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

ОТЧЕТ

по лабораторной работе №1

по дисциплине «Статистические методы обработки экспериментальных данных»

Тема: Формирование и первичная обработка выборки. Ранжированный и интервальный ряды.

Студентка гр. 8382	____________________	Звегинцева Е.Н.
Студент гр. 8382	____________________	Мирончик П.Д.
Преподаватель	____________________	Середа В.И.

Санкт-Петербург
2022

Цель работы.

Ознакомление с основными правилами формирования выборки и подготовки выборочных данных к статистическому анализу.

Основные теоретические положения.

Ранжированный ряд – это распределение отдельных единиц совокупности в порядке возрастания или убывания исследуемого признака. Ранжирование позволяет легко разделить количественные данные по группам, сразу обнаружить наименьшее и наибольшее значения признака, выделить значения, которые чаще всего повторяются.

Вариационный ряд – последовательность значений заданной выборки $x^m = (x_1, \ldots, x_m)$, расположенных в порядке неубывания:

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

Интервальный ряд распределения – это таблица, состоящая из двух столбцов (строк) – интервалов варьирующего признака X_i и числа единиц совокупности, попадающих в данный интервал (частот - f_i), или долей этого числа в общей численности совокупностей (частостей - d_i).

Полигоном частот называют ломанную, отрезки которой соединяют точки $(x_1, n_1), (x_2, n_2), \ldots, (x_k, n_k)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им частоты n_i. Точки (x_i, n_i) соединяют отрезками прямых и получают полигон частот.

Гистограммой частот (частостей) называется ступенчатая фигура, состоящая из прямоугольников с основаниями, равными интервалам значений h_i и высотами, равными отношению частот (или частостей) к шагу:

$$\frac{m_i}{h_i}\left(\frac{\omega_i}{h_i} = \frac{m_i}{n * h_i}\right)$$

Эмпирической функцией распределения величины называют функцию,

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{Xi \leq x\}},$$ где

Xi – i-ый элемент выборки. Т.е. значение функции в точке x равно числу элементов выборки, не превосходящих x, деленному на объем выборки. В данной работе эмпирическая функция строится для интервального ряда абсолютных и относительных частот, поэтому вместо отдельных элементов выборки в формулу подставляются сформированные интервалы, а в случае относительных частот не осуществляется деление на N.

Постановка задачи.

Осуществить формирование репрезентативной выборки заданного объема из имеющейся генеральной совокупности экспериментальных данных. Осуществить последовательное преобразование полученной выборки в ранжированный, вариационный и интервальный ряды. Применительно к интервальному ряду построить и отобразить графически полигон, гистограмму и эмпирическую функцию распределения для абсолютных и относительных частот. Полученные результаты содержательно проинтерпретировать.

Выполнение работы.

Генеральная совокупность состоит из данных наблюдений относительно объемного веса $пи$ $(\frac{г}{см^3})$ при влажности 10% и модуля упругости E $(\frac{кг}{см^2})$ при сжатии вдоль волокон древесины резонансной ели, представленных преподавателем. Из нее была сгенерирована выборка из 114 наблюдений.

