

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

ОТЧЕТ
по лабораторной работе №2
по дисциплине «Статистические методы
обработки экспериментальных данных»
Тема: Обработка выборочных данных. Нахождение точечных оценок
параметров распределения.

Студент гр. 8383

Бабенко Н.С.

Студент гр. 8383

Сахаров В.М.

Преподаватель

Середа А.-В.И.

Санкт-Петербург

2022

Цель работы

Получение практических навыков нахождения точечных статистических оценок параметров распределения.

Основные теоретические положения

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

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

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

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

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

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

Выборочная дисперсия определяется по формуле:

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

Исправленная выборочная дисперсия определяется по формуле:

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

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

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

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

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

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

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

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

Постановка задачи

Для заданных выборочных данных вычислить с использованием метода моментов и условных вариант точечные статистические оценки математического ожидания, дисперсии, среднеквадратического отклонения, асимметрии и эксцесса исследуемой случайной величины. Полученные результаты содержательно проинтерпретировать.

Выполнение работы

Переменная π

Интервальный ряд из первой лабораторной работы для переменной π и с посчитанными накопленными частотами представлен в таблице 1.

Таблица 1

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

Объем выборки $n = 100$

Условные варианты можно найти с помощью формулы:

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

Условные моменты k -го порядка можно найти по формуле:

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

Результаты вычислений представлены в табл. 2.

Таблица 2

u	n	u	n * u	n * u²	n * u³	n * u⁴	n * (u + 1)⁴
343	0.04	-3	-0.12	0.36	-1.08	3.24	0.64
387	0.09	-2	-0.18	0.36	-0.72	1.44	0.09
431	0.27	-1	-0.27	0.27	-0.27	0.27	0.0
475	0.35	0	0.0	0.0	0.0	0.0	0.35
519	0.17	1	0.17	0.17	0.17	0.17	2.72
563	0.06	2	0.12	0.24	0.48	0.96	4.86
604	0.02	3	0.06	0.18	0.54	1.62	5.12
Σ	1	—	-0.22	1.58	-0.88	7.7	13.78

Проверить вычисления можно с помощью последнего столбца:

$$\begin{aligned} & \sum n_j * u_j^4 + 4 * \sum n_j * u_j^3 + 6 * \sum n_j * u_j^2 + 4 * \sum n_j * u_j + 1 = \\ & = 7.7 + 4 * -0.88 + 6 * 1.58 + 4 * -0.22 + 1 = 13.78 \end{aligned}$$

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

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

$$\bar{x}_B = \overline{M}_1 = \overline{M}_1^* h + C = 465.32$$

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

$$D_B = \overline{m}_2 = \left(\overline{M}_2^* - \left(\overline{M}_1^* \right)^2 \right) h^2 = 2965.1776$$

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

$$\bar{x}_B = \frac{1}{N} \sum_{i=1}^k x_i n_i = 465.26$$

$$D_B = \frac{1}{N} \sum_{i=1}^k (x_i - \bar{x}_B)^2 n_i = 2968.35$$

Исправленная оценка дисперсии:

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

Были найдены статистические оценки СКО:

$$\sigma_B = \sqrt{D_B} = \sqrt{2968.35} = 54.3$$

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

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

Были найдены статистические оценки коэффициентов асимметрии и эксцесса:

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

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

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

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

Статистическая оценка коэффициента асимметрии:

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

Статистическая оценка коэффициента эксцесса:

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

Коэффициент асимметрии положительный, следовательно, это правосторонняя асимметрия, и $\bar{x}_B > M_o$, но полученный коэффициент незначительный и

