. пособие: - СПб. : Изд-во CП6ГЭТУ "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. Егоров В.А. и др. Анализ однородных статистически данных: учеб. пособие: — СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2005.

4. Котельников Р.Б. Анализ результатов наблюдений.

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

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

7. Методические указания к лабораторным работам: учеб.-метод. пособие / сост.: А.-В.И. Середа. СПб. 2016. 15 с.

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

9. Кластеризация // machinelearning.ru URL: http://www.machinelearning.ru/wiki/index.php?title=Линейная\_регрессия (дата обращения: 05.04.2021).

10. Линейная регрессия // machinelearning.ru URL: http://www.machinelearning.ru/wiki/index.php?title=Линейная\_регрессия (дата обращения: 05.04.2021).

**приложение А**

**Программа для формирования и первичной обработки выборки, построения, ранжированного и интервального рядов.**

**lab1.py**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv('sample.csv', header=None).iloc[:, 0]

ranked\_series = df.sort\_values()

variation\_series = ranked\_series.apply(lambda x: sum(ranked\_series == x))

relative\_var\_series = variation\_series.apply(lambda x: x / len(df))

variation\_df = pd.DataFrame({'Значение': ranked\_series, 'Частота': variation\_series,

'Относительная частота': relative\_var\_series}).drop\_duplicates()

k = 1.33 + 3.31 \* np.log10(len(df))

k = int(np.round(k, 0))

h = (max(ranked\_series) - min(ranked\_series)) / k

h = int(np.round(h, 0))

means = []

nums = []

relative\_nums = []

distrib\_nums = []

low\_range = []

up\_range = []

for i in np.arange(np.min(ranked\_series), np.max(ranked\_series), h):

cond = (i <= variation\_df['Значение']) & \

(variation\_df['Значение'] < i + h) \

if i + h < max(ranked\_series) \

else (i <= variation\_df['Значение']) & (variation\_df['Значение'] <= i + h)

means.append((i \* 2 + h) / 2)

nums.append(variation\_df['Частота'][cond].sum())

relative\_nums.append(variation\_df['Относительная частота'][cond].sum())

distrib\_nums.append(variation\_df['Относительная частота'][variation\_df['Значение'] < i + h].sum())

low\_range.append(i)

up\_range.append(i+h)

if nums[-1] <= 3:

nums[-2] += nums[-1]

del nums[-1]

relative\_nums[-2] += relative\_nums[-1]

del relative\_nums[-1]

means[-2] = (means[-2] + means[-1]) / 2

del means[-1]

up\_range[-2] = up\_range[-1]

del up\_range[-1]

del low\_range[-1]

distrib\_nums[-2] = distrib\_nums[-1]

del distrib\_nums[-1]

if \_\_name\_\_ == '\_\_main\_\_':

ranked\_series.to\_csv('Ранжированный\_ряд.csv', index=0, header=None)

variation\_df.to\_csv('Вариационный\_ряд.csv', index=0, header=None)

inter\_df = pd.DataFrame({'Средние значения': means, 'Частота': nums}, dtype=np.int64)

inter\_df.to\_csv('Интервальный\_Ряд.csv', index=0, header=None)

relative\_inter\_df = pd.DataFrame({'Средние значения': means, 'Относительная частота': relative\_nums},

dtype=np.int64)

relative\_inter\_df.to\_csv('Интервальный\_ряд\_относительные\_частоты.csv', index=0, header=None)

distrib\_df = pd.DataFrame({'Средние значения': means, 'Относительная частота': distrib\_nums}, dtype=np.int64)

distrib\_df.to\_csv('Функция\_распределения.csv', index=0, header=None)

fig = inter\_df.plot(x='Средние значения', y='Частота', title='Полигон для абсолютной частоты')

plt.show()

fig = inter\_df.plot(x='Средние значения', y='Частота', kind='bar', title='Гистограмма для абсолютной частоты')

plt.show()

fig = relative\_inter\_df.plot(x='Средние значения', y='Относительная частота',

title='Полигон для относительной частоты')

plt.show()

fig = relative\_inter\_df.plot(x='Средние значения', y='Относительная частота', kind='bar',

title='Гистограмма для относительной частоты')

plt.show()

fig = distrib\_df.plot(drawstyle="steps", x='Средние значения', y='Относительная частота', title='Эмпирическая функция распределения')

plt.show()

**приложение Б**

**Программа для** **нахождения точечных оценок параметров распределения.**

**lab2.py**

from lab1 import means, nums, h, df, low\_range, up\_range, variation\_df, ranked\_series

import pandas as pd

import numpy as np

inter\_df = pd.DataFrame({'Средние значения': means, 'Частота': nums}, dtype=np.int64)

