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.

Студент гр. 8383 _____ Бабенко Н.С.

Студент гр. 8383 _____ Сахаров В.М.

Преподаватель _____ Середа А.-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 $x^m = (x_1, \dots, x_m)$, 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 $(x_1, n_1), (x_2, n_2), \dots, (x_k, n_k)$. 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.

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

В качестве генеральной совокупности были выбраны данные наблюдений относительно объемного веса nu $(\frac{г}{см^3})$ при влажности 10% и модуля упругости E $(\frac{кг}{см^2})$ при сжатии вдоль волокон древесины резонансной ели. Далее была сформирована репрезентативная выборка заданного объема из имеющейся генеральной совокупности экспериментальных данных при помощи библиотеки scikit-learn. Объём выборки: 100. Выборка представлена в таблице 1.

Таблица 1

№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***	№	***nu***	***E***
1	481	135.2	***21***	418	131.4	***41***	513	159.3	***61***	450	122.3	***81***	475	143.6
2	445	124.7	***22***	378	103.8	***42***	489	149.8	***62***	468	128.9	***82***	518	144.4
3	550	147.9	***23***	521	154.9	***43***	474	132.5	***63***	441	122.8	***83***	566	175.7
4	465	140.9	***24***	394	117.7	***44***	379	94.6	***64***	460	140.7	***84***	464	131.3
5	566	168.5	***25***	504	145.3	***45***	472	135.6	***65***	480	117.7	***85***	394	112.1
6	497	147.3	***26***	440	126.7	***46***	544	169.6	***66***	429	112.9	***86***	480	146.1
7	478	136.6	***27***	465	114.8	***47***	507	142.4	***67***	457	126.4	***87***	321	86.1
8	521	139.6	***28***	418	109.3	***48***	409	116.7	***68***	464	143.2	***88***	502	132.5
9	352	84.9	***29***	418	118.6	***49***	498	164.0	***69***	431	125.0	***89***	460	122.4
10	422	117.9	***30***	465	127.7	***50***	468	142.0	***70***	424	119.0	***90***	458	104.7
11	506	153.5	***31***	447	117.5	***51***	593	187.4	***71***	502	137.2	***91***	362	111.7
12	443	122.9	***32***	433	131.5	***52***	523	152.6	***72***	465	140.7	***92***	503	148.5
13	434	140.4	***33***	460	136.8	***53***	478	126.6	***73***	492	137.5	***93***	446	144.0
14	422	108.6	***34***	382	98.8	***54***	438	122.2	***74***	446	128.4	***94***	421	115.1
15	569	157.4	***35***	532	160.6	***55***	423	115.9	***75***	482	136.4	***95***	407	110.5
16	439	119.2	***36***	482	148.2	***56***	408	110.0	***76***	510	140.6	***96***	448	137.7
17	437	129.4	***37***	472	122.6	***57***	386	105.8	***77***	434	122.3	***97***	490	139.9
18	461	138.6	***38***	532	158.7	***58***	428	130.3	***78***	623	195.7	***98***	482	141.2
19	351	89.0	***39***	473	137.9	***59***	560	169.8	***79***	468	141.2	***99***	463	129.2
20	390	91.4	***40***	525	148.3	***60***	483	130.3	***80***	471	119.7	***100***	459	145.4

Выборка относительно переменной *nu* представлена в таблице 2.