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

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

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

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

Медиана заданного распределения:

$$M_e = x_0 + \frac{0.5n - n_{m-1}^n}{n_m} h,$$

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

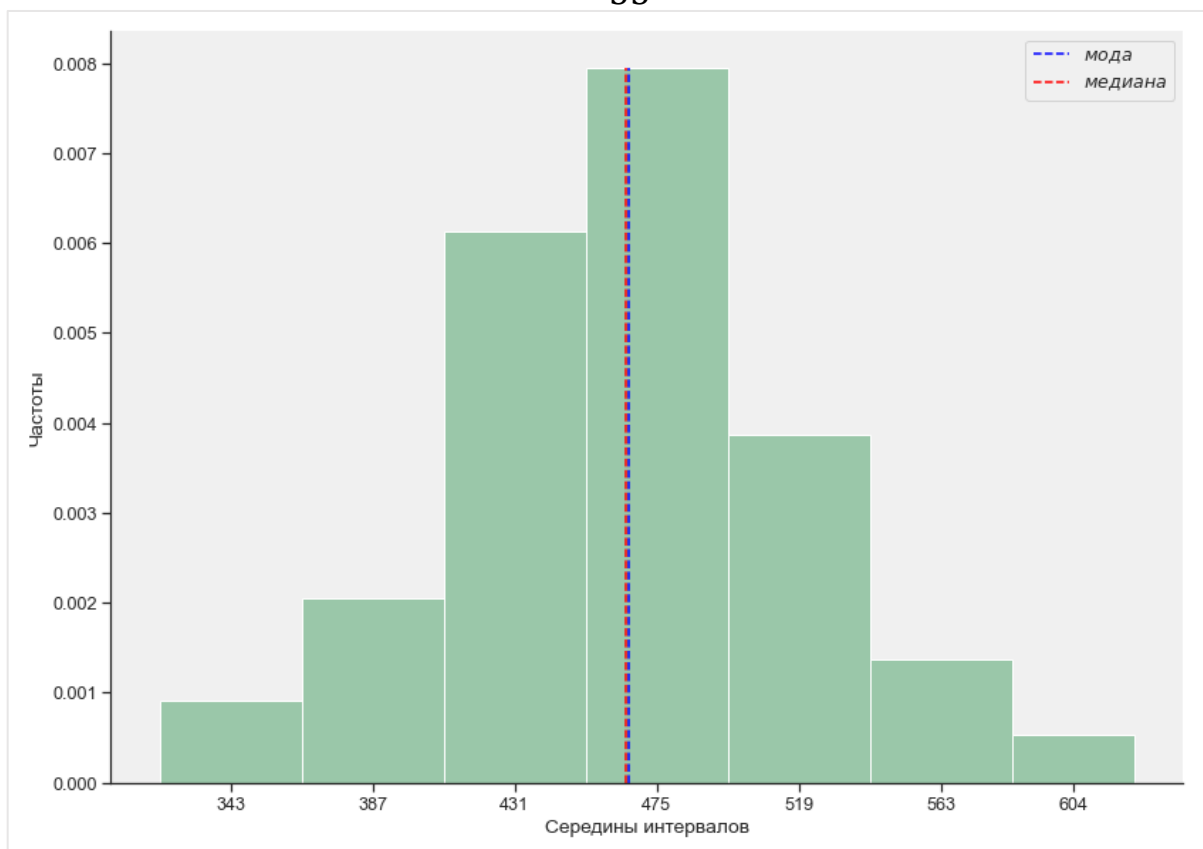


Рисунок 1 – Гистограмма относительных частот

На рисунке можно увидеть, что мода смещена относительно центра модального интервала в сторону левого интервала с большей частотой. Медиана

также смещена левее, так как по левую сторону находится большее количество вариантов.

Переменная E

Интервальный ряд из первой лабораторной работы для переменной E и с посчитанными накопленными частотами представлен в таблице 3.

Таблица 3

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

Результаты вычислений условных моментов представлены в табл. 4.

Таблица 4

v	n	u	n * u	n * u²	n * u³	n * u⁴	n * (u + 1)⁴
92	0.06	-3	-0.18	0.54	-1.62	4.86	0.96
108	0.14	-2	-0.28	0.56	-1.12	2.24	0.14
124	0.32	-1	-0.32	0.32	-0.32	0.32	0.0
140	0.33	0	0.0	0.0	0.0	0.0	0.33
156	0.09	1	0.09	0.09	0.09	0.09	1.44
172	0.04	2	0.08	0.16	0.32	0.64	3.24
188	0.02	3	0.06	0.18	0.54	1.62	5.12
Σ	1	—	-0.55	1.85	-2.11	9.77	11.23

