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	____________________	Киреев К.А.
Студент гр. 8383	____________________	Муковский Д.В.
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:

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

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

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

Выборочная дисперсия определяется по формуле:

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

Исправленная выборочная дисперсия определяется по формуле:

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

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

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

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

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

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

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

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

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

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

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

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

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

Для интервального ряда, полученного в лабораторной работе №1 были найдены середины интервалов и накопленные частоты. Интервальный ряд представлен в таблице 1.

Таблица 1

Границы интервалов	***Середины интервалов***	***Абсолютная частота***	***Относительная частота***	***Накопленная частота***
[320, 357)	338.5	5	0.048	0.048
[357, 394)	375.5	8	0.077	0.125
[394, 431)	412.5	23	0.221	0.346
[431, 468)	449.5	25	0.240	0.586
[468, 505)	486.5	24	0.231	0.817
[505, 542)	523.5	15	0.144	0.961
[542, 576)	559	4	0.039	1

Объем выборки $n = 104$

Количество интервалов $k = 1 + 3.31 * \lg n = 7$

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

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

Условные моменты k-го порядка:

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

Результаты вычислений представлены в табл. 2.

Таблица 2

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

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

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

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

Эмпирические начальные и центральные моменты вычислены ниже:

$$\overline{x_{\text{в}}} = \overline{M_1} = \overline{M_1^*}h + C = 453.829$$

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

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

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

Найдем выборочное среднее и дисперсию с помощью стандартных формул.

Статистическая оценка математического ожидания:

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

Статистическая оценка дисперсии:

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

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

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

Статистические оценки СКО:

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

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

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

Статистические оценки коэффициентов асимметрии и эксцесса можно вычислить по формулам:

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

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

Центральные эмпирические моменты третьего и четвертого порядков были найдены выше.

Статистическая оценка коэффициента асимметрии:

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

Статистическая оценка коэффициента эксцесса:

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

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

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

Вычислим моду и медиану заданного распределения для интервального ряда.

Мода заданного распределения:

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

где n_m – частота модального интервала,

n_{m-1} – частота предыдущего интервала,

n_{m+1} – частота следующего интервала,

x_0 – нижняя граница модального интервала

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

Медиана заданного распределения:

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

где n_m − частота медианного интервала,

n_{m-1}^n − накопленная частота предыдущего интервала,

x_0 − нижняя граница медианного интервала

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

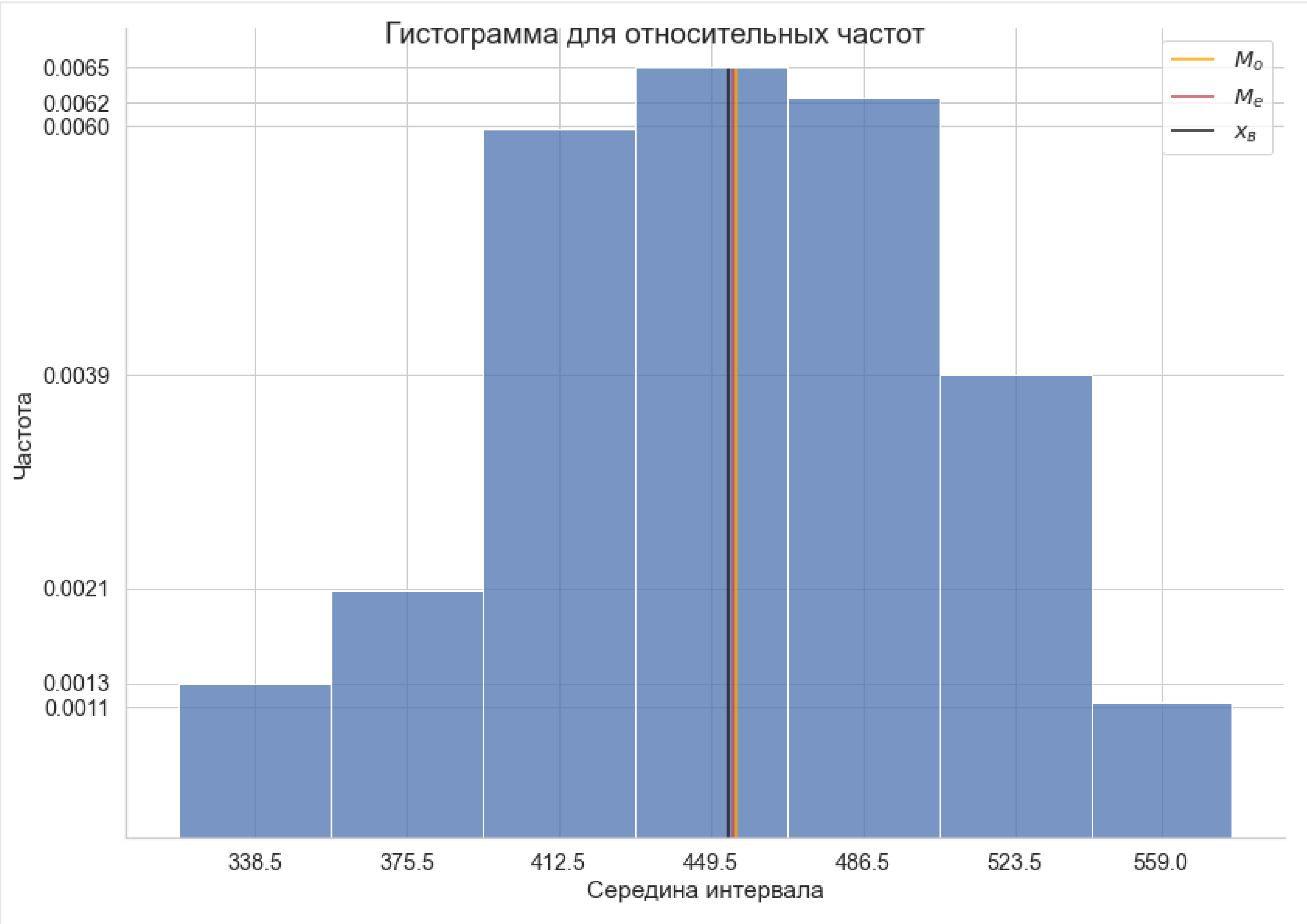

Рисунок 1 – Гистограмма относительных частот с отмеченными модой, медианой и выборочным средним

Видно, что мода смещена относительно центра модального интервала в сторону правого интервала с большей частотой. Медиана также смещена правее, так как по правую сторону находится большее количество вариант.

В целом можно заметить, что медиана, мода и выборочное среднее примерно равны, поэтому можно предположить, что анализируемая переменная имеет примерно нормальное распределение.

Выводы

В ходе выполнения данной лабораторной работы были получены практические навыки нахождения точечных статистических оценок параметров распределения.

Для интервального ряда из лабораторной работы №1 были найдены середины интервалов и накопленные частоты, далее для полученных вариант были вычислены условные варианты. Были вычислены условные эмпирические моменты через условные варианты, и с их помощью вычислены начальные и центральные эмпирические моменты. Корректность вычислений была проверена контрольной суммой, которая дала понять, что вычисления были верны.

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

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

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

ПРИЛОЖЕНИЕ А
ИСХОДНЫЙ КОД

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

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

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

# %%
usl_mom = int_row.copy()
usl_mom = usl_mom.iloc[:, [1,3]]
usl_mom['u'] = np.arange(-3,4,1)
usl_mom['nu'] = usl_mom['rf']*usl_mom['u']
usl_mom['nu2'] = usl_mom['rf']*pow(usl_mom['u'], 2)
usl_mom['nu3'] = usl_mom['rf']*pow(usl_mom['u'], 3)
usl_mom['nu4'] = usl_mom['rf']*pow(usl_mom['u'], 4)
usl_mom['nu4+'] = usl_mom['rf']*pow(usl_mom['u']+1, 4)

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

# %%
```

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

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

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

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

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

# %%
As, Ex

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

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

# %%
```

```
original.head()

# %%
sns.set_theme(style="whitegrid", palette='deep', context='notebook', font_scale=1.3)
ax = sns.displot(data=original, x='nu', bins=np.array([320, 357, 394, 431, 468, 505, 542, 576]),
                 kind='hist', height=8.27, aspect=11.7/8.27, stat='density')
plt.vlines(raw_mode, 0, int_row.loc[3, 'rf']/h, colors='orange', linestyles='solid', label='$M_o$')
plt.vlines(raw_median, 0, int_row.loc[3, 'rf']/h, colors='r', linestyles='solid', label='$M_e$')
plt.vlines(int_mean, 0, int_row.loc[3, 'rf']/h, colors='k', linestyles='solid', label='$x_в$')
ax.set_axis_labels('Середина интервала', 'Частота')
ax.set(xticks=int_row['avg_inter'], yticks=round((int_row['rf']/h), 4))
ax.fig.suptitle('Гистограмма для относительных частот')
plt.legend()
plt.savefig('pics/1.png')

# %%
```