Полученная исходная выборка представлена в табл.1

Таблица 1

№	nu	E	№	nu	E	№	nu	E	№	nu	E	№	nu	E
1	405	107.5	24	418	125.7	47	528	163.4	70	476	143.0	93	391	98.2
2	521	154.9	25	477	139.7	48	462	135.7	71	464	143.2	94	482	150.1
3	382	98.1	26	468	142.0	49	453	119.5	72	449	124.5	95	453	126.4
4	362	111.7	27	545	145.3	50	393	103.2	73	331	84.6	96	465	140.9
5	444	130.0	28	441	140.8	51	487	146.0	74	444	121.4	97	458	121.7
6	465	114.8	29	543	155.4	52	430	95.3	75	423	104.1	98	482	139.9
7	513	163.6	30	474	132.5	53	371	91.9	76	446	130.3	99	448	125.6
8	362	84.3	31	321	86.1	54	441	126.1	77	451	128.6	100	411	112.9
9	550	147.9	32	438	120.7	55	495	150.9	78	434	140.4	101	452	119.7
10	453	138.2	33	409	121.0	56	510	129.4	79	469	131.5	102	504	143.8
11	482	141.2	34	451	124.3	57	496	141.7	80	452	131.0	103	454	131.1
12	431	125.0	35	462	125.2	58	525	162.1	81	397	108.6	104	422	122.9
13	437	129.2	36	541	162.3	59	377	96.0	82	470	124.0	105	443	137.4
14	352	87.7	37	460	140.7	60	512	169.9	83	351	102.9	106	510	153.9
15	351	89.0	38	358	98.3	61	382	113.9	84	430	104.3	107	448	125.0
16	498	164.0	39	353	98.0	62	480	153.9	85	541	146.8	108	442	123.4
17	510	162.3	40	508	159.0	63	391	107.5	86	467	140.5	109	423	115.9
18	547	154.7	41	379	94.6	64	448	137.3	87	394	112.1	110	504	145.3
19	477	146.0	42	458	128.0	65	540	156.7	88	406	110.1	111	471	143.9
20	428	131.6	43	548	162.3	66	421	126.9	89	477	135.8	112	493	154.5
21	458	124.4	44	447	117.5	67	453	124.2	90	424	119.0	113	544	166.7
22	470	146.7	45	446	128.4	68	441	122.8	91	475	143.6	114	460	122.4
23	499	144.5	46	483	130.3	69	452	116.1	92	352	84.9			

Далее выборка была сокращена до величины nu и была упорядочена по возрастанию для получения ранжированного ряда. Результат представлен на рис.1

```
Исходный ряд:
 [405 521 382 362 444 465 513 362 550 453 482 431 437 352 351 498 510 547
 477 428 458 470 499 418 477 468 545 441 543 474 321 438 409 451 462 541
 460 358 353 508 379 458 548 447 446 483 528 462 453 393 487 430 371 441
 495 510 496 525 377 512 382 480 391 448 540 421 453 441 452 476 464 449
 331 444 423 446 451 434 469 452 397 470 351 430 541 467 394 406 477 424
 475 352 391 482 453 465 458 482 448 411 452 504 454 422 443 510 448 442
 423 504 471 493 544 460]
Ранжированный ряд:
 [321 331 351 351 352 352 353 358 362 362 371 377 379 382 382 391 391 393
 394 397 405 406 409 411 418 421 422 423 423 424 428 430 430 431 434 437
 438 441 441 441 442 443 444 444 446 446 447 448 448 448 449 451 451 452
 452 452 453 453 453 453 454 458 458 458 460 460 462 462 464 465 465 467
 468 469 470 470 471 474 475 476 477 477 477 480 482 482 482 483 487 493
 495 496 498 499 504 504 508 510 510 510 512 513 521 525 528 540 541 541
 543 544 545 547 548 550]
```

Рисунок 1 –выполнение формирования ранжированного ряда из исходного программой.

Отсюда можно сделать вывод, что наименьшее значение в выборке $x_{min} = 321$, а наибольшее значение $x_{max} = 550$.

Преобразование полученной выборки в вариационный ряд с абсолютными частотами, путем подсчета уникальных значений для нашей величины, представлено на рис.2. Также были рассчитаны относительные частоты, как доля значений от общего числа наблюдений, что можно наблюдать на рис.3

```
Вариационный ряд - абсолютные частоты:
 [(321, 1) (331, 1) (351, 2) (352, 2) (353, 1) (358, 1) (362, 2) (371, 1)
 (377, 1) (379, 1) (382, 2) (391, 2) (393, 1) (394, 1) (397, 1) (405, 1)
 (406, 1) (409, 1) (411, 1) (418, 1) (421, 1) (422, 1) (423, 2) (424, 1)
 (428, 1) (430, 2) (431, 1) (434, 1) (437, 1) (438, 1) (441, 3) (442, 1)
 (443, 1) (444, 2) (446, 2) (447, 1) (448, 3) (449, 1) (451, 2) (452, 3)
 (453, 4) (454, 1) (458, 3) (460, 2) (462, 2) (464, 1) (465, 2) (467, 1)
 (468, 1) (469, 1) (470, 2) (471, 1) (474, 1) (475, 1) (476, 1) (477, 3)
 (480, 1) (482, 3) (483, 1) (487, 1) (493, 1) (495, 1) (496, 1) (498, 1)
 (499, 1) (504, 2) (508, 1) (510, 3) (512, 1) (513, 1) (521, 1) (525, 1)
 (528, 1) (540, 1) (541, 2) (543, 1) (544, 1) (545, 1) (547, 1) (548, 1)
 (550, 1)]
```