Проверим вычисления с помощью последнего столбца:

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

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

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

$$\bar{x}_B = \overline{M_1} = \overline{M_1^*}h + C = 133.7$$

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

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

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

$$\bar{x}_B = \frac{1}{N} \sum_{i=1}^k x_i n_i = 133.8$$

$$D_B = \frac{1}{N} \sum_{i=1}^k (x_i - \bar{x}_B)^2 n_i = 396.16$$

Исправленная оценка дисперсии:

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

Были найдены статистические оценки СКО:

$$\sigma_B = \sqrt{D_B} = \sqrt{396.16} = 19.9$$

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

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

Были найдены статистические оценки коэффициентов асимметрии и эксцесса:

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

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

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

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

Статистическая оценка коэффициента асимметрии:

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

Статистическая оценка коэффициента эксцесса:

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

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

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

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

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

Медиана заданного распределения:

$$M_e = x_0 + \frac{0.5n - n_{m-1}^n}{n_m} h,$$

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

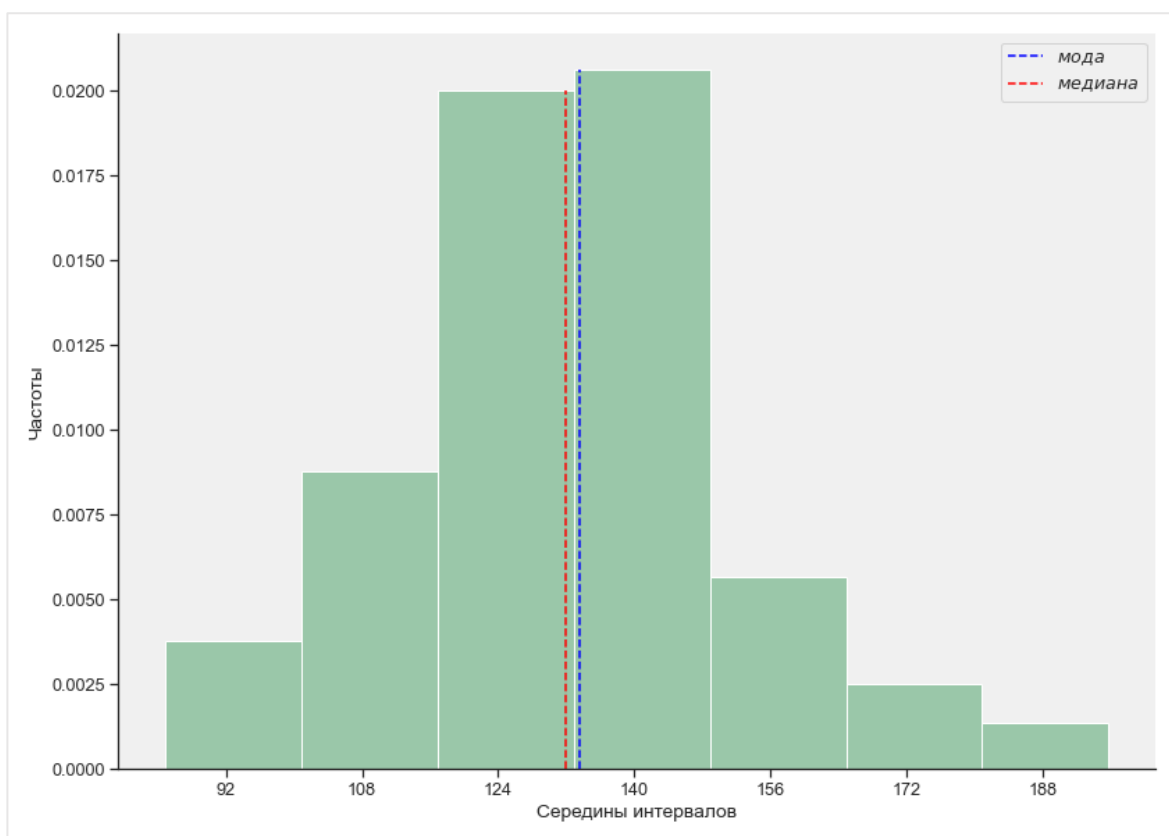


Рисунок 2 – Гистограмма относительных частот для переменной Е

На рисунке можно увидеть, что мода смещена относительно центра модального интервала в сторону левого интервала с большей частотой. Медиана же смещена правее, так как по правую сторону находится большее количество вариантов.