Таблица 2

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	481	***21***	418	***41***	513	***61***	450	***81***	475
2	445	***22***	378	***42***	489	***62***	468	***82***	518
3	550	***23***	521	***43***	474	***63***	441	***83***	566
4	465	***24***	394	***44***	379	***64***	460	***84***	464
5	566	***25***	504	***45***	472	***65***	480	***85***	394
6	497	***26***	440	***46***	544	***66***	429	***86***	480
7	478	***27***	465	***47***	507	***67***	457	***87***	321
8	521	***28***	418	***48***	409	***68***	464	***88***	502
9	352	***29***	418	***49***	498	***69***	431	***89***	460
10	422	***30***	465	***50***	468	***70***	424	***90***	458
11	506	***31***	447	***51***	593	***71***	502	***91***	362
12	443	***32***	433	***52***	523	***72***	465	***92***	503
13	434	***33***	460	***53***	478	***73***	492	***93***	446
14	422	***34***	382	***54***	438	***74***	446	***94***	421
15	569	***35***	532	***55***	423	***75***	482	***95***	407
16	439	***36***	482	***56***	408	***76***	510	***96***	448
17	437	***37***	472	***57***	386	***77***	434	***97***	490
18	461	***38***	532	***58***	428	***78***	623	***98***	482
19	351	***39***	473	***59***	560	***79***	468	***99***	463
20	390	***40***	525	***60***	483	***80***	471	***100***	459

В таблице 3 представлено преобразование выборки в ранжированный ряд.

Таблица 3

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	321	***21***	423	***41***	457	***61***	473	***81***	504
2	351	***22***	424	***42***	458	***62***	474	***82***	506
3	352	***23***	428	***43***	459	***63***	475	***83***	507
4	362	***24***	429	***44***	460	***64***	478	***84***	510
5	378	***25***	431	***45***	460	***65***	478	***85***	513
6	379	***26***	433	***46***	460	***66***	480	***86***	518
7	382	***27***	434	***47***	461	***67***	480	***87***	521
8	386	***28***	434	***48***	463	***68***	481	***88***	521

9	390	***29***	437	***49***	464	***69***	482	***89***	523
10	394	***30***	438	***50***	464	***70***	482	***90***	525
11	394	***31***	439	***51***	465	***71***	482	***91***	532
12	407	***32***	440	***52***	465	***72***	483	***92***	532
13	408	***33***	441	***53***	465	***73***	489	***93***	544
14	409	***34***	443	***54***	465	***74***	490	***94***	550
15	418	***35***	445	***55***	468	***75***	492	***95***	560
16	418	***36***	446	***56***	468	***76***	497	***96***	566
17	418	***37***	446	***57***	468	***77***	498	***97***	566
18	421	***38***	447	***58***	471	***78***	502	***98***	569
19	422	***39***	448	***59***	472	***79***	502	***99***	593
20	422	***40***	450	***60***	472	***80***	503	***100***	623

Из таблицы 3 видно, что наименьшее значение в выборке $x_{min} = 321$, а наибольшее значение $x_{max} = 623$.

В таблице 4 представлено преобразование полученной выборки в вариационный ряд с абсолютными $\boldsymbol{n_i}$ и относительными $\overline{\boldsymbol{n_i}}$ частотами соответственно.

Таблица 4

$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\overline{\boldsymbol{n_i}}$	$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\overline{\boldsymbol{n_i}}$	$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\overline{\boldsymbol{n_i}}$	$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\overline{\boldsymbol{n_i}}$
1	321	1	0.01	***26***	439	1	0.01	***51***	481	1	0.01	***76***	593	1	0.01
2	351	1	0.01	***27***	440	1	0.01	***52***	482	3	0.03	***77***	623	1	0.01
3	352	1	0.01	***28***	441	1	0.01	***53***	483	1	0.01				
4	362	1	0.01	***29***	443	1	0.01	***54***	489	1	0.01				
5	378	1	0.01	***30***	445	1	0.01	***55***	490	1	0.01				
6	379	1	0.01	***31***	446	2	0.02	***56***	492	1	0.01				
7	382	1	0.01	***32***	447	1	0.01	***57***	497	1	0.01				
8	386	1	0.01	***33***	448	1	0.01	***58***	498	1	0.01				
9	390	1	0.01	***34***	450	1	0.01	***59***	502	2	0.02				
10	394	2	0.02	***35***	457	1	0.01	***60***	503	1	0.01				
11	407	1	0.01	***36***	458	1	0.01	***61***	504	1	0.01				
12	408	1	0.01	***37***	459	1	0.01	***62***	506	1	0.01				
13	409	1	0.01	***38***	460	3	0.03	***63***	507	1	0.01				
14	418	3	0.03	***39***	461	1	0.01	***64***	510	1	0.01				
15	421	1	0.01	***40***	463	1	0.01	***65***	513	1	0.01				
16	422	2	0.02	***41***	464	2	0.02	***66***	518	1	0.01				
17	423	1	0.01	***42***	465	4	0.04	***67***	521	2	0.02				
18	424	1	0.01	***43***	468	3	0.03	***68***	523	1	0.01				

