

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

**ОТЧЕТ**  
**по лабораторной работе №4**  
**по дисциплине «Статистические методы обработки экспериментальных**  
**данных»**  
**Тема: Элементы корреляционного анализа. Проверка статистической**  
**гипотезы о равенстве коэффициента корреляции нулю.**

Студент гр. 8383

\_\_\_\_\_

Киреев К.А.

Студент гр. 8383

\_\_\_\_\_

Муковский Д.В.

Преподаватель

\_\_\_\_\_

Середа А.-В.И.

Санкт-Петербург

2022

## Цель работы

Освоение основных понятий, связанных с корреляционной зависимостью между случайными величинами, доверительными интервалами, статистическими гипотезами и проверкой их «справедливости».

## Основные теоретические положения

Рассмотрим систему двух случайных величин  $\{X; Y\}$ . Эти случайные величины могут быть независимыми:

$$f(x, y) = f_1(x) \cdot f_2(y)$$

В противном случае между ними может быть:

- Функциональная зависимость:

$$y = g(x)$$

- Статистическая зависимость:

$$\varphi(x/y) = \frac{f(x, y)}{f_2(y)}$$

$$\phi(y/x) = \frac{f(x, y)}{f_1(x)}$$

Частным случаем статистической зависимости является корреляционная зависимость. Корреляционной называют статистическую зависимость двух случайных величин, при которой изменение значения одной из случайных величин приводит к изменению математического ожидания другой случайной величины:

$$M(X/y) = q_1(y)$$

$$M(Y/x) = q_2(x)$$

Корреляционный момент:

$$\mu_{xy} = M\{[x - M(X)] \cdot [y - M(Y)]\}$$

Коэффициент корреляции:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y}$$

Для коэффициента корреляции справедливо соотношение:

$$|r_{xy}| \leq 1$$

Случайные величины называют коррелированными, если их корреляционный момент или их коэффициент корреляции отличен от нуля. В противном случае эти величины некоррелированные. Если случайные величины  $X$  и  $Y$  коррелированы, то они зависимы.

Коэффициент корреляции служит мерой тесноты линейной зависимости между случайными величинами  $X$  и  $Y$ . При  $|r_{xy}| = 1$  эта зависимость становится функциональной.

Значение  $\bar{r}_{xy}$  – статистической оценки  $r_{xy}$  – коэффициента корреляции можно вычислить по формуле:

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y} \sum_{j=1}^{K_x} n_{ij} y_i x_j - N \bar{x}_B \bar{y}_B}{N S_x S_y}$$

При  $N > 50$  в случае нормального распределения системы случайных величин  $\{X; Y\}$  для оценки значения  $\bar{r}_{xy}$  можно использовать соотношение:

$$\bar{r}_{xy} - 3 \frac{1 - \bar{r}_{xy}^2}{\sqrt{N}} \leq r_{xy} \leq \bar{r}_{xy} + 3 \frac{1 + \bar{r}_{xy}^2}{\sqrt{N}}$$

Для построения доверительного интервала с помощью преобразования Фишера перейдём к случайной величине  $z$ :

$$\bar{z} = 0.5 \ln \frac{1 + \bar{r}_{xy}}{1 - \bar{r}_{xy}}$$

Распределение  $z$  при неограниченном возрастании объёма выборки асимптотически нормальное со значением СКО:

$$\bar{\sigma}_z = \frac{1}{\sqrt{N - 3}}$$

Доверительный интервал для генерального значения:

$$(\bar{z} - \lambda(\gamma) \bar{\sigma}_z; \bar{z} + \lambda(\gamma) \bar{\sigma}_z), \text{ где}$$

$$\Phi(\lambda(\gamma)) = \frac{\gamma}{2}$$

Для пересчёта интервала в доверительный интервал для коэффициента корреляции с тем же значением  $\gamma$  необходимо воспользоваться обратным преобразованием Фишера:

$$r = th(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Гипотеза  $H_0: r_{xy} = 0$ . Альтернативой будет гипотеза  $H_1: r_{xy} \neq 0$ . Если основная гипотеза отвергается, то это означает, что выборочный коэффициент корреляции  $\bar{r}_{xy}$  значимо отличается от нуля (значим). В качестве критерия проверки статистической гипотезы о значимости выборочного коэффициента корреляции можно принять случайную величину:

$$T = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}}$$

При справедливости нулевой гипотезы случайная величина  $T$  распределена по закону Стьюдента с  $k = K - 2$  степенями свободы. Критическая область для данного критерия двусторонняя. Если  $|T_{\text{набл}}| \leq t_{\text{крит}}(\alpha, k)$  – нет оснований отвергать гипотезу  $H_0$ . Если  $|T_{\text{набл}}| > t_{\text{крит}}(\alpha, k)$  – основная гипотеза  $H_0$  с выборочными данными должна быть отвергнута.