Выводы

В ходе выполнения данной лабораторной работы были получены практические навыки нахождения точечных статистических оценок параметров распределения. Для интервального ряда для обеих переменных из лабораторной работы №1 были вычислены условные эмпирические моменты через условные варианты. Была проведена корректность вычислений через контрольную сумму, вычисления оказались верны для обеих переменных. Были посчитаны выборочное среднее и дисперсия с помощью стандартных формул и с помощью условных вариантов. Статистические оценки, вычисленные по стандартным формулам и с помощью условных вариантов совпали для обеих переменных.

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

Для интервальных рядов была вычислена мода и медиана. Мода для обеих переменных оказалась смещена относительно центра модального интервала в сторону левого интервала с большей частотой. Для переменной *пи*: медиана также смещена левее, так как по левую сторону находится большее количество вариантов, для переменной *Е*: медиана смещена правее, так как уже по правую сторону находится большее количество вариантов.

ПРИЛОЖЕНИЕ А

ИСХОДНЫЙ КОД

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
original = pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/lab1/data/data2.csv')
var_row = pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/lab1/data/data4.csv')
var_row.to_csv('data/var_row.csv', index=False)
n = 100
h = 44
int_row = pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/data/interval.csv')
int_row['cum_sum'] = np.round(np.cumsum(int_row['rf']), 3)
int_row.to_csv('data/int_row.csv', index=False)
usl_mom = int_row.copy()
usl_mom = usl_mom.iloc[:, [1,3]]
usl_mom['u'] = np.arange(-3,4,1)
usl_mom['nu'] = usl_mom['rf']*usl_mom['u']
usl_mom['nu2'] = usl_mom['rf']*pow(usl_mom['u'], 2)
usl_mom['nu3'] = usl_mom['rf']*pow(usl_mom['u'], 3)
usl_mom['nu4'] = usl_mom['rf']*pow(usl_mom['u'], 4)
usl_mom['nu4+'] = usl_mom['rf']*pow(usl_mom['u']+1, 4)
usl_mom
usl_mom_f = usl_mom.append(np.round(usl_mom.sum(), 3), ignore_index=True)
usl_mom_f.to_csv('data/usl_mom.csv', index=False)
moms = usl_mom_f.iloc[7, [3,4,5,6]]
checker = moms[3]+4*moms[2]+6*moms[1]+4*moms[0]+1
'True' if checker == usl_mom_f.loc[7, ['nu4+']][0] else 'False'
checker
M1 = moms[0]*h+475
m2 = (moms[1] - pow(moms[0],2))*pow(h,2)
m3 = (moms[2] - 3*moms[1]*moms[0] + 2*pow(moms[0],3))*pow(h,3)
m4 = (moms[3] - 4*moms[2]*moms[0] + 6*moms[1]*pow(moms[0],2) - 3*pow(moms[0],4))*pow(h,4)
(M1, m2, m3, m4)
int_mean = (int_row['avg_inter']*int_row['af']).sum()/n
int_var = (((int_row['avg_inter']-int_mean)**2)*int_row['af']).sum()/n
s = int_var*(n/(n-1))
std_s = np.sqrt(s)
std_var = np.sqrt(int_var)
```