19	428	1	0.01	***44***	471	1	0.01	***69***	525	1	0.01			
20	429	1	0.01	***45***	472	2	0.02	***70***	532	2	0.02			
21	431	1	0.01	***46***	473	1	0.01	***71***	544	1	0.01			
22	433	1	0.01	***47***	474	1	0.01	***72***	550	1	0.01			
23	434	2	0.02	***48***	475	1	0.01	***73***	560	1	0.01			
24	437	1	0.01	***49***	478	2	0.02	***74***	566	2	0.02			
25	438	1	0.01	***50***	480	2	0.02	***75***	569	1	0.01			

Из таблицы 4 можно увидеть моду выборки, которой является варианта $x_{42} = 465$ с абсолютной частотой равной 4.

Чтобы преобразовать вариационный ряд в интервальный ряд сначала нужно вычислить количество интервалов разбиения с помощью формулы Стерджесса:

$$k = 1 + 3.31 * \lg N = 7$$

Далее вычислена ширина интервала с помощью формулы:

$$h = \frac{x_{max} - x_{min}}{k} = \frac{623 - 321}{7} \approx 44$$

В таблице 5 представлен полученный интервальный ряд.

Таблица 5

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

Далее для интервального ряда абсолютных частот были построены полигон и гистограмма.

Полигон представлен на рис. 1.