### **Постановка задачи**

Из заданной генеральной совокупности сформировать выборку по второму признаку. Провести статистическую обработку второй выборки в объеме лабораторных работ №1 и №2, с целью определения точечных статистических оценок параметров распределения исследуемого признака (математического ожидания, дисперсии, среднеквадратичного отклонения, асимметрии и эксцесса). Для системы двух случайных величин  $X$  (первый признак) и  $Y$  (второй признак) сформировать двумерную выборку и найти статистическую оценку коэффициента корреляции, построить доверительный интервал для коэффициента корреляции и осуществить проверку статистической гипотезы о равенстве коэффициента корреляции нулю. Полученные результаты содержательно проинтерпретировать.

## Порядок выполнения работы

- Провести статистическую обработку второй выборки в объеме лабораторных работ №1 и №2, с целью определения точечных статистических оценок параметров распределения исследуемого признака (математического ожидания, дисперсии, среднеквадратичного отклонения, асимметрии, эксцесса, моды и медианы). Оформить результаты в виде таблицы, сделать выводы.
- Построить двумерный интервальный вариационный ряд, оформить в виде таблицы.
- По полученному двумерному интервальному вариационному ряду построить корреляционную таблицу, сделать выводы.
- Исходя из результатов корреляционной таблицы вычислить статистическую оценку корреляционного момента.
- Вычислить коэффициент корреляции, сделать выводы.
- Построить доверительный интервал для коэффициента корреляции при уровне значимости  $\gamma \in \{0.95, 0.99\}$ , сделать выводы.
- Осуществить проверку статистической гипотезы о равенстве коэффициента корреляции нулю при заданном уровне значимости  $\alpha=0.05$  сделать выводы.

## Выполнение работы

### *Статистическая обработка выборки по второму признаку*

Формирование репрезентативной выборки заданного объема из имеющейся генеральной совокупности экспериментальных данных представлено в таблице 1. Объём выборки: 104.

*Таблица 1*

№	<i>ni</i>	<i>E</i>	№	<i>ni</i>	<i>E</i>	№	<i>ni</i>	<i>E</i>	№	<i>ni</i>	<i>E</i>	№	<i>ni</i>	<i>E</i>
1	460	124.5	25	394	112.1	49	411	112.9	73	428	131.6	97	378	103.8
2	525	148.3	26	434	118.6	50	451	124.3	74	510	140.6	98	576	170.1

*Продолжение таблицы 1*

3	503	146.6	27	518	151.3	51	466	130.3	75	478	126.6	99	452	116.1
4	482	148.2	28	522	143.8	52	433	130.0	76	421	115.1	100	543	155.4
5	470	124.0	29	511	149.5	53	492	137.5	77	510	153.9	101	538	165.0
6	400	114.6	30	437	124.3	54	503	148.5	78	351	102.9	102	523	172.8
7	398	109.0	31	352	87.7	55	451	128.6	79	493	149.7	103	434	108.7
8	514	174.6	32	406	112.4	56	415	107.1	80	411	115.2	104	458	128.0
9	518	154.0	33	448	125.9	57	459	145.4	81	422	108.6			
10	383	109.7	34	493	129.7	58	442	123.4	82	402	120.8			
11	412	117.9	35	468	128.9	59	424	117.1	83	438	126.7			
12	320	64.5	36	345	95.9	60	397	108.6	84	485	138.6			
13	473	137.9	37	523	152.6	61	414	113.5	85	496	155.3			
14	438	134.1	38	498	144.3	62	437	129.2	86	453	126.4			
15	359	71.9	39	482	139.9	63	512	169.9	87	377	96.1			
16	569	157.4	40	487	146.0	64	525	165.9	88	540	156.7			
17	423	115.9	41	331	84.6	65	546	177.0	89	502	137.2			
18	460	140.7	42	416	120.5	66	422	122.9	90	408	110.0			
19	372	81.7	43	358	98.3	67	495	150.9	91	417	124.3			
20	383	107.4	44	463	144.9	68	452	131.0	92	474	132.5			
21	409	116.7	45	462	138.8	69	465	140.7	93	480	153.9			
22	444	130.0	46	413	110.8	70	391	107.5	94	483	130.3			
23	463	136.7	47	506	153.5	71	426	128.2	95	472	135.6			
24	482	150.1	48	465	140.9	72	482	136.4	96	477	146.0			

В таблице 2 представлена выборка только для  $E$ .