```

int_mean
int_var
s
std_s
std_var
As = m3/(pow(s, 3))
Ex = (m4/(pow(s, 4))) - 3
As, Ex
raw_mode = 453+h*(8/26)
raw_median = 453+(((0.5*n)-40)/35)*h
raw_mode
raw_median
int_mean
sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style('ticks', {"axes.facecolor": ".94"})
ax = sns.displot(data=original, x='nu', bins=np.array([321, 365, 409, 453,
497, 541, 585, 623]),
                kind='hist', height=8.27, aspect=11.7/8.27, stat='densi-
ty')
plt.vlines(raw_mode, 0, int_row.loc[3, 'rf']/h, colors='b', linestyle='-
-', label='$мода$')
plt.vlines(raw_median, 0, int_row.loc[3, 'rf']/h, colors='r', lin-
estyle='--', label='$медиана$')
# plt.vlines(int_mean, 0, int_row.loc[3, 'rf']/h, colors='k', lin-
estyle='--', label='$x_в$')
ax.set_axis_labels('Средины интервалов', 'Частоты')
ax.set(xticks=int_row['avg_inter'])
plt.legend()
plt.savefig('pics/1.png')
original = pd.read_csv('c:/Us-
ers/gandh/dev/unv/smoed/NB/lab1/data2/data2.csv')
var_row = pd.read_csv('c:/Us-
ers/gandh/dev/unv/smoed/NB/lab1/data2/data4.csv')
var_row.to_csv('data/var_row2.csv', index=False)
n = 100
h = 16
int_row = pd.read_csv('c:/Users/gandh/dev/unv/smoed/NB/data/inter-
val2.csv')
int_row['cum_sum'] = np.round(np.cumsum(int_row['rf']), 3)
int_row.to_csv('data/int_row2.csv', index=False)
usl_mom = int_row.copy()
usl_mom = usl_mom.iloc[:, [1,3]]
usl_mom['u'] = np.arange(-3,4,1)
usl_mom['nu'] = usl_mom['rf']*usl_mom['u']

```

```

usl_mom['nu2'] = usl_mom['rf']*pow(usl_mom['u'], 2)
usl_mom['nu3'] = usl_mom['rf']*pow(usl_mom['u'], 3)
usl_mom['nu4'] = usl_mom['rf']*pow(usl_mom['u'], 4)
usl_mom['nu4+'] = usl_mom['rf']*pow(usl_mom['u']+1, 4)
usl_mom
usl_mom_f = usl_mom.append(np.round(usl_mom.sum(), 3), ignore_index=True)
usl_mom_f.to_csv('data/usl_mom2.csv', index=False)
usl_mom_f
moms = usl_mom_f.iloc[7, [3,4,5,6]]
checker = moms[3]+4*moms[2]+6*moms[1]+4*moms[0]+1
'True' if checker == usl_mom_f.loc[7, ['nu4+']][0] else 'False'
checker
M1 = moms[0]*h+140
m2 = (moms[1] - pow(moms[0],2))*pow(h,2)
m3 = (moms[2] - 3*moms[1]*moms[0] + 2*pow(moms[0],3))*pow(h,3)
m4 = (moms[3] - 4*moms[2]*moms[0] + 6*moms[1]*pow(moms[0],2) -
3*pow(moms[0],4))*pow(h,4)
M1, m2, m3, m4
int_mean = (int_row['avg_inter']*int_row['af']).sum()/n
int_var = (((int_row['avg_inter']-int_mean)**2)*int_row['af']).sum()/n
s = int_var*(n/(n-1))
std_s = np.sqrt(s)
std_var = np.sqrt(int_var)
int_mean
int_var
s
std_s
std_var
As = m3/(pow(s, 3))
Ex = (m4/(pow(s, 4))) - 3
As, Ex
original.mean()
raw_mode = 132.9+h*(1/25)
raw_median = 116.9+(((0.5*n)-20)/32)*h
raw_mode
raw_median
int_mean
sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style('ticks', {"axes.facecolor": ".94"})
ax = sns.displot(data=original, x='E', bins=np.array([84.9, 100.9, 116.9,
132.9, 148.9, 164.9, 180.9, 195.7]),
kind='hist', height=8.27, aspect=11.7/8.27, stat='densi-
ty')

```

```

plt.vlines(raw_mode, 0, int_row.loc[3, 'rf']/h, colors='b', linestyle='--', label='$мода$')
plt.vlines(raw_median, 0, int_row.loc[2, 'rf']/h, colors='r', linestyle='--', label='$медиана$')
# plt.vlines(int_mean, 0, int_row.loc[2, 'rf']/h, colors='k', linestyle='--', label='$x_v$')
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=int_row['avg_inter'])
plt.legend()
plt.savefig('pics/2.png')

```