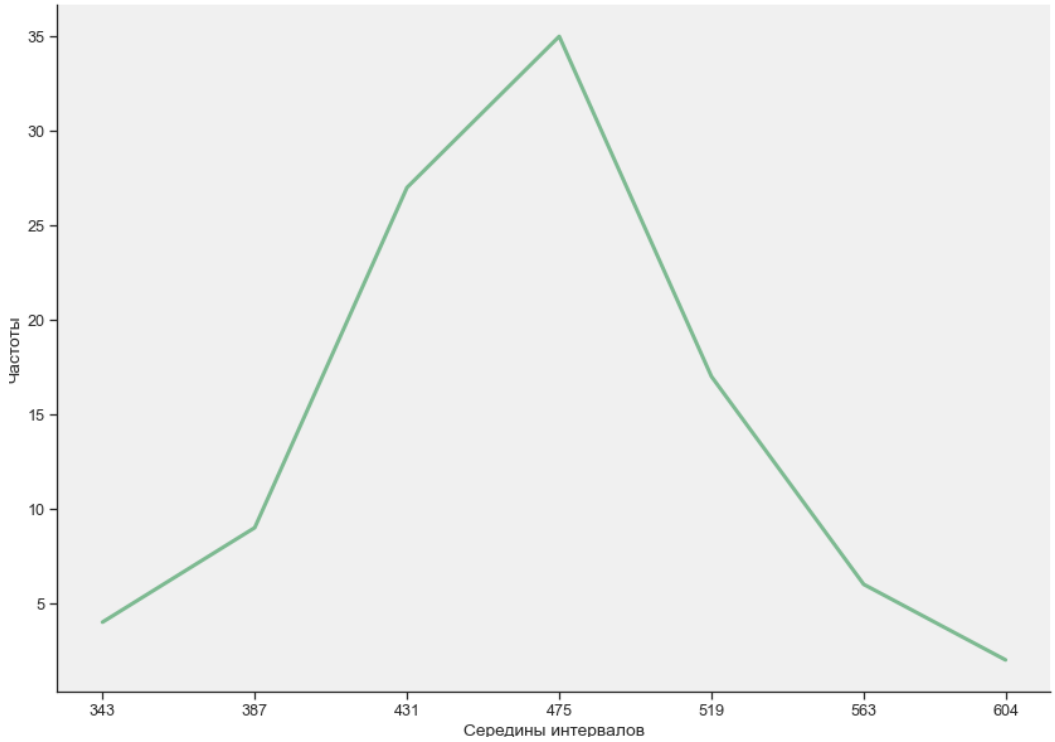

Рисунок 1 – Полигон для абсолютных частот

Полигон представляет собой ломаную, соединяющую точки, соответствующие срединным значениям интервалов и абсолютным частотам этих интервалов. Видно, что на пике значение равно 35, что сходится с данными таблицы 5.

Гистограмма, представлена на рис. 2.

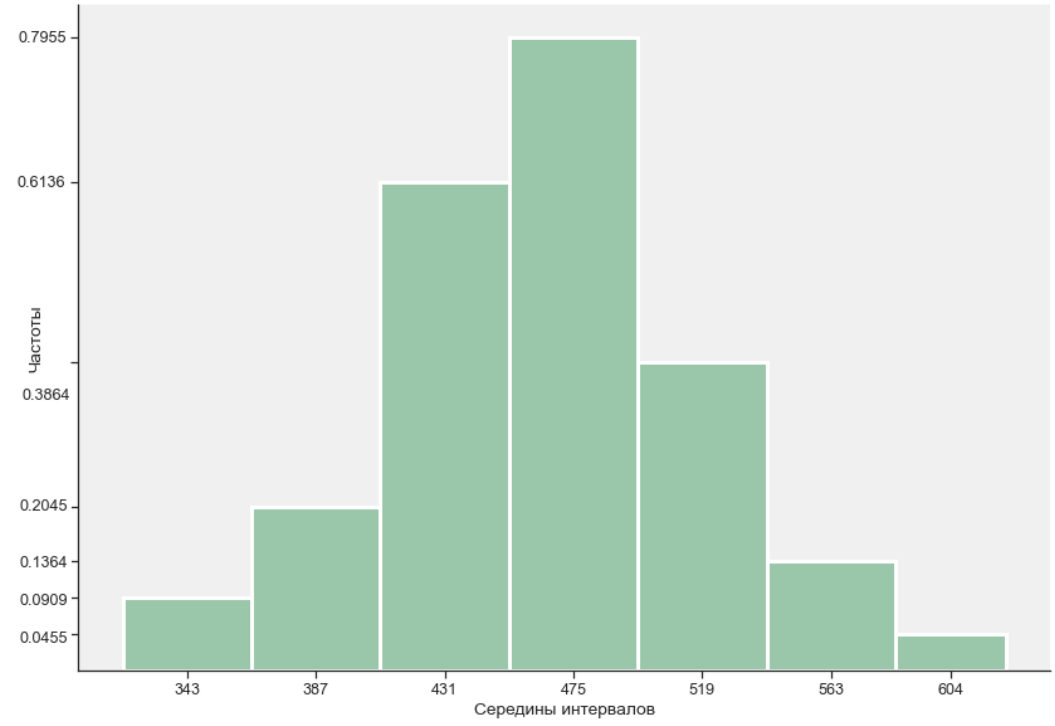

Рисунок 2 – Гистограмма для абсолютных частот

Гистограмма представляет собой фигуру, состоящую из прямоугольников, основания которых это длина интервалов h, а высота равна отношению частоты к длине интервала, то есть площадь прямоугольника обозначает частоту интервала.

Графики для интервального ряда относительных частот представлены ниже.

Полигон для относительных частот представлен на рис. 3.