C = inter\_df.iloc[4, 0]

inter\_df['Условные варианты'] = inter\_df['Средние значения'].apply(lambda x: (x - C) / h)

moments = []

for num\_of\_moment in range(1, 5):

col = 'Условный момент {}'.format(num\_of\_moment)

inter\_df[col] = inter\_df.iloc[:, 1:3].apply(lambda x: x[0]\*(x[1]\*\*num\_of\_moment), axis=1)

moments.append(inter\_df[col].sum() / len(df))

if \_\_name\_\_ == '\_\_main\_\_':

print(col, ': ', moments[-1])

inter\_df['Проверка'] = inter\_df.iloc[:, 1:3].apply(lambda x: x[0]\*((x[1]+1)\*\*4), axis=1)

inter\_df.to\_csv('Таблица.csv', index=0)

start\_moment\_1\_usl = moments[0]\*h + C

central\_moment\_2\_usl = (moments[1] - moments[0]\*\*2)\*(h\*\*2)

central\_moment\_3\_usl = (moments[2] - 3\*moments[1]\*moments[0] + 2\*(moments[0]\*\*3))\*(h\*\*3)

central\_moment\_4\_usl = (moments[3] - 4\*moments[2]\*moments[0] + 6\*moments[1]\*(moments[0]\*\*2) - 3\*(moments[0]\*\*4))\*(h\*\*4)

start\_moment\_1\_emp = inter\_df.iloc[:, :2].apply(lambda x: x[0]\*x[1], axis=1).sum() / len(df)

central\_moment\_2\_emp = inter\_df.iloc[:, :2].apply(lambda x: ((x[0] - start\_moment\_1\_emp)\*\*2)\*x[1], axis=1).sum() / len(df)

s = np.sqrt((len(df)/(len(df)-1)) \* central\_moment\_2\_emp)

asim = central\_moment\_3\_usl / (s\*\*3)

ecs = central\_moment\_4\_usl / (s\*\*4) - 3

max\_low\_val = 0

max\_num = 0

max\_low\_val\_prev = 0

max\_low\_val\_next = 0

sum\_median\_prev\_nums = 0

for i, (n, l, u) in enumerate(list(zip(nums, low\_range, up\_range))):

if n > max\_num:

max\_num = n

max\_low\_val = l

try:

max\_low\_val\_prev = nums[i-1]

max\_low\_val\_next = nums[i+1]

except Exception:

max\_low\_val\_prev = 0

max\_low\_val\_next = 0

if i < int(len(nums)/2):

sum\_median\_prev\_nums += n

moda = max\_low\_val + h \* ((max\_num - max\_low\_val\_prev)/(2\*max\_num - max\_low\_val\_prev - max\_low\_val\_next))

median\_int = int(len(nums)/2)

median\_num = nums[median\_int]

x\_0\_median = low\_range[median\_int]

median = x\_0\_median + h\*((0.5 \* sum(nums) - sum\_median\_prev\_nums) / median\_num)

if \_\_name\_\_ == '\_\_main\_\_':

print('Начальный условный момент 1го порядка: ', start\_moment\_1\_usl)

print('Центральный условный момент 2го порядка: ', central\_moment\_2\_usl)

print('Центральный условный момент 3го порядка: ', central\_moment\_3\_usl)

print('Центральный условный момент 3го порядка: ', central\_moment\_4\_usl)

print('Начальный эмпирический момент 1го порядка: ', start\_moment\_1\_emp)

print('Центральный эмпирический момент 2го порядка: ', central\_moment\_2\_emp)

print('Асимметрия: ', asim)

print('Эксцесса: ', ecs)

print('Мода: ', moda)

print('Медиана: ', median)

**приложение В**

**Программа для** **нахождения интервальных оценок параметров распределения и проверки статистической гипотезы о нормальном распределении.**

**lab3.py**

from lab2 import means, nums, h, df, low\_range, up\_range, start\_moment\_1\_emp, central\_moment\_2\_emp, s

import pandas as pd

import numpy as np

import math

print('Мат ожидание: ', start\_moment\_1\_emp)

print('Дисперсия: ', central\_moment\_2\_emp)

print('Среднекв отклонение: ', np.sqrt(central\_moment\_2\_emp))

print('Исправленная выб. дисперсия: ', s)

print('n=', len(df))

t = 1.98

print('t=', t)

sl = s \* t / np.sqrt(len(df))

i1 = start\_moment\_1\_emp - sl

i2 = start\_moment\_1\_emp + sl

print('Доверительный интервал для мат. ож. - ({}, {})'.format(i1, i2))

q = 0.14

print('Доверительный интервал для среднекв. отклонения - ({}, {})'.format(s \* (1 - q), s \* (1 + q)))

low\_range.append(up\_range[-1])

