МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»

институт информационных технологий и технологического образования
кафедра информационных технологий и электронного обучения

Основная профессиональная образовательная программа
Направление подготовки 09.03.01 Информатика и вычислительная техника
Направленность (профиль) «Технологии разработки программного обеспечения»
форма обучения – очная

Зачётная лабораторная работа

по дисциплине «Анализ данных и основы Data Science»

Ряды распределения и математические характеристики

Выполнила обучающаяся 2 курса
Яблонская Евгения

Руководитель:
д.п.н, профессор
_____ Власова Е.З.

«_____» _____ 2023 г.

Санкт-Петербург
2023

Цель: собрать статистические данные, после чего провести их обработку и анализ.

Постановка задачи:

1. Осуществить поиск/сбор данных.
2. Составить ряд распределения и изобразить его графически.
3. Вычислить математические характеристики вариационного ряда.
4. Рассчитать теоретические частоты для нормального распределения.
5. Определить, является ли распределение нормальным, используя критерий Колмогорова.

Оборудование: ПК, табличный процессор Excel.

Математические модели:

Средняя арифметическая взвешенная

$$\overline{x} = \frac{\sum xm}{\sum m}$$

Среднеквадратичное отклонение

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2 \cdot m_i}{\sum m_i}}$$

Нормированное отклонение от средней

$$t_i = \frac{x_i - \overline{x}}{\sigma_x}$$

Значение функции q(t)

$$q(t) = \frac{1}{2\pi} \cdot \mathrm{e}^{\frac{-t^2}{2}}$$

Теоретические частоты

$$f_m = q(t)\frac{Nd}{\sigma}$$

Критерий

$$\lambda = \frac{D_{max}}{\sqrt{N}}$$

Коэффициент вариации

$$\mathrm{C} = \frac{\overline{\mathrm{x}}}{\sigma} \cdot 100\%$$

Коэффициент асимметрии

$$A_s = \frac{\overline{\mathrm{x}} - Mo}{\sigma}$$

Эксцесс

$$\mathrm{E} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4 \cdot m_i}{n \cdot \sigma^4} - 3$$

Все вычисления представлены в Приложении 1.

Результаты:

1. Данные для обработки представлены на Рисунке 1.

Условие: в школе собрали данные о полученных по ЕГЭ баллов у 80-ти учеников за настоящий год. Нужно провести обработку данных.

Данные			
93	39	96	68
3	67	98	16
73	18	14	73
66	90	83	6
54	31	74	50
93	65	19	48
37	63	78	92
43	23	8	93
34	42	77	21
66	43	35	84
98	36	10	81
87	87	87	88
49	49	81	24
35	79	95	66
91	9	64	51
23	76	28	97
78	61	61	65
3	88	42	4
73	97	97	73
26	2	91	90

Рисунок 1. Данные

2. Ряд распределения и его графическое изображение.

Был составлен интервальный вариационный ряд, представленный на Рисунке 2.

Интервальный вариационный ряд						
Вычисления			начала интервалов (включительно) или нижняя граница	конец интервалов или верхняя граница	частоты, mi	доля интервалов, wi
Минимум:			2	14	8	0,1
2			14	26	8	0,1
Максимум:			26	38	8	0,1
98			38	50	8	0,1
			50	62	5	0,0625
Количество интервалов по формуле Стёрджесса:			62	74	13	0,1625
7,178657555			74	86	11	0,1375
			86	98	19	0,2375
Следовательно, k =	8					
начало первого интервала =	2			Итого	80	1
конец последнего интервала =	98					
Длина каждого интервала:						
12						

Рисунок 2. Интервальный вариационный ряд

Графические изображение вариационного ряда представлены на Рисунках 3-6.