*Таблица 2*

$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$	$i$	$x_i$
1	124.5	25	112.1	49	112.9	73	131.6	97	103.8
2	148.3	26	118.6	50	124.3	74	140.6	98	170.1
3	146.6	27	151.3	51	130.3	75	126.6	99	116.1
4	148.2	28	143.8	52	130.0	76	115.1	100	155.4
5	124.0	29	149.5	53	137.5	77	153.9	101	165.0
6	114.6	30	124.3	54	148.5	78	102.9	102	172.8

*Продолжение таблицы 2*

7	109.0	31	87.7	55	128.6	79	149.7	103	108.7
8	174.6	32	112.4	56	107.1	80	115.2	104	128.0
9	154.0	33	125.9	57	145.4	81	108.6		
10	109.7	34	129.7	58	123.4	82	120.8		
11	117.9	35	128.9	59	117.1	83	126.7		
12	64.5	36	95.9	60	108.6	84	138.6		
13	137.9	37	152.6	61	113.5	85	155.3		
14	134.1	38	144.3	62	129.2	86	126.4		
15	71.9	39	139.9	63	169.9	87	96.1		
16	157.4	40	146.0	64	165.9	88	156.7		
17	115.9	41	84.6	65	177.0	89	137.2		
18	140.7	42	120.5	66	122.9	90	110.0		
19	81.7	43	98.3	67	150.9	91	124.3		
20	107.4	44	144.9	68	131.0	92	132.5		
21	116.7	45	138.8	69	140.7	93	153.9		
22	130.0	46	110.8	70	107.5	94	130.3		
23	136.7	47	153.5	71	128.2	95	135.6		
24	150.1	48	140.9	72	136.4	96	146.0		

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	64.5	25	114.6	49	128.6	73	140.9	97	157.4
2	71.9	26	115.1	50	128.9	74	143.8	98	165.0
3	81.7	27	115.2	51	129.2	75	144.3	99	165.9
4	84.6	28	115.9	52	129.7	76	144.9	100	169.9
5	87.7	29	116.1	53	130.0	77	145.4	101	170.1
6	95.9	30	116.7	54	130.0	78	146.0	102	172.8
7	96.1	31	117.1	55	130.3	79	146.0	103	174.6
8	98.3	32	117.9	56	130.3	80	146.6	104	177.0
9	102.9	33	118.6	57	131.0	81	148.2		
10	103.8	34	120.5	58	131.6	82	148.3		
11	107.1	35	120.8	59	132.5	83	148.5		
12	107.4	36	122.9	60	134.1	84	149.5		
13	107.5	37	123.4	61	135.6	85	149.7		
14	108.6	38	124.0	62	136.4	86	150.1		
15	108.6	39	124.3	63	136.7	87	150.9		

*Продолжение таблицы 3*

16	108.7	40	124.3	64	137.2	88	151.3		
17	109.0	41	124.3	65	137.5	89	152.6		
18	109.7	42	124.5	66	137.9	90	153.5		
19	110.0	43	125.9	67	138.6	91	153.9		
20	110.8	44	126.4	68	138.8	92	153.9		
21	112.1	45	126.6	69	139.9	93	154.0		
22	112.4	46	126.7	70	140.6	94	155.3		
23	112.9	47	128.0	71	140.7	95	155.4		
24	113.5	48	128.2	72	140.7	96	156.7		

В таблице 3 можно заметить, что наименьшее значение в выборке  $x_{min} = 64.5$ , а наибольшее значение  $x_{max} = 177$ .

В таблицах 4 и 5 представлено преобразование полученной выборки в вариационный ряд с абсолютными и относительными частотами соответственно.

*Таблица 4*

$i$	$x_i$	$n_i$	$i$	$x_i$	$n_i$	$i$	$x_i$	$n_i$	$i$	$x_i$	$n_i$
1	64.5	1	25	115.1	1	49	129.7	1	73	146.6	1
2	71.9	1	26	115.2	1	50	130.0	2	74	148.2	1
3	81.7	1	27	115.9	1	51	130.3	2	75	148.3	1
4	84.6	1	28	116.1	1	52	131.0	1	76	148.5	1
5	87.7	1	29	116.7	1	53	131.6	1	77	149.5	1
6	95.9	1	30	117.1	1	54	132.5	1	78	149.7	1
7	96.1	1	31	117.9	1	55	134.1	1	79	150.1	1
8	98.3	1	32	118.6	1	56	135.6	1	80	150.9	1
9	102.9	1	33	120.5	1	57	136.4	1	81	151.3	1
10	103.8	1	34	120.8	1	58	136.7	1	82	152.6	1
11	107.1	1	35	122.9	1	59	137.2	1	83	153.5	1
12	107.4	1	36	123.4	1	60	137.5	1	84	153.9	2
13	107.5	1	37	124.0	1	61	137.9	1	85	154.0	1
14	108.6	2	38	124.3	3	62	138.6	1	86	155.3	1
15	108.7	1	39	124.5	1	63	138.8	1	87	155.4	1
16	109.0	1	40	125.9	1	64	139.9	1	88	156.7	1
17	109.7	1	41	126.4	1	65	140.6	1	89	157.4	1