table = pd.DataFrame({'x': low\_range})

table['x-x\_mean'] = table['x'].apply(lambda x: x - start\_moment\_1\_emp)

table['v'] = table['x-x\_mean'].apply(lambda x: x / s)

def gauss\_integral(z):

return np.sqrt(np.pi) \* math.erf(z / np.sqrt(2)) / np.sqrt(2)

def gauss(x):

return np.exp(-np.power(x, 2.) / 2) / np.sqrt(2 \* np.pi)

table['f(v)'] = table['v'].apply(lambda x: gauss\_integral(x) / np.sqrt(2 \* np.pi))

table.to\_csv('Таблица1.csv', index=0)

v\_means = []

deltas = []

for i in range(0, len(table['v']) - 1):

v\_means.append((table['v'][i] + table['v'][i + 1]) / 2)

deltas.append(table['f(v)'][i + 1] - table['f(v)'][i])

table2 = pd.DataFrame({'v\_means': v\_means, 'p\_2': deltas})

table2['f(v\_mean)'] = table2['v\_means'].apply(gauss)

table2['p\_1'] = table2['f(v\_mean)'].apply(lambda x: x \* h / s)

table2['n\_1'] = table2['p\_1'].apply(lambda x: x \* len(df))

table2['n\_2'] = table2['p\_2'].apply(lambda x: x \* len(df))

table2.to\_csv('Таблица2.csv', index=0)

sol1 = pd.DataFrame({'nums': nums, 'n': table2['n\_1']})

sol1['nums-n'] = sol1.apply(lambda x: x[0]-x[1], axis=1)

sol1['(nums-n)^2'] = sol1['nums-n'].apply(lambda x: x\*\*2)

sol1['(nums-n)^2/n']=sol1.apply(lambda x: x[3]/x[1], axis=1)

print('Solution 1 summ: ', sol1['(nums-n)^2/n'].sum())

sol1.to\_csv('solution1.csv', index=0)

sol2 = pd.DataFrame({'nums': nums, 'n': table2['n\_2']})

sol2['nums-n'] = sol2.apply(lambda x: x[0]-x[1], axis=1)

sol2['(nums-n)^2'] = sol2['nums-n'].apply(lambda x: x\*\*2)

sol2['(nums-n)^2/n']=sol2.apply(lambda x: x[3]/x[1], axis=1)

print('Solution 2 summ: ', sol2['(nums-n)^2/n'].sum())

sol2.to\_csv('solution2.csv', index=0)

**приложение Г**

**Программа для** **нахождения элементов корреляционного анализа и проверки статистической гипотезы о равенстве коэффициента корреляции нулю.**

**lab1\_y.py**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv('sample.csv', header=None).iloc[:, 1]

ranked\_series = df.sort\_values()

variation\_series = ranked\_series.apply(lambda x: sum(ranked\_series == x))

relative\_var\_series = variation\_series.apply(lambda x: x / len(df))

variation\_df = pd.DataFrame({'Значение': ranked\_series, 'Частота': variation\_series,

'Относительная частота': relative\_var\_series}).drop\_duplicates()

k = 1.33 + 3.31 \* np.log10(len(df))

k = int(np.round(k, 0))

h = (max(ranked\_series) - min(ranked\_series)) / k

h = int(np.round(h, 0))

means = []

nums = []

relative\_nums = []

distrib\_nums = []

low\_range = []

up\_range = []

for i in np.arange(np.min(ranked\_series), np.max(ranked\_series), h):

cond = (i <= variation\_df['Значение']) & \

(variation\_df['Значение'] < i + h) \

if i + h < max(ranked\_series) \

else (i <= variation\_df['Значение']) & (variation\_df['Значение'] <= i + h)

means.append((i \* 2 + h) / 2)

nums.append(variation\_df['Частота'][cond].sum())

relative\_nums.append(variation\_df['Относительная частота'][cond].sum())

distrib\_nums.append(variation\_df['Относительная частота'][variation\_df['Значение'] < i + h].sum())

low\_range.append(i)

up\_range.append(i+h)

if nums[-1] <= 3:

nums[-2] += nums[-1]

del nums[-1]

relative\_nums[-2] += relative\_nums[-1]

del relative\_nums[-1]

means[-2] = (means[-2] + means[-1]) / 2

del means[-1]

up\_range[-2] = up\_range[-1]

del up\_range[-1]

del low\_range[-1]

distrib\_nums[-2] = distrib\_nums[-1]

del distrib\_nums[-1]

if \_\_name\_\_ == '\_\_main\_\_':

ranked\_series.to\_csv('Ранжированный\_ряд.csv', index=0, header=None)

variation\_df.to\_csv('Вариационный\_ряд.csv', index=0, header=None)