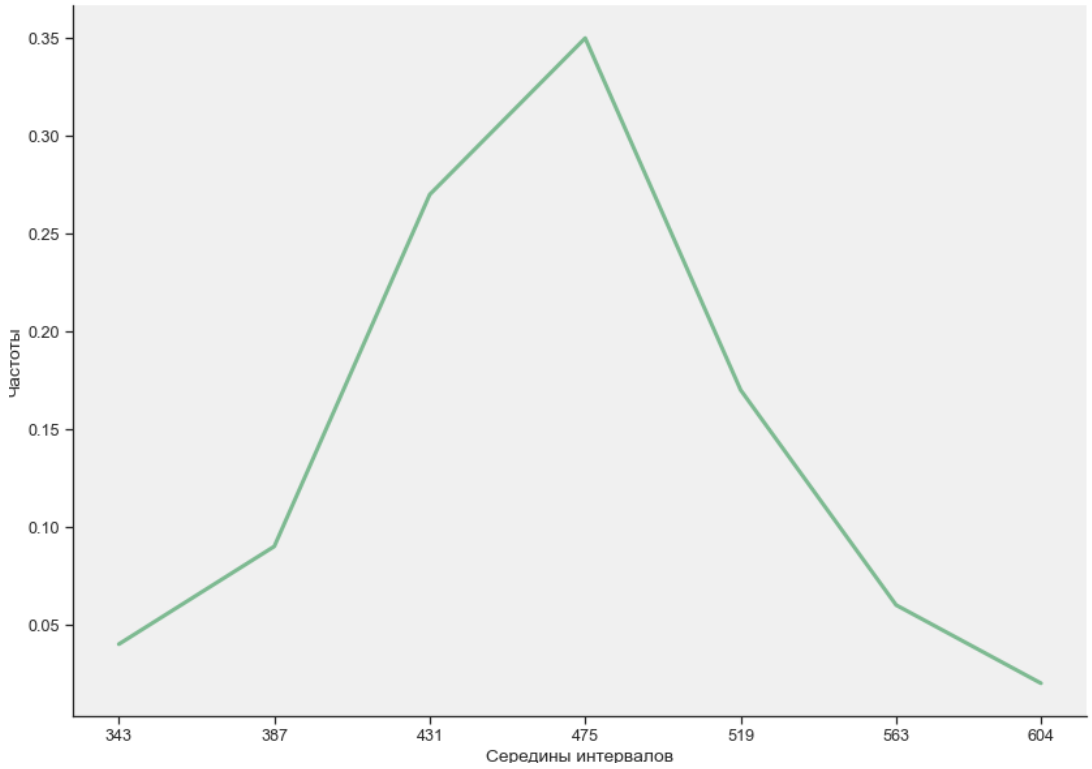

Рисунок 3 – Полигон для относительных частот

Гистограмма для относительных частот, представлена на рис. 4.