*Продолжение таблицы 4*

18	110.0	1	42	126.6	1	66	140.7	2	90	165.0	1
19	110.8	1	43	126.7	1	67	140.9	1	91	165.9	1
20	112.1	1	44	128.0	1	68	143.8	1	92	169.9	1
21	112.4	1	45	128.2	1	69	144.3	1	93	170.1	1
22	112.9	1	46	128.6	1	70	144.9	1	94	172.8	1
23	113.5	1	47	128.9	1	71	145.4	1	95	174.6	1
24	114.6	1	48	129.2	1	72	146.0	2	96	177.0	1

*Таблица 5*

$i$	$x_i$	$\overline{n}_i$	$i$	$x_i$	$\overline{n}_i$	$i$	$x_i$	$\overline{n}_i$	$i$	$x_i$	$\overline{n}_i$
1	64.5	0.0096	25	115.1	0.0096	49	129.7	0.0096	73	146.6	0.0096
2	71.9	0.0096	26	115.2	0.0096	50	130.0	0.0192	74	148.2	0.0096
3	81.7	0.0096	27	115.9	0.0096	51	130.3	0.0192	75	148.3	0.0096
4	84.6	0.0096	28	116.1	0.0096	52	131.0	0.0096	76	148.5	0.0096
5	87.7	0.0096	29	116.7	0.0096	53	131.6	0.0096	77	149.5	0.0096
6	95.9	0.0096	30	117.1	0.0096	54	132.5	0.0096	78	149.7	0.0096
7	96.1	0.0096	31	117.9	0.0096	55	134.1	0.0096	79	150.1	0.0096
8	98.3	0.0096	32	118.6	0.0096	56	135.6	0.0096	80	150.9	0.0096
9	102.9	0.0096	33	120.5	0.0096	57	136.4	0.0096	81	151.3	0.0096
10	103.8	0.0096	34	120.8	0.0096	58	136.7	0.0096	82	152.6	0.0096
11	107.1	0.0096	35	122.9	0.0096	59	137.2	0.0096	83	153.5	0.0096
12	107.4	0.0096	36	123.4	0.0096	60	137.5	0.0096	84	153.9	0.0192
13	107.5	0.0096	37	124.0	0.0096	61	137.9	0.0096	85	154.0	0.0096
14	108.6	0.0192	38	124.3	0.0288	62	138.6	0.0096	86	155.3	0.0096
15	108.7	0.0096	39	124.5	0.0096	63	138.8	0.0096	87	155.4	0.0096
16	109.0	0.0096	40	125.9	0.0096	64	139.9	0.0096	88	156.7	0.0096
17	109.7	0.0096	41	126.4	0.0096	65	140.6	0.0096	89	157.4	0.0096
18	110.0	0.0096	42	126.6	0.0096	66	140.7	0.0192	90	165.0	0.0096
19	110.8	0.0096	43	126.7	0.0096	67	140.9	0.0096	91	165.9	0.0096
20	112.1	0.0096	44	128.0	0.0096	68	143.8	0.0096	92	169.9	0.0096
21	112.4	0.0096	45	128.2	0.0096	69	144.3	0.0096	93	170.1	0.0096
22	112.9	0.0096	46	128.6	0.0096	70	144.9	0.0096	94	172.8	0.0096
23	113.5	0.0096	47	128.9	0.0096	71	145.4	0.0096	95	174.6	0.0096
24	114.6	0.0096	48	129.2	0.0096	72	146.0	0.0192	96	177.0	0.0096

С помощью формулы Стерджесса было вычислено количество интервалов:

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

Получено нечетное количество интервалов.

Ширина интервала  $h$  была вычислена по формуле:

$$h = \frac{x_{max} - x_{min}}{k} = \frac{177 - 64.5}{7} \approx 16.1$$

В таблице 6 представлен полученный интервальный ряд.

Таблица 6

<b>Границы интервалов</b>	<b>Средины интервалов</b>	<b>Абсолютная частота</b>	<b>Относительная частота</b>
[64.5, 80.6)	72.55	2	0.019
[80.6, 96.7)	88.65	5	0.048
[96.7, 112.8)	104.75	15	0.144
[112.8, 128.9)	120.85	27	0.26
[128.9, 145.0)	136.95	27	0.26
[145.0, 161.1)	153.05	21	0.202
[161.1, 177.0)	169.05	7	0.067

- Графики для интервального ряда абсолютных частот

Полигон представлен на рис. 1.