inter\_df = pd.DataFrame({'Средние значения': means, 'Частота': nums}, dtype=np.int64)

inter\_df.to\_csv('Интервальный\_Ряд.csv', index=0, header=None)

relative\_inter\_df = pd.DataFrame({'Средние значения': means, 'Относительная частота': relative\_nums},

dtype=np.int64)

relative\_inter\_df.to\_csv('Интервальный\_ряд\_относительные\_частоты.csv', index=0, header=None)

distrib\_df = pd.DataFrame({'Средние значения': means, 'Относительная частота': distrib\_nums}, dtype=np.int64)

distrib\_df.to\_csv('Функция\_распределения.csv', index=0, header=None)

fig = inter\_df.plot(x='Средние значения', y='Частота', title='Полигон для абсолютной частоты')

plt.show()

fig = inter\_df.plot(x='Средние значения', y='Частота', kind='bar', title='Гистограмма для абсолютной частоты')

plt.show()

fig = relative\_inter\_df.plot(x='Средние значения', y='Относительная частота',

title='Полигон для относительной частоты')

plt.show()

fig = relative\_inter\_df.plot(x='Средние значения', y='Относительная частота', kind='bar',

title='Гистограмма для относительной частоты')

plt.show()

fig = distrib\_df.plot(x='Средние значения', y='Относительная частота', title='Эмпирическая функция распределения', drawstyle='steps')

plt.show()

**lab2\_y.py**

from lab1\_y import means, nums, h, df, low\_range, up\_range, variation\_df, ranked\_series, k

import pandas as pd

import numpy as np

inter\_df = pd.DataFrame({'Средние значения': means, 'Частота': nums}, dtype=np.int64)

C = inter\_df.iloc[4, 0]

inter\_df['Условные варианты'] = inter\_df['Средние значения'].apply(lambda x: (x - C) / h)

moments = []

for num\_of\_moment in range(1, 5):

col = 'Условный момент {}'.format(num\_of\_moment)

inter\_df[col] = inter\_df.iloc[:, 1:3].apply(lambda x: x[0]\*(x[1]\*\*num\_of\_moment), axis=1)

moments.append(inter\_df[col].sum() / len(df))

if \_\_name\_\_ == '\_\_main\_\_':

print(col, ': ', moments[-1])

inter\_df['Проверка'] = inter\_df.iloc[:, 1:3].apply(lambda x: x[0]\*((x[1]+1)\*\*4), axis=1)

inter\_df.to\_csv('Таблица.csv', index=0)

start\_moment\_1\_usl = moments[0]\*h + C

central\_moment\_2\_usl = (moments[1] - moments[0]\*\*2)\*(h\*\*2)

central\_moment\_3\_usl = (moments[2] - 3\*moments[1]\*moments[0] + 2\*(moments[0]\*\*3))\*(h\*\*3)

central\_moment\_4\_usl = (moments[3] - 4\*moments[2]\*moments[0] + 6\*moments[1]\*(moments[0]\*\*2) - 3\*(moments[0]\*\*4))\*(h\*\*4)

start\_moment\_1\_emp = inter\_df.iloc[:, :2].apply(lambda x: x[0]\*x[1], axis=1).sum() / len(df)

central\_moment\_2\_emp = inter\_df.iloc[:, :2].apply(lambda x: ((x[0] - start\_moment\_1\_emp)\*\*2)\*x[1], axis=1).sum() / len(df)

s = np.sqrt((len(df)/(len(df)-1)) \* central\_moment\_2\_emp)

asim = central\_moment\_3\_usl / (s\*\*3)

ecs = central\_moment\_4\_usl / (s\*\*4) - 3

max\_low\_val = 0

max\_num = 0

max\_low\_val\_prev = 0

max\_low\_val\_next = 0

sum\_median\_prev\_nums = 0

for i, (n, l, u) in enumerate(list(zip(nums, low\_range, up\_range))):

if n > max\_num:

max\_num = n

max\_low\_val = l

try:

max\_low\_val\_prev = nums[i-1]

max\_low\_val\_next = nums[i+1]

except Exception:

max\_low\_val\_prev = 0

max\_low\_val\_next = 0

if i < int(len(nums)/2):

sum\_median\_prev\_nums += n

moda = max\_low\_val + h \* ((max\_num - max\_low\_val\_prev)/(2\*max\_num - max\_low\_val\_prev - max\_low\_val\_next))

median\_int = int(len(nums)/2)

median\_num = nums[median\_int]

x\_0\_median = low\_range[median\_int]

median = x\_0\_median + h\*((0.5 \* sum(nums) - sum\_median\_prev\_nums) / median\_num)

if \_\_name\_\_ == '\_\_main\_\_':

print('Начальный условный момент 1го порядка: ', start\_moment\_1\_usl)

print('Центральный условный момент 2го порядка: ', central\_moment\_2\_usl)

print('Центральный условный момент 3го порядка: ', central\_moment\_3\_usl)

print('Центральный условный момент 3го порядка: ', central\_moment\_4\_usl)

print('Начальный эмпирический момент 1го порядка: ', start\_moment\_1\_emp)

print('Центральный эмпирический момент 2го порядка: ', central\_moment\_2\_emp)

print('Асимметрия: ', asim)

print('Эксцесса: ', ecs)

print('Мода: ', moda)

print('Медиана: ', median)