Рисунок 2 – Вариационный ряд с абсолютными частотами

```
Вариационный ряд - относительные частоты:
 [(321, 0.00877193) (331, 0.00877193) (351, 0.01754386) (352, 0.01754386)
 (353, 0.00877193) (358, 0.00877193) (362, 0.01754386) (371, 0.00877193)
 (377, 0.00877193) (379, 0.00877193) (382, 0.01754386) (391, 0.01754386)
 (393, 0.00877193) (394, 0.00877193) (397, 0.00877193) (405, 0.00877193)
 (406, 0.00877193) (409, 0.00877193) (411, 0.00877193) (418, 0.00877193)
 (421, 0.00877193) (422, 0.00877193) (423, 0.01754386) (424, 0.00877193)
 (428, 0.00877193) (430, 0.01754386) (431, 0.00877193) (434, 0.00877193)
 (437, 0.00877193) (438, 0.00877193) (441, 0.02631579) (442, 0.00877193)
 (443, 0.00877193) (444, 0.01754386) (446, 0.01754386) (447, 0.00877193)
 (448, 0.02631579) (449, 0.00877193) (451, 0.01754386) (452, 0.02631579)
 (453, 0.03508772) (454, 0.00877193) (458, 0.02631579) (460, 0.01754386)
 (462, 0.01754386) (464, 0.00877193) (465, 0.01754386) (467, 0.00877193)
 (468, 0.00877193) (469, 0.00877193) (470, 0.01754386) (471, 0.00877193)
 (474, 0.00877193) (475, 0.00877193) (476, 0.00877193) (477, 0.02631579)
 (480, 0.00877193) (482, 0.02631579) (483, 0.00877193) (487, 0.00877193)
 (493, 0.00877193) (495, 0.00877193) (496, 0.00877193) (498, 0.00877193)
 (499, 0.00877193) (504, 0.01754386) (508, 0.00877193) (510, 0.02631579)
 (512, 0.00877193) (513, 0.00877193) (521, 0.00877193) (525, 0.00877193)
 (528, 0.00877193) (540, 0.00877193) (541, 0.01754386) (543, 0.00877193)
 (544, 0.00877193) (545, 0.00877193) (547, 0.00877193) (548, 0.00877193)
 (550, 0.00877193)]
```

Рисунок 3 – Вариационный ряд с относительными частотами

Наиболее часто встречающаяся величина - $x \quad = 453$

Далее был сформирован интервальный ряд. Были рассчитаны коэффициент k по формуле Стерджесса для определения количества интервалов и ширина интервала h.

$$k = 1. + 3.322 * \log(n), \text{ где } n \text{ – объем выборки}$$

$$h = \frac{x_{max} - x_{min}}{k}$$

Аналогично предыдущему пункту для каждого интервала были вычислены абсолютные и относительные частоты. Полученные значения можно увидеть на рис.4

```
Интервальный ряд: k=7.833037916139762 (формула Стерджесса), h=29.23514509334261=29 (шаг)
Интервальный ряд:
Интервал        Абс. частоты     Отн. частоты
[321, 350)      2                0.01754
[350, 379)      10               0.08772
[379, 408)      10               0.08772
[408, 437)      13               0.11404
[437, 466)      36               0.31579
[466, 495)      19               0.16667
[495, 524)      13               0.11404
[524, 553)      11               0.09649
```

Рисунок 4 – Интервальный ряд, вычисленный программой

Для проверки достоверности вычислений, можно сложить все абсолютные частоты, получив 114, что соответствует объему выборки, а при сложении относительных частот получим 1. Наибольшее количество наблюдений мы наблюдаем в интервале [437, 466)

Строим полигоны, представленные на рис.5 и 6, применительно к интервальному ряду для абсолютных и относительных частот раздельно.