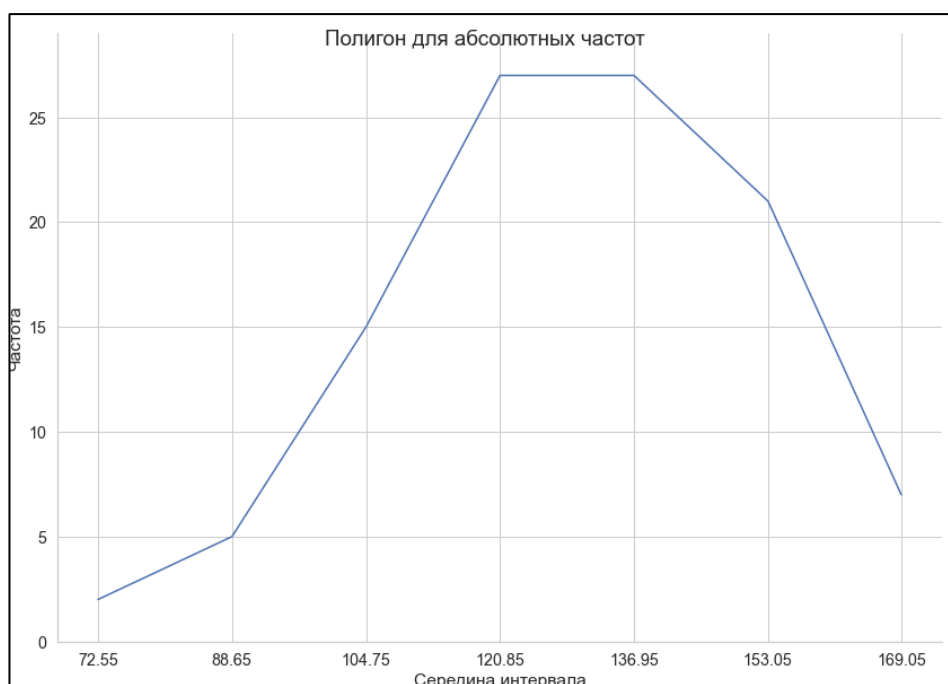


Рисунок 1 – Полигон для абсолютных частот

Гистограмма, представлена на рис. 2.

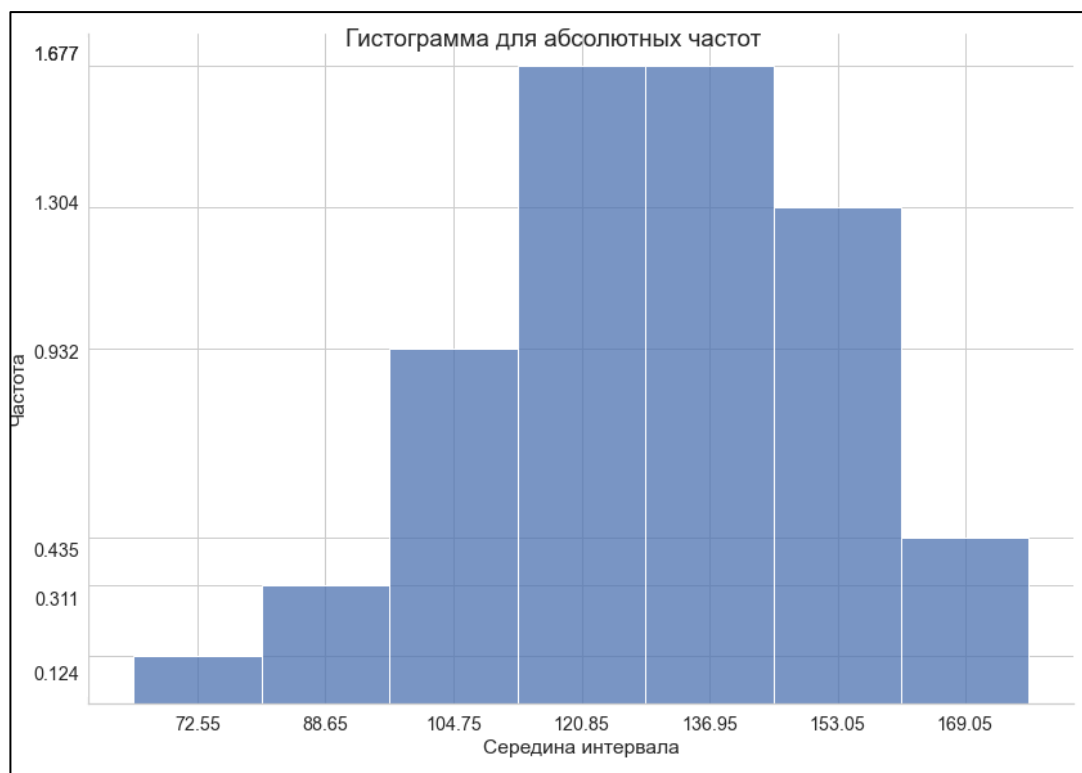


Рисунок 2 – Гистограмма для абсолютных частот

- Графики для интервального ряда относительных частот

Полигон представлен на рис. 3.