**lab4.py**

import pandas as pd

import numpy as np

from lab2\_y import inter\_df as inter\_df\_y, means as means\_y, df as df\_y, up\_range as up\_range\_y, \

low\_range as low\_Range\_y, h as h\_y, s as s\_y, moments as m\_y, k, start\_moment\_1\_emp as mean\_y, nums as nums\_y

from lab2 import inter\_df as inter\_df\_x, means as means\_x, df as df\_x, up\_range as up\_range\_x, \

low\_range as low\_Range\_x, h as h\_x, s as s\_x, moments as m\_x, start\_moment\_1\_emp as mean\_x, nums as nums\_x

df = pd.read\_csv('sample.csv', header=None)

rows = []

for u\_x, l\_x in zip(up\_range\_x, low\_Range\_x):

cols = []

cond\_x = (l\_x <= df.iloc[:, 0]) & \

(df.iloc[:, 0] < u\_x) \

if u\_x < max(df.iloc[:, 0]) \

else (l\_x <= df.iloc[:, 0]) & (df.iloc[:, 0] <= u\_x)

rng = df.iloc[:, 1][cond\_x]

for u\_y, l\_y in zip(up\_range\_y, low\_Range\_y):

cond\_y = (l\_y <= rng) & \

(rng < u\_y) \

if u\_y < max(rng) \

else (l\_y <= rng) & (rng <= u\_y)

cols.append(sum(cond\_y))

rows.append(cols)

rows = np.array(rows)

cor\_table = pd.DataFrame(data=rows, index=means\_x, columns=means\_y)

cor\_table.to\_csv('Двумерный интервальный ряд.csv')

C\_x = means\_x[int(len(means\_x) / 2)]

C\_y = means\_y[int(len(means\_y) / 2)]

v\_x = (np.array(means\_x) - C\_x) / h\_x

v\_y = (np.array(means\_y) - C\_y) / h\_y

cor\_table.columns = v\_x

cor\_table = cor\_table.set\_index(v\_y)

cor\_table.to\_csv('Корялиционная таблица.csv')

rows = []

for i in range(len(cor\_table)):

cols = []

for j in range(len(cor\_table.columns)):

cols.append(cor\_table.to\_numpy()[i][j] \* v\_y[j] \* v\_x[i])

rows.append(cols)

help\_table = pd.DataFrame(data=rows, index=v\_x, columns=v\_y)

help\_table.to\_csv('Вспомогательная таблица.csv')

sums = help\_table.to\_numpy().sum(axis=0)

s = sums.sum()

cor = (s - len(df)\*m\_x[0]\*m\_y[0]) / (len(df) \* (s\_x/h\_x)\*(s\_y/h\_y))

zb = np.log((1+cor)/(1 - cor)) / 2

o\_b = 1 / np.sqrt(len(df) - 3)

z\_r = zb + 1.96 \*o\_b

z\_l = zb - 1.96 \*o\_b

T\_n = cor \* np.sqrt(len(df) - 2) / np.sqrt(1 - cor\*\*2)

if \_\_name\_\_=='\_\_main\_\_':

print('Суммы столбцов:', sums)

print('Сумма: ', s)

print("Коэф корреляции: ", cor)

print("S\_v: ", (s\_x/h\_x))

print("S\_u: ", (s\_y/h\_y))

print("v\_b: ", m\_x[0])

print("u\_b: ", m\_y[0])

print(

'({}:{})'.format((np.exp(2 \* z\_l) - 1) / (np.exp(2 \* z\_l) + 1), (np.exp(2 \* z\_r) - 1) / (np.exp(2 \* z\_r) + 1)))

print('k: ', k - 2)

print('T\_krit: ', 1.943)

print('T\_n: ', T\_n)

print('Отвергаем' if T\_n > 1.943 else 'Принимаем')

**приложение Д**

**Программа для нахождения элементов регрессионного анализа и построения выборочные прямых среднеквадратической регрессии, поиска корреляционного отношения.**

**lab5.py**

from lab4 import df, cor, s\_x, s\_y, mean\_x, mean\_y, means\_x, means\_y, k, cor\_table, nums\_x, nums\_y