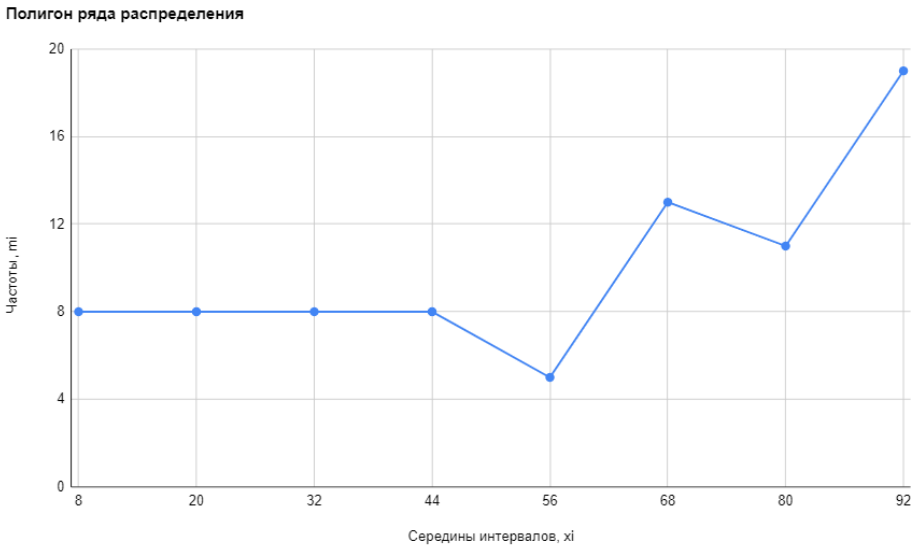

Рисунок 3. Полигон

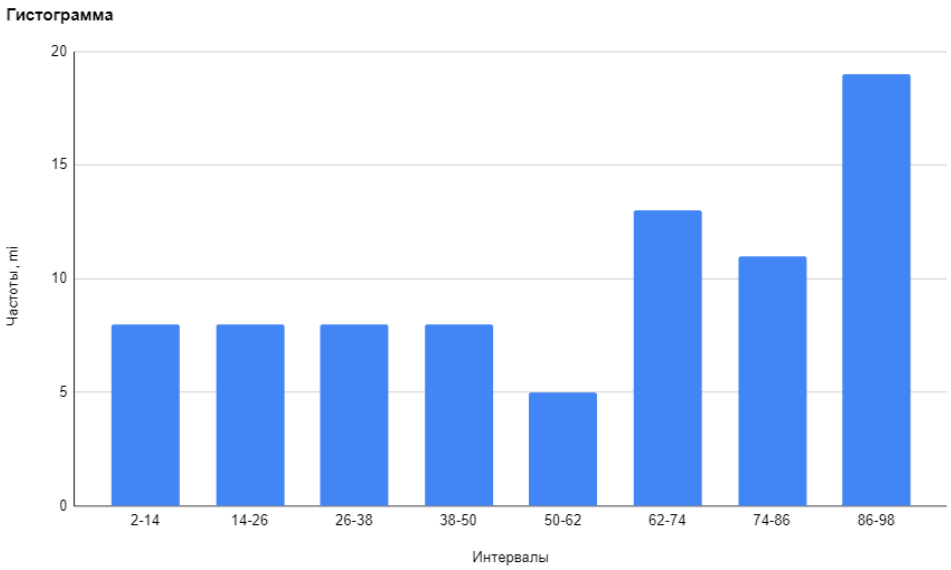

Рисунок 4. Гистограмма

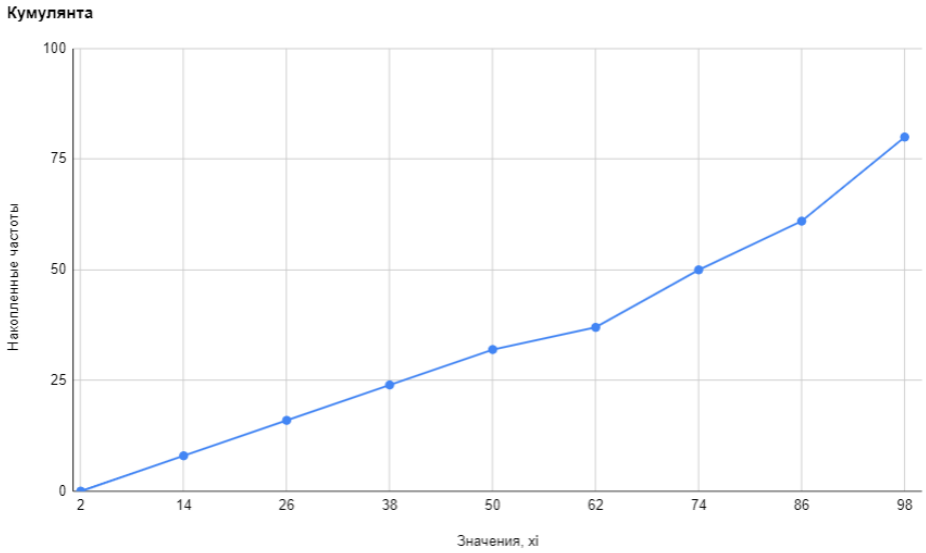

Рисунок 5. Кумулянта

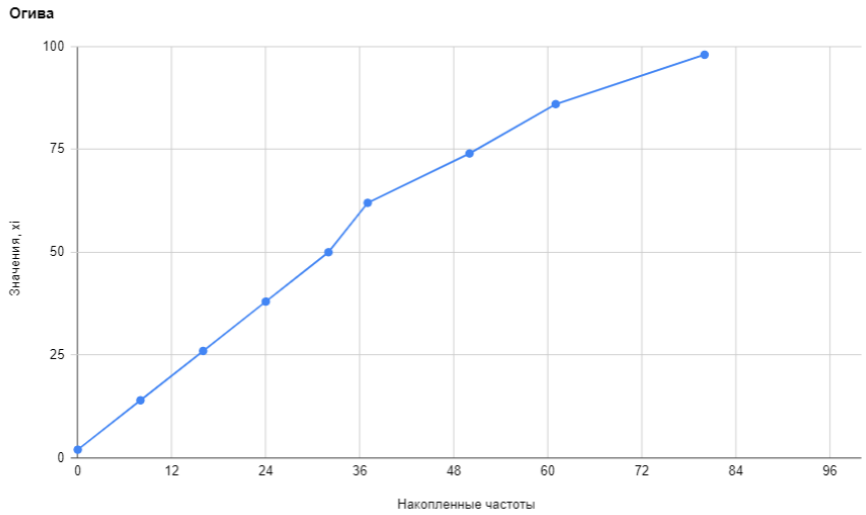

Рисунок 6. Огива

3. Вычисление математических характеристик вариационного ряда.

Математические характеристики построенного вариационного ряда представлены на Рисунках 7 и 8.

1. Среднее значение признака (n=80)
57,80

2. Дисперсия
839,160

2. Среднее квадратичное отклонение
28,968

3. Коэффициент вариации
50,12%

4. Учитывая минимальное и максимальное
Интервал:
[2; 98]

5. Коэффициент ассиметрии.
Мода:
73
Коэффициент ассиметрии:
-0,525

6. Эксцесс
μ4
1227223,48
Эксцесс E
-1,2573

7. Медиана
65

Рисунок 7. Математические характеристики

25-й перцентиль	20,25	25-ый перцентиль находится между числами на позициях 20 и 21 в упорядоченных по возрастанию данных. Это будут числа 34 и 35, следовательно, 25-ый перцентиль равен 34+(35-34)*0,25=34,25
50-й перцентиль	40,5	50-ый перцентиль находится между числами на позициях 40 и 41 в упорядоченных по возрастанию данных. Это будут числа 65 и 65, следовательно, 50-ый перцентиль - 65.
90-й перцентиль	72,9	90-ый находится между числами на позициях 72 и 73. Это будут числа 93 и 93, следовательно, 90-ый перцентиль - 93.

Рисунок 8. Перцентили

4. Расчёт теоретических частот представлен на Рисунке 9.