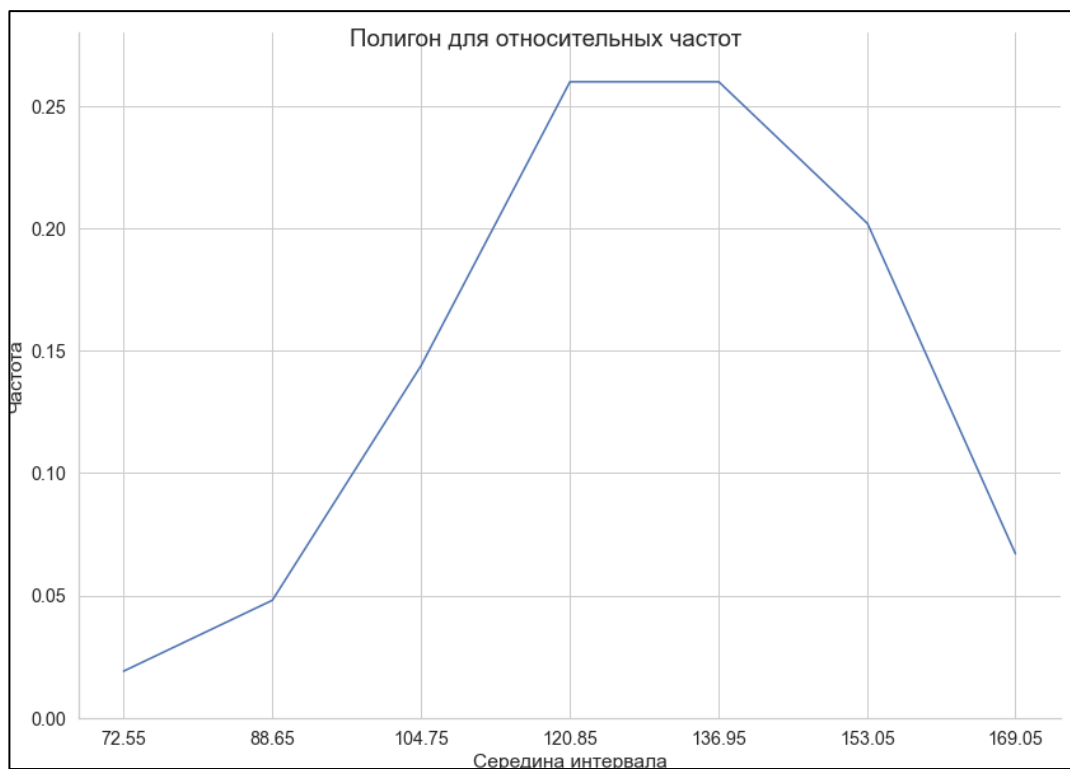


Рисунок 3 – Полигон для относительных частот

Гистограмма, представлена на рис. 4.