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

def mean\_sq\_regression(s1, s2, mean1, mean2, cor, pref='y(x)='):

def inner\_foo(x, pr=False):

if pr:

print(pref + '{} \* x'.format(cor \* s1 / s2), '{0:+}'.format(mean1 - cor \* s1 / s2 \* mean2))

a1 = cor \* s1 / s2

a2 = mean1 - cor \* s1 / s2 \* mean2

return a2 + a1 \* x

return inner\_foo

msr\_x = mean\_sq\_regression(s\_x, s\_y, mean\_x, mean\_y, cor, 'x(y)=')

msr\_y = mean\_sq\_regression(s\_y, s\_x, mean\_y, mean\_x, cor)

ax = df.plot.scatter(x=0, y=1)

y1 = np.array([60, 200])

x2 = np.array([300, 600])

ax.plot(x2, msr\_y(x2, pr=True), label='y(x)')

ax.plot(msr\_x(y1, pr=True), y1, label='x(y)')

line1 = np.array([[x2[0], msr\_y(x2)[0]], [x2[1], msr\_y(x2)[1]]])

line2 = np.array([[msr\_x(y1)[0], y1[0]], [msr\_x(y1)[1], y1[1]]])

t, s = np.linalg.solve(np.array([line1[1] - line1[0], line2[0] - line2[1]]).T, line2[0] - line1[0])

ax.plot(\*((1 - t) \* line1[0] + t \* line1[1]), 'o', color='red')

ax.set\_title('Regression')

ax.set\_xlabel('x')

ax.set\_ylabel('y')

plt.legend()

plt.show()

cors = cor\_table.to\_numpy()

rows\_y = []

vars\_y = []

for row in range(len(cor\_table)):

cols = []

var = []

for col in range(len(cor\_table.columns)):

cols.append(cors[row][col] \* means\_y[col])

rows\_y.append(sum(cols) / nums\_x[row])

for col in range(len(cor\_table.columns)):

var.append(((means\_y[col] - rows\_y[-1]) \*\* 2) \* cors[row][col])

vars\_y.append(sum(var) / nums\_x[row])

cols\_x = []

vars\_x = []

for col in range(len(cor\_table.columns)):

rows = []

var = []

for row in range(len(cor\_table)):

rows.append(cors[row][col] \* means\_x[row])

cols\_x.append(sum(rows) / nums\_y[col])

for row in range(len(cor\_table)):

var.append(((means\_x[col] - cols\_x[-1]) \*\* 2) \* cors[row][col])

vars\_x.append(sum(var) / nums\_y[col])

cor\_table['n\_y'] = nums\_x

cor\_table['mean\_y\_gr'] = rows\_y

cor\_table['D\_y\_gr'] = vars\_y

cor\_table.to\_csv('Table1.csv')

pd.DataFrame({'n\_x': nums\_y, 'x\_mean\_gr': cols\_x, 'D\_x\_gr': vars\_x}).T.to\_csv('Table1\_last\_rows.csv')

x\_t = pd.DataFrame({'n\_x': nums\_y, 'x\_mean\_gr': cols\_x, 'D\_x\_gr': vars\_x})

D\_ingr\_yx = (cor\_table['D\_y\_gr'].to\_numpy() \* x\_t['n\_x'].to\_numpy()).sum() / len(df)

D\_ingr\_xy = (x\_t['D\_x\_gr'].to\_numpy() \* cor\_table['n\_y'].to\_numpy()).sum() / len(df)

D\_betwgr\_yx = (((cor\_table['mean\_y\_gr'] - mean\_y) \*\* 2).to\_numpy() \* x\_t['n\_x'].to\_numpy()).sum() / len(df)

D\_betwgr\_xy = (((x\_t['x\_mean\_gr'] - mean\_x) \*\* 2).to\_numpy() \* cor\_table['n\_y'].to\_numpy()).sum() / len(df)

D\_gen\_xy = D\_ingr\_xy + D\_betwgr\_xy

D\_gen\_yx = D\_ingr\_yx + D\_betwgr\_yx

n\_xy = np.sqrt(D\_betwgr\_xy / D\_gen\_xy)

n\_yx = np.sqrt(D\_betwgr\_yx / D\_gen\_yx)

x = df.iloc[:, 0]

y = df.iloc[:, 1]

system = []

b = []

for i in range(3):

line = []

for j in range(3):

line.append((x \*\* (4 - i - j)).sum())

system.append(line)

b.append((y \* (x \*\* (2 - i))).sum())

res = np.linalg.solve(np.array(system), np.array(b))

def sq\_regr(x):

return res[0] \* x \*\* 2 + res[1] \* x + res[2]