Интервалы баллов по ЕГЭ	Количество предприятий, fi	Середина интервала, xi	xi*fi	Среднее значение xi	((xi-x ср.знач.)^2) *fi	Среднее квадратичное отклонение σ	Нормированное отклонение от средней ti	Значение функции q(t)	Теоретические частоты, fm (d=12, N=80)
2-14	8	8	64	57,800	19840,3	28,968	-1,72	0,091	3
14-26	8	20	160		11430,7		-1,30	0,170	6
26-38	8	32	256		5325,1		-0,89	0,268	9
38-50	8	44	352		1523,5		-0,48	0,356	12
50-62	5	56	280		16,2		-0,06	0,398	13
62-74	13	68	884		1352,5		0,35	0,375	12
74-86	11	80	880		5421,2		0,77	0,297	10
86-98	19	92	1748		22223,2		1,18	0,199	7
ИТОГ	80	-	4624	-	67132,8	-	-	-	72

Рисунок 9. Расчёт теоретических частот

5. Вычисление критерий Колмогорова для данного вариационного ряда представлен на Рисунке 10.

Накопленные эмпирические частоты, Fi	Накопленные теоретические частоты, Fm	Di=|Fi-Fm|	Вычисления
8	3	5	Dmax
16	9	7	8
24	18	6	λ
32	30	2	0,89
37	43	6	
50	55	5	
61	65	4	
80	72	8	
-	-	-	

Рисунок 10. Критерий Колмогорова

Анализ:

По таблице с Рисунка 10 видно, что вычисленные теоретические частоты отличаются от эмпирических.

По таблице значений критерия Колмогорова вероятность того, что исследуемые данные имеют нормальный закон распределения: $P(\lambda)=P(0,89)=0,4067$. Таким образом, делаем вывод, что распределение исследуемых данных не происходит по нормальному закону.

Вычисленный коэффициент асимметрии Пирсона равен $As=-0,525$, то есть $As<0$ и As по модулю превосходит 0,5. Следовательно, имеется левосторонняя асимметрия, при том существенная.

Вывод:

В ходе лабораторной работе были проведены анализ и обработка собранных статистических данных.