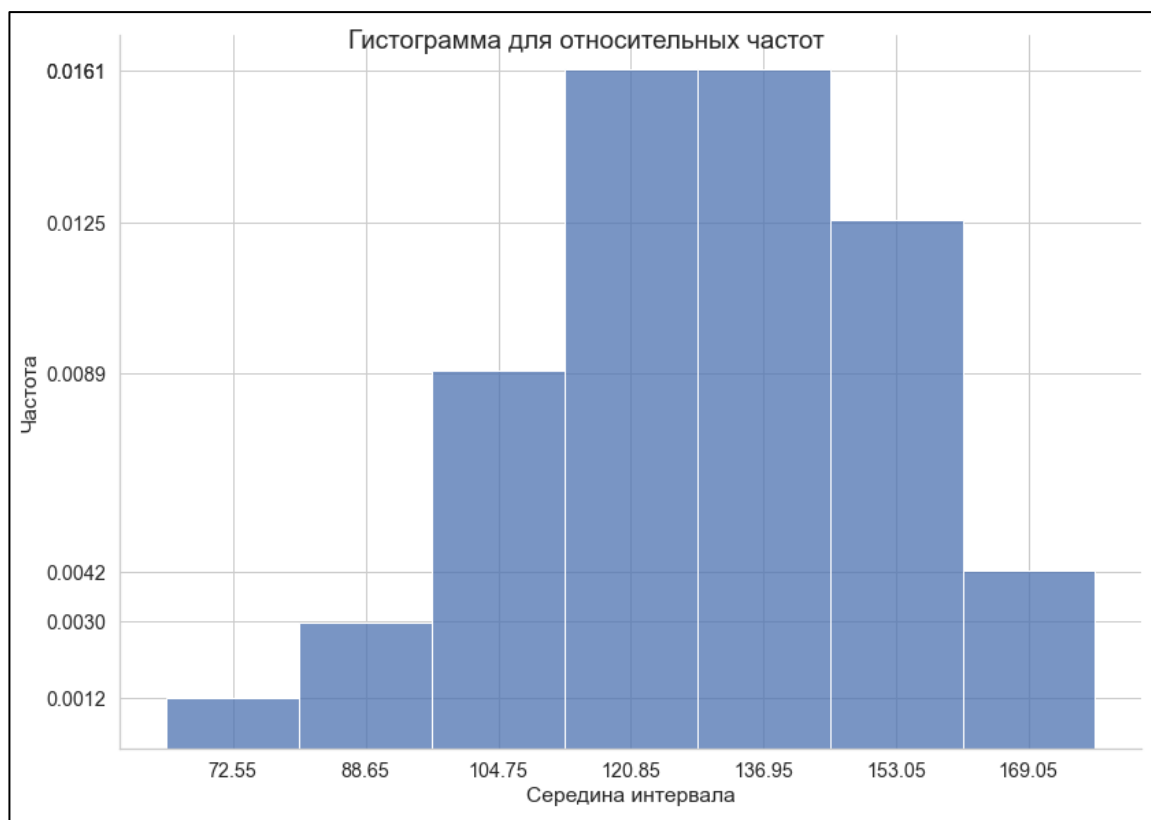


Рисунок 4 – Гистограмма для относительных частот

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

Функция распределения:

$$F(x) = \begin{cases} 0, & x = 72.55 \\ 0.019, & x = 88.650 \\ 0.067, & x = 104.75 \\ 0.211, & x = 120.85 \\ 0.471, & x = 136.95 \\ 0.731, & x = 153.05 \\ 0.933, & x = 169.05 \\ 1, & x > 169.05 \end{cases}$$

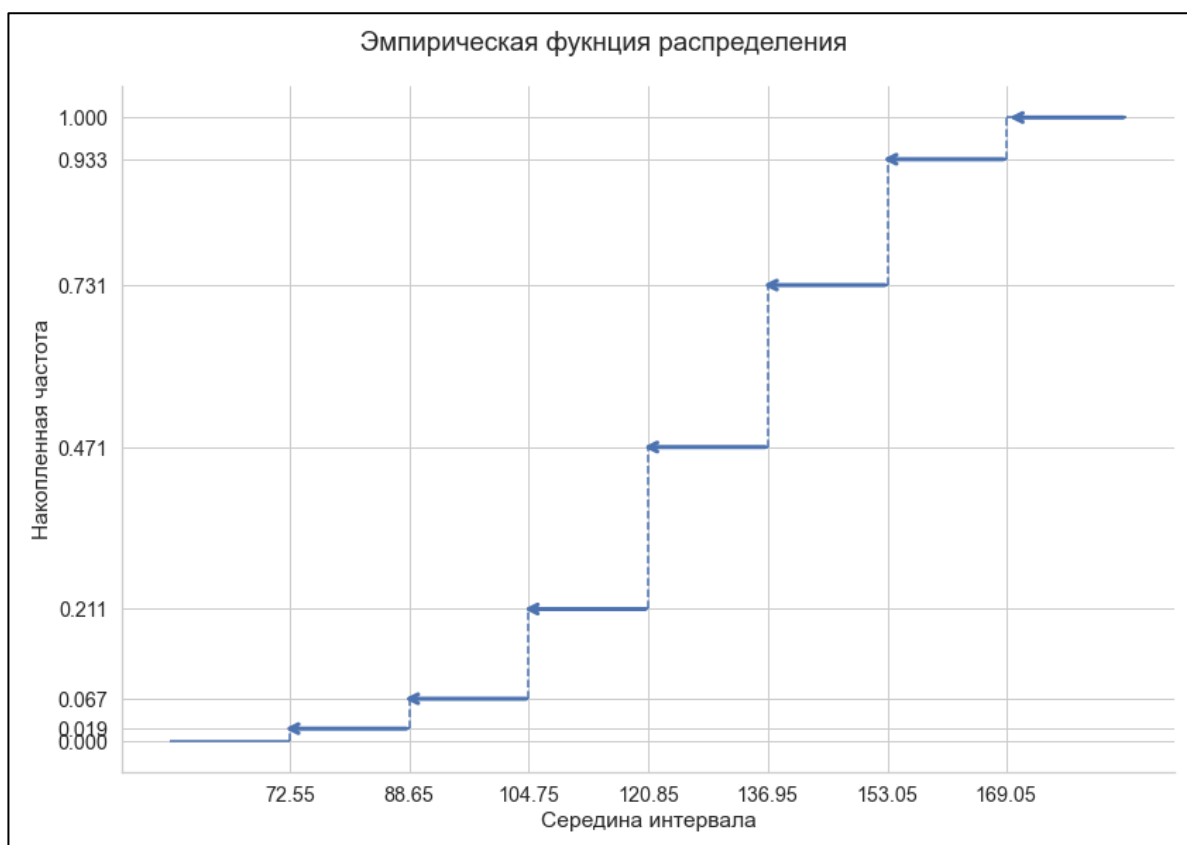


Рисунок 5 – График эмпирической функции распределения

Для интервального ряда были найдены середины интервалов и накопленные частоты. Интервальный ряд представлен в таблице 7.