ax = df.plot.scatter(x=0, y=1)

y1 = np.array([60, 200])

x2 = np.array([300, 600])

ax.plot(x2, msr\_y(x2), label='y(x)')

ax.plot(msr\_x(y1), y1, label='x(y)')

x3 = np.linspace(300, 600)

ax.plot(x3, sq\_regr(x3), label='y(x)^2')

ax.plot(\*((1 - t) \* line1[0] + t \* line1[1]), 'o', color='red')

ax.set\_title('Regression')

ax.set\_xlabel('x')

ax.set\_ylabel('y')

y = df.iloc[:, 1]

x = df.iloc[:, 0]

z = np.log(y)

a1 = (len(df) \* (x\*z).sum() - x.sum() \* z.sum())/(len(df)\*(x\*x).sum()-x.sum()\*\*2)

a0 = z.mean() - a1 \* x.mean()

b = a1

a = np.exp(a0)

ax.plot(x3, a \* np.exp(b\*x3), label='y(x) = b1\*exp(b2\*x)')

plt.legend()

plt.show()

if \_\_name\_\_=='\_\_main\_\_':

print('Остаточная дисперсия y: ', (np.array(means\_y) - msr\_y(np.array(means\_x))).sum() / k)

print('Остаточная дисперсия x: ', (np.array(means\_x) - msr\_x(np.array(means\_y))).sum() / k)

print('D внутригр xy = {}'.format(D\_ingr\_xy))

print('D внутригр yx = {}'.format(D\_ingr\_yx))

print('D межгр xy = {}'.format(D\_betwgr\_xy))

print('D межгр yx = {}'.format(D\_betwgr\_yx))

print('D общая xy = {}'.format(D\_gen\_xy))

print('D общая yx = {}'.format(D\_gen\_yx))

print('n xy = {}'.format(n\_xy))

print('n yx = {}'.format(n\_yx))

print('a={}, b={}, c={}'.format(\*res))

print('b0={}, b1={}'.format(a, b))

**приложение Е**

**Программа для метода k-cредних**

**lab6.py**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import matplotlib.colors as colors

import random

from itertools import combinations

np.random.seed(10)

random.seed(11)

df = pd.read\_csv('sample.csv', header=None)

df.columns = ['x', 'y']

ax = df.plot.scatter(x=0, y=1)

ax.set\_title('Выборка')

plt.show()

df = (df - df.min(axis=0)) / (df.max(axis=0) - df.min(axis=0))

# df = (df - df.mean(axis=0)) / df.std(axis=0)

ax = df.plot.scatter(x=0, y=1)

ax.set\_title('Нормализованная выборка')

plt.show()

up\_limit = np.sqrt(len(df) / 2).astype(np.int64)

print('Верхняя граница: {}'.format(up\_limit))

def f1():

distances = df.apply(lambda x: np.min(dists\_to\_centroids(x, centroids)) \*\* 2, axis=1)

return distances.sum()

def get\_metrics():

f2 = []

f3 = []

for i, centroid in enumerate(centroids.to\_numpy()):

cluster\_dists = []

f3.append(df[cl\_centroids == i].var().mean())

for comb in combinations(df[cl\_centroids == i].to\_numpy(), 2):

cluster\_dists.append(np.linalg.norm(comb[0] - comb[1]) \*\* 2)

f2.append(sum(cluster\_dists))

f2 = sum(f2)

f3 = sum(f3)

print('----\nF1 = {}\nF2 = {}\nF3 = {}\n----'.format(f1(), f2, f3))

def dists\_to\_centroids(point, cur\_centroids):

return cur\_centroids.apply(lambda x: np.linalg.norm(x - point), axis=1)

def get\_closest\_centroids(points, cur\_centroids):

return points.apply(lambda x: np.argmin(dists\_to\_centroids(x, cur\_centroids)), axis=1)

def move\_centroids(points, closest\_centroids, num\_of\_centroids):

return np.array([points[closest\_centroids == c].mean(axis=0) for c in range(num\_of\_centroids)])

for N in range(2, up\_limit+1):

print(N)

dict\_colors = {i: name for i, (name, col) in enumerate(random.choices(list(colors.CSS4\_COLORS.items()), k=N))}

list\_colors = [name for name, col in random.choices(list(colors.CSS4\_COLORS.items()), k=N)]

centroids = df.sample(N)

i = 1

while True:

prev\_centroids = centroids.copy()

cl\_centroids = get\_closest\_centroids(df, centroids)

centroids[:] = move\_centroids(df, cl\_centroids, N)

ax = df.plot.scatter(x=0, y=1, c=cl\_centroids.apply(lambda x: dict\_colors[x]))

ax.scatter(centroids.x, centroids.y, c='red')

ax.set\_title('K means {}, шаг - {}'.format(N, i))

plt.show()