Рисунок 5 – Полигон для абсолютной частоты

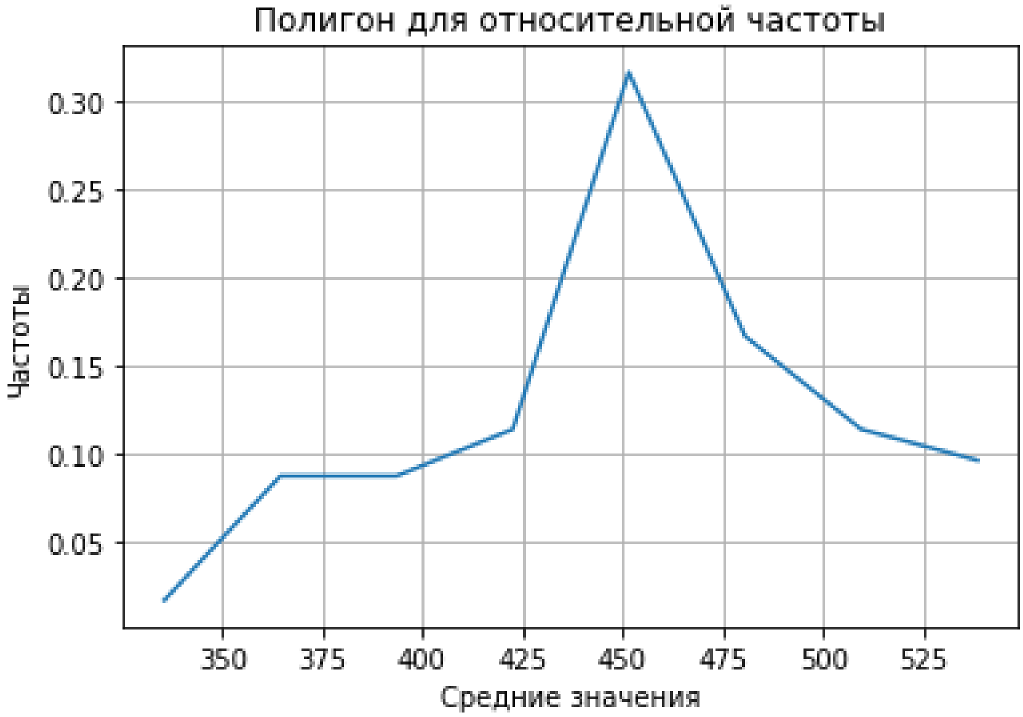

Рисунок 6 – Полигон для относительной частоты

Гистограммы, построенные применительно к интервальному ряду для абсолютных и относительных частот представлены на рис. 7 и 8.