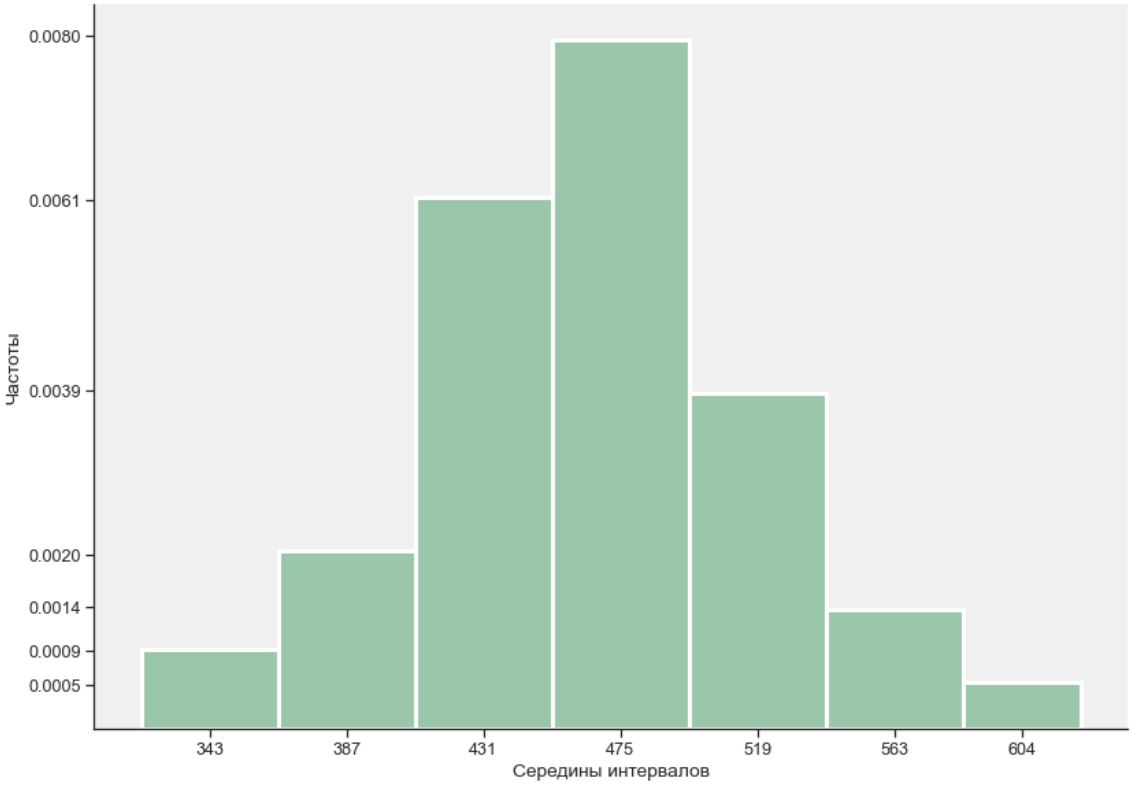

Рисунок 4 – Гистограмма для относительных частот

Эмпирическая функция распределения, построенная применительно к интервальному ряду для относительных частот представлен на рис. 5.

Функция распределения:

$$F(343) = 0$$

$$F(387) = 0.04$$

$$F(431) = 0.13$$

$$F(475) = 0.40$$

$$F(519) = 0.75$$

$$F(563) = 0.92$$

$$F(604) = 0.98$$

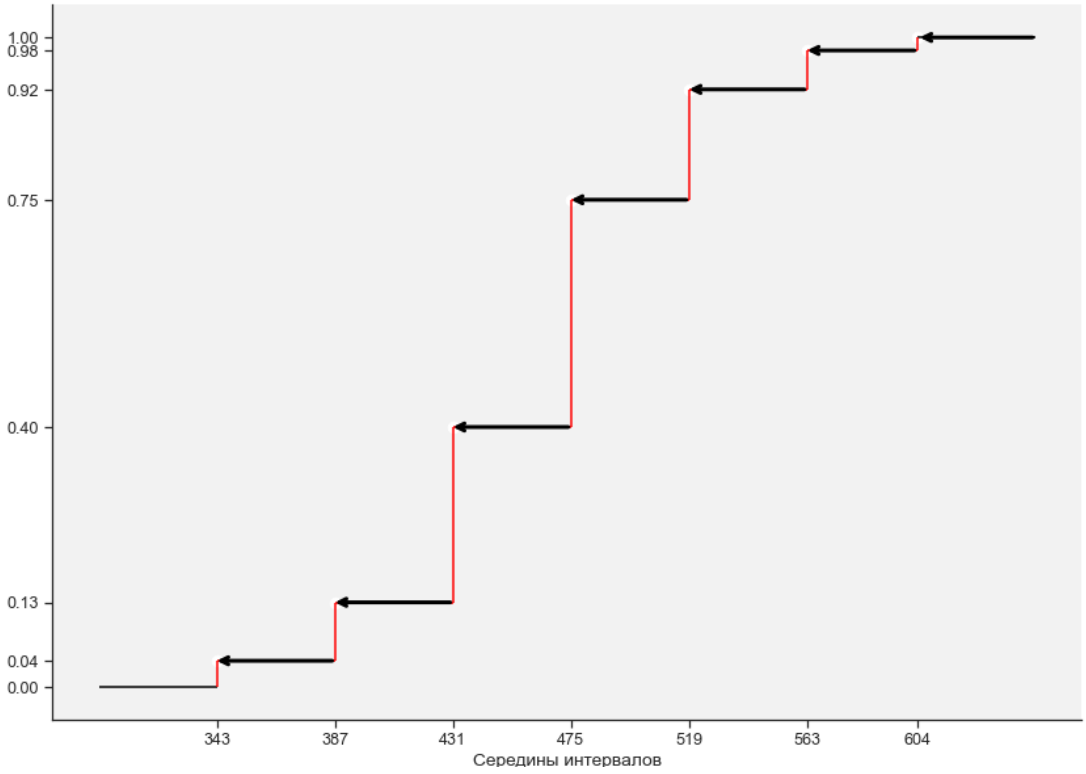

Рисунок 5 – График эмпирической функции распределения

Выводы

В ходе данной лабораторной работы была выбрана выборка, которая представляет собой данные наблюдений относительно объемного веса nu $(\frac{\text{г}}{\text{см}^3})$ при влажности 10% и модуля упругости E $(\frac{\text{кг}}{\text{см}^2})$ при сжатии вдоль волокон древесины резонансной ели. Выборка была преобразована в ранжированный, вариационный и интервальный ряды.

С помощью ранжированного ряда удалось определить минимальный и максимальный элемент выборки $x_{min} = 321$, $x_{max} = 623$, так как его элементы находятся в порядке возрастания. Далее при преобразовании ряда в вариационный ряд (объединение одинаковых элементов) удалось определить моду – значение в выборке, которое встречается наиболее часто, для данной выборки это $x_{42} = 465$ с абсолютной $n_{42} = 4$ и относительной частотой $\overline{n_{42}} = 0.04$. Далее при преобразовании интервального ряда из вариационного с помощью высчитанных значений количества интервалов $k = 7$ (нечетное) и последующего $h =$

44 можно было заметить, что наибольшая частота попаданий в интервал равная $n = 35$ находится в интервале $[453, 497)$.

Построенные графики также помогают увидеть наглядное представление ряда распределения. Видно, например, что в интервале $[453, 497)$ больше всего значений. Полигон строится как ломаная, которая соединяет точки, соответствующие срединным значениям интервалов и частотам этих интервалов, поэтому его форма не меняется для абсолютных и относительных частот, а меняется ось ординат, где как раз откладывают соответствующие абсолютные или относительные частоты. Гистограмма же — это фигура, состоящая из прямоугольников, площадь которых как раз и обозначает соответствующие частоты. Можно проверить, что для гистограммы абсолютных частот общая площадь прямоугольников равна объему выборки, а для гистограммы относительных частот она равна единице. Эмпирическая функция распределения же показывает отношение накопленных частот до середины интервалов к объему выборки $n = 100$, где опять же видно, как на интервале $[497, 541)$ с серединой равной 519 накопленная частота резко увеличивается.

ПРИЛОЖЕНИЕ А
ИСХОДНЫЙ КОД

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df.to_csv('data/data1.csv', index=False)
n = len(df)
df2 = df.drop('E', axis=1)
df2.to_csv('data/data2.csv', index=False)
df2.head()
df2 = df2.sort_values(by=['nu'], ignore_index = True)
df2.to_csv('data/data3.csv', index=False)
df2.head()
df2.min()
df2.max()
X = df2['nu']
X.mode()
table_af = X.value_counts().sort_index()
table_rf = X.value_counts(normalize=True).sort_index()
table_af = pd.DataFrame({'nu': table_af.index, 'af': table_af.values})
table_rf = pd.DataFrame({'nu': table_rf.index, 'rf': table_rf.values})
table_rf2 = table_rf.copy()
table_rf2['rf'] = np.round(table_rf2['rf'], 4)
table_af.to_csv('data/data4.csv', index=False)
table_rf2.to_csv('data/data5.csv', index=False)
k = 1+3.31*np.log10(n)
k = int(np.floor(k))
min(X)
max(X)
h = (max(X)-min(X))/k
h = int(np.ceil(h))
```