# print('step {}'.format(i))

i += 1

if ((prev\_centroids - centroids).mean(axis=0).abs() < [0.0001, 0.0001]).all():

break

cl\_centroids = get\_closest\_centroids(df, centroids)

get\_metrics()

rows = []

for i, centroid in enumerate(centroids.to\_numpy()):

rows.append([i + 1, '({} : {})'.format(\*centroid), sum(cl\_centroids == i)])

res = pd.DataFrame(rows, columns=['Номер кластера', 'Центр кластера', 'Количество элементов в кластере'])

res.to\_csv('Таблица{}.csv'.format(N), index=False)

**приложение Ж**

**Программа для метода поиска сгущений.**

**lab7.py**

import numpy as np

import sys

import math

import pandas as pd

import matplotlib.pyplot as plt

import matplotlib.colors as colors

import random

from itertools import combinations

from scipy.spatial import distance

random.seed(5)

np.random.seed(5)

df = pd.read\_csv('sample.csv', header=None)

df.columns = ['x', 'y']

ax = df.plot.scatter(x=0, y=1)

ax.set\_title('Выборка')

plt.show()

df = (df - df.mean(axis=0)) / df.std(axis=0)

noise = np.random.normal(0, 0.1, [len(df), 2])

# df = df + noise

ax = df.plot.scatter(x=0, y=1)

ax.set\_title('Нормализованная выборка')

xlim = ax.get\_xlim()

ylim = ax.get\_ylim()

plt.show()

distances = distance.cdist(df, df)

distances = set(distances.flatten().tolist()) - {0}

R\_min = min(distances)

R\_max = max(distances)

def plot\_step(step\_points, circle, title):

ax = step\_points.plot.scatter(x=0, y=1)

ax.set\_title(title)

ax.set\_xlim(xlim)

ax.set\_ylim(ylim)

c = plt.Circle(circle, R, alpha=0.1)

ax.add\_patch(c)

plt.show()

points = df.copy()

R = np.float32(1.2) # 0.8

i = 0

df['Clusters'] = 0

while len(points):

circle = points.iloc[:, :2].sample(1)

j = 0

while True:

prev\_circle = circle

points\_in\_circle = points.apply(lambda x: np.linalg.norm(x - circle) <= R, axis=1)

circle = points[points\_in\_circle].mean(axis=0)

plot\_step(points, circle, 'R = {}, Шаг {}.{}'.format(R.round(1), i + 1, j + 1))

j += 1

if ((circle - prev\_circle).abs() < 0.0001).all().all():

break

points\_in\_circle = points.apply(lambda x: np.linalg.norm(x - circle) <= R, axis=1)

df.loc[points\_in\_circle.index, 'Clusters'] = points\_in\_circle \* i

points = points[~ points\_in\_circle]

i += 1

list\_colors = np.array([name for name, col in colors.CSS4\_COLORS.items()])

np.random.shuffle(list\_colors)

ax = df.plot.scatter(x=0, y=1, c=list\_colors[df['Clusters']], s=50)

for c in range(i):

circle = df[df['Clusters'] == c].iloc[:, :2].mean(axis=0)

ax.scatter(circle[0], circle[1], c='red')

ax.set\_title('Результат кластеризации, R = {}'.format(R.round(1)))

plt.show()

f1 = []

f2 = []

f3 = []

for c in range(i):

if np.isnan(df[df['Clusters'] == c].iloc[:, :2].var().mean()):

continue

circle = df[df['Clusters'] == c].iloc[:, :2].mean(axis=0)

cluster\_dists = []

f3.append(df[df['Clusters'] == c].iloc[:, :2].var().mean())

for comb in combinations(df[df['Clusters'] == c].iloc[:, :2].to\_numpy(), 2):

cluster\_dists.append(np.linalg.norm(comb[0] - comb[1]) \*\* 2)

f2.append(sum(cluster\_dists))

f1.append(sum(np.linalg.norm(df[df['Clusters'] == c].iloc[:, :2] - circle, axis=1) \*\* 2))

f2 = sum(f2)

f3 = sum(f3)

f1 = sum(f1)

print('R = {}:\nF1 = {}\nF2 = {}\nF3 = {}'.format(R, f1, f2, f3))

rows = []

for centroid\_id in range(i):

centroid = df[df['Clusters'] == centroid\_id].iloc[:, :2]

rows.append([centroid\_id + 1, '({} : {})'.format(\*(centroid.mean(axis=0))), len(centroid)])

res = pd.DataFrame(rows, columns=['Номер кластера', 'Центр кластера', 'Количество элементов в кластере'])

res.to\_csv('Таблица.csv', index=False)