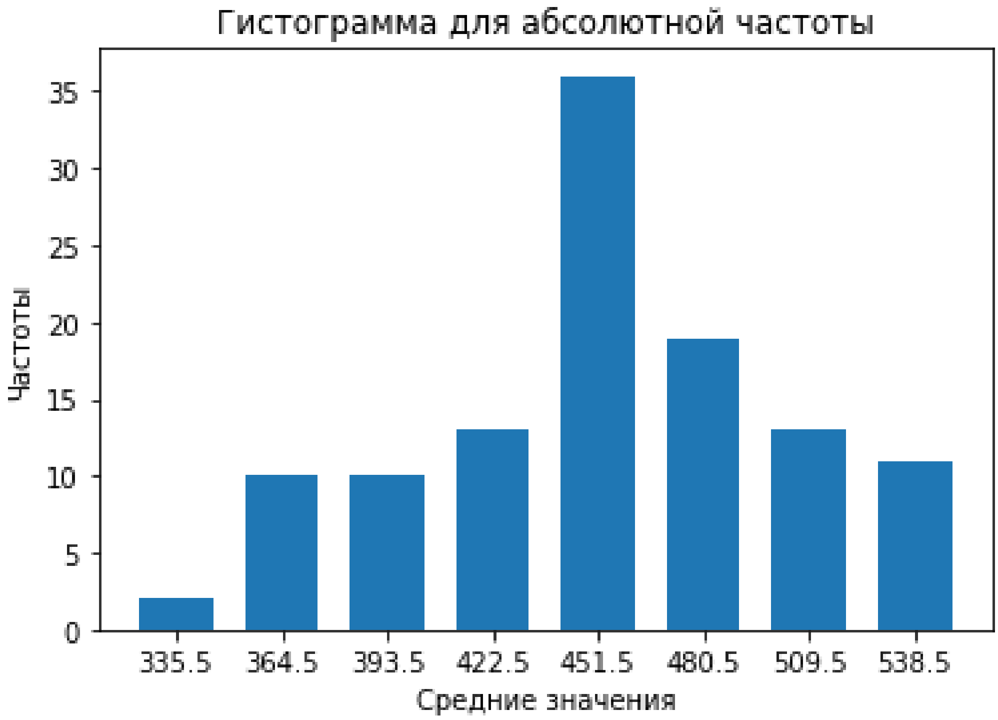

Рисунок 7 – Гистограмма для абсолютной частоты

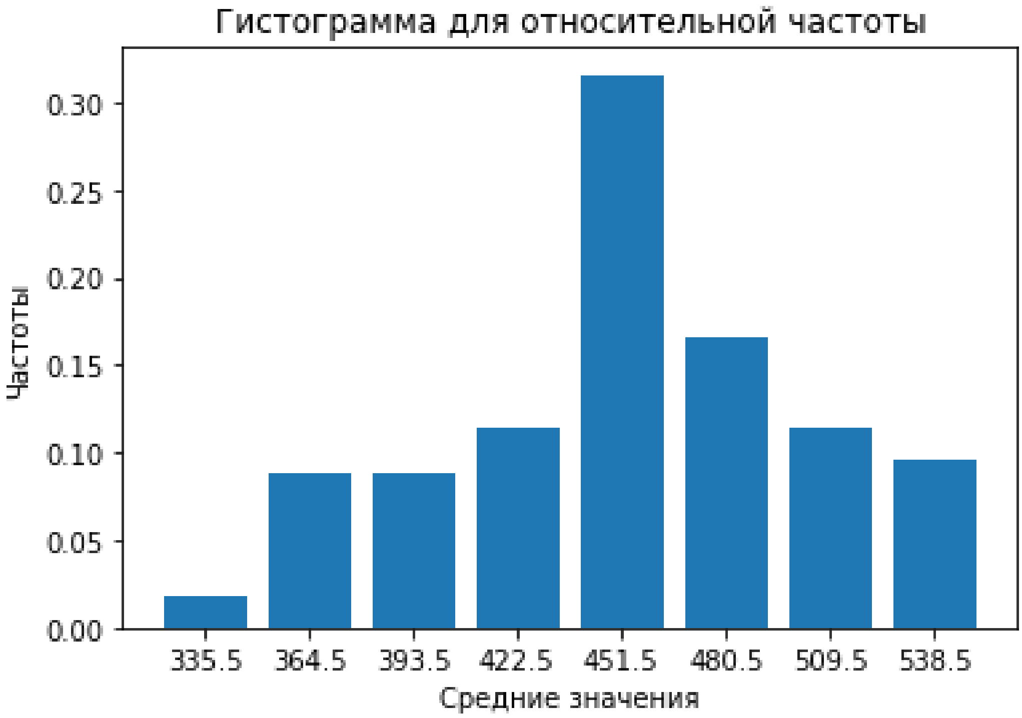

Рисунок 8 – Гистограмма для относительной частоты

Как мы можем заметить, и гистограммы, и полигоны, для абсолютных и относительных частот совпадают по форме, в связи с тем, что при преобразовании, отношения величин сохраняются.

Из данных графиков можно предположить, что выборка сформирована под действием нормального закона.

Далее строим эмпирические функции распределения применительно к интервальному ряду для абсолютных и относительных частот, представленные на рис.9 и 10.

Рисунок 9 – График эмпирической функции распределения для абсолютной частоты

Рисунок 10 – График эмпирической функции распределения для относительной частоты

Выводы.

В ходе данной лабораторной работы была сформирована выборка данных и осуществлена её первичная подготовка с помощью средств языка Python.

Выборка приведена к ранжированному, вариационному и интервальному видам. Используя полученный интервальный ряд построен полигон, гистограмма и эмпирическая функция распределения для абсолютных и относительных частот.

По гистограмме был сделан вывод о формировании выборки под действием нормального закона.