```
data_interval = pd.concat([table_af, table_rf], ignore_index=True,
axis=1).drop(2, axis=1)
data_interval.columns = ['nu', 'af', 'rf']
data_interval.to_csv('data/data6.csv', index=False)
ivs = np.hstack((np.arange(min(X), max(X), h), np.array(max(X))))
data_interval['inter'] = pd.cut(data_interval['nu'], bins=ivs,
                                right=False)
data_interval.iloc[76, 3] = data_interval.iloc[75, 3]
data_interval['inter'].value_counts().sort_index()
f_inter = data_interval.groupby(['inter'])[['af', 'rf']].apply(sum).re-
set_index()
f_inter['avg_inter'] = np.array([np.mean([ivs[i], ivs[i+1]], axis=0) for
i in range(k)])
f_inter = f_inter[['inter', 'avg_inter', 'af', 'rf']]
f_inter['rf'] = np.round(f_inter['rf'], 2)
f_inter.to_csv('data/data7.csv', index=False)
sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style('ticks', {"axes.facecolor": ".94"})
ax = sns.relplot(data=f_inter, x='avg_inter', y='af', kind='line',
                 height=8.27, aspect=11.7/8.27, linewidth=3)
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'])
plt.savefig('pics/3.png')
ax = sns.displot(data=df, x='nu', bins=ivs, kind='hist',
                 height=8.27, aspect=11.7/8.27, linewidth=3)
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'], yticks=f_inter['af'])
plt.savefig('pics/4.png')
f_inter['sum_rf'] = f_inter['rf'].cumsum()
f_inter
f_inter
ax = sns.relplot(data=f_inter, x='avg_inter', y='sum_rf', s=80,
                 kind='scatter', height=8.27, aspect=11.7/8.27,
color='w')
```

```
for i in range(6):
    plt.hlines(f_inter['sum_rf'][i], f_inter['avg_inter'][i], f_in-
ter['avg_inter'][i+1], color='r')
plt.hlines(1, 604, 624, color='r')
for i in range(6):
    plt.vlines(f_inter['avg_inter'][i+1], f_inter['sum_rf'][i], f_in-
ter['sum_rf'][i+1], color='r', linestyle='-')
plt.vlines(343, 0, 0.04, color='r', linestyle='-')
for i in range(6):
    plt.annotate('', xy=(f_inter['avg_inter'][i]-1, f_in-
ter['sum_rf'][i]),
                 xytext=(f_inter['avg_inter'][i+1], f_in-
ter['sum_rf'][i]),
                 arrowprops=dict(arrowstyle="->", color='r', lin-
ewidth=3))
plt.annotate('', xy=(604, 1),
                 xytext=(624, 1),
                 arrowprops=dict(arrowstyle="->", color='r', lin-
ewidth=3))
ax.set_axis_labels('Середины интервалов', '')
ax.set(xticks=f_inter['avg_inter'])
plt.savefig('pics/5.png')
ax = sns.relplot(data=f_inter, x='avg_inter', y='rf', kind='line',
                 height=8.27, aspect=11.7/8.27, linewidth=3)
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'])
plt.savefig('pics/6.png')
ax = sns.displot(data=df, x='nu', bins=ivs, kind='hist', linewidth=3,
                 height=8.27, aspect=11.7/8.27, stat='density')
ax.set_axis_labels('Середины интервалов', 'Частоты')
ax.set(xticks=f_inter['avg_inter'], yticks=round((f_inter['rf']/h), 4))
plt.savefig('pics/7.png')
```