ПРИЛОЖЕНИЕ А
ИСХОДНЫЙ КОД

```
from tkinter import ROUND
import numpy as np
from math import *
import matplotlib.pyplot as plt

ROUND_ACC = 5

def m_round(a):
    return round(a, ROUND_ACC)


def m_np_round(a):
    return np.round(a, ROUND_ACC)


def make_spaced(_data):
    _n = len(_data)
    _k = 1 + 3.322 * log10(_n)
    _max = max(_data)
    _min = min(_data)
    _h = (_max - _min) / _k
    print("Интервальный ряд: k={} (формула Спенсера), h={}={} (шаг)".for-
mat(_k, _h, round(_h)))
    _h = round(_h)
    _from, _upto = _min, _min + _h
    _spaces = []
    _abs_freqs = []
    _rel_freqs = []
    _space_count = 0
    for i in _data:
        if i >= _upto:
            _spaces.append([_from, _upto])
            _abs_freqs.append(_space_count)
            _rel_freqs.append(_space_count / _n)
            _from = _upto
            _upto = _from + _h
            _space_count = 1
        else:
            _space_count += 1

    _spaces.append([_from, _upto])
    _abs_freqs.append(_space_count)
    _rel_freqs.append(_space_count / _n)
    return np.array(_spaces), np.array(_abs_freqs), np.array(_rel_freqs)


source_dataset = np.genfromtxt('dataset.csv', delimiter=',')
```

```
source_len = len(source_dataset)
data = source_dataset[:,0].astype(int)

print("Исходный ряд:\n", data)

data = np.sort(data)
ranked_data = np.array(data)
print("Ранжированный ряд:\n", ranked_data)

uniq, abs_freqs = np.unique(data, return_counts=True)
abs_freqs_table = np.array([tu-
ple([uniq[i], abs_freqs[i]]) for i in range(len(uniq))], dtype='i, i')
print("Вариационный ряд - абсолютные частоты:\n", abs_freqs_table)

rel_freqs = (abs_freqs / source_len).astype(float)
rel_freqs_table = np.array([tu-
ple([uniq[i], rel_freqs[i]]) for i in range(len(uniq))], dtype='i, f')
print("Вариационный ряд - относительные частоты:\n", rel_freqs_table)

spaces, spaces_abs_freq, spaces_rel_freq = make_spaced(data)
print("Интервальный ряд:\n{}\t{}\t{}".format("Интервал", "Абс. ча-
стоты", "Отн. частоты"))
for i in range(len(spaces)):
    print("[{}, {})\t{}\t\t{}".for-
mat(spaces[i][0], spaces[i][1], spaces_abs_freq[i], m_round(spaces_rel_freq[
i])))

spaces_median = np.median(spaces, axis=1)
spaces_median_str = [str(i) for i in spaces_median]

plt.plot(spaces_median, spaces_abs_freq)
plt.title("Полигон для абсолютной частоты")
plt.xlabel("Средние значения")
plt.ylabel("Частоты")
plt.grid()
plt.show()

plt.plot(spaces_median, spaces_rel_freq)
plt.title("Полигон для относительной частоты")
plt.xlabel("Средние значения")
plt.ylabel("Частоты")
plt.grid()
plt.show()

plt.bar(spaces_median_str, spaces_abs_freq, width=0.7)
plt.title("Гистограмма для абсолютной частоты")
plt.xlabel("Средние значения")
plt.ylabel("Частоты")
plt.show()
```

```
plt.bar(spaces_median_str, spaces_rel_freq)
plt.title("Гистограмма для относительной частоты")
plt.xlabel("Средние значения")
plt.ylabel("Частоты")
plt.show()

space_width = spaces_median[1] - spaces_median[0]
spaces_median_t = np.array([spaces_median[0] - space_width, *spaces_me-
dian, spaces_median[-1] + space_width])

spaces_distrib = [0, 0]
t = 0
for i in spaces_abs_freq:
    t += i
    spaces_distrib.append(t)

plt.step(spaces_median_t, spaces_distrib)
plt.title("Эмпирическая ф.р. для абсолютной частоты")
plt.xlabel("Средние значения")
plt.ylabel("Абсолютная частота")
plt.xticks(spaces_median_t)
plt.grid()
plt.show()

spaces_distrib = [0, 0]
t = 0
for i in spaces_abs_freq:
    t += i / np.sum(spaces_abs_freq)
    spaces_distrib.append(t)

plt.step(spaces_median_t, spaces_distrib)
plt.title("Эмпирическая ф.р. для относительной частоты")
plt.xlabel("Средние значения")
plt.ylabel("Относительная частота")
plt.xticks(spaces_median_t)
plt.grid()
plt.show()
```