Таблица 7

<b>Границы интервалов</b>	<b>Середины интервалов</b>	<b>Абсолютная частота</b>	<b>Относительная частота</b>	<b>Накопленная частота</b>
[64.5, 80.6)	72.55	2	0.019	0.019
[80.6, 96.7)	88.65	5	0.048	0.067
[96.7, 112.8)	104.75	15	0.144	0.211
[112.8, 128.9)	120.85	27	0.26	0.471
[128.9, 145.0)	136.95	27	0.26	0.731
[145.0, 161.1)	153.05	21	0.202	0.933
[161.1, 177.0)	169.05	7	0.067	1

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. 8.

Таблица 8

$v$	$n$	$u$	$n * u$	$n * u^2$	$n * u^3$	$n * u^4$	$n * (u + 1)^4$
72.55	0.019	-3	-0.057	0.171	-0.513	1.539	0.304
88.65	0.048	-2	-0.096	0.192	-0.384	0.768	0.048
104.75	0.144	-1	-0.144	0.144	-0.144	0.144	0
120.85	0.26	0	0	0	0	0	0.26
136.95	0.26	1	0.26	0.26	0.26	0.26	4.16
153.05	0.202	2	0.404	0.808	1.616	3.232	16.362
169.05	0.067	3	0.201	0.603	1.809	5.427	17.152
$\Sigma$	1	—	0.568	2.178	2.644	11.37	38.286

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

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

$$11.37 + 4 * 2.644 + 6 * 2.178 + 4 * 0.568 + 1 = 38.286$$

Эмпирические начальные и центральные моменты вычислены ниже:

$$\bar{x}_B = \overline{M}_1 = \overline{M}_1^* h + C = 129.9948$$

$$D_B = \overline{m}_2 = \left( \overline{M}_2^* - (\overline{M}_1^*)^2 \right) h^2 = 480.932$$

$$\overline{m}_3 = \left( \overline{M}_3^* - 3 \overline{M}_2^* \overline{M}_1^* + 2 (\overline{M}_1^*)^3 \right) h^3 = -2924.6818$$

$$\overline{m}_4 = \left( \overline{M}_4^* - 4 \overline{M}_3^* \overline{M}_1^* + 6 \overline{M}_2^* (\overline{M}_1^*)^2 + 2 (\overline{M}_1^*)^4 \right) h^4 = 622622.816$$

Найдем выборочное среднее и дисперсию с помощью стандартных формул.

Статистическая оценка математического ожидания:

$$\bar{x}_B = \frac{1}{N} \sum_{i=1}^k x_i n_i = 129.98$$

Статистическая оценка дисперсии:

$$D_B = \frac{1}{N} \sum_{i=1}^k (x_i - \bar{x}_B)^2 n_i = 481.8$$

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

$$s^2 = \frac{N}{N-1} D_B = \frac{104}{103} * 481.8 = 486.5$$

Статистические оценки СКО:

$$\sigma_B = \sqrt{D_B} = \sqrt{481.8} = 21.95$$

$$s = \sqrt{s^2} = \sqrt{486.5} = 22.06$$

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

Статистические оценки коэффициентов асимметрии и эксцесса можно вычислить по формулам:

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

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

Центральные эмпирические моменты третьего и четвертого порядков были найдены выше.

Статистическая оценка коэффициента асимметрии:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3} = -0.0000254$$

Статистическая оценка коэффициента эксцесса:

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

Коэффициент асимметрии отрицателен, следовательно, в данном случае это левосторонняя асимметрия, которая характеризуется удлинненным левым хвостом, а также неравенством  $\bar{x}_B < M_0$ , но полученное значение незначительно и скос распределения небольшой. Коэффициент эксцесса также отрицателен,

следовательно, эмпирическое распределение является более низким и пологим относительно нормального распределения.

### ***Двумерный интервальный вариационный ряд***

Был построен двумерный интервальный вариационный ряд. Двумерный интервальный ряд представлен в таблице 9.

*Таблица 9*

Y	X							n <sub>y</sub>
	338.5	375.5	412.5	449.5	486.5	523.5	559	
72.55	1	1	-	-	-	-	-	2
88.65	3	2	-	-	-	-	-	5
104.75	1	5	8	1	-	-	-	15
120.85	-	-	14	11	2	-	-	27
136.95	-	-	1	12	12	2	-	27
153.05	-	-	-	1	10	8	2	21
169.05	-	-	-	-	-	5	2	7
n <sub>x</sub>	5	8	23	25	24	15	4	n = 104

### ***Корреляционная таблица***

При вычислении выборочного коэффициента корреляции необходимо будет посчитать двойную сумму:

$$\sum_{i=1}^{K_y} \sum_{j=1}^{K_x} n_{ij} y_i x_j = \sum_{i=1}^{K_y} y_i \sum_{j=1}^{K_x} n_{ij} x_j = \sum_{j=1}^{K_x} x_j \sum_{i=1}^{K_y} n_{ij} y_i$$

Данные вычисления были произведены в корреляционной таблице, представленной в таблице 10.



Таблица 10

Y	X																				$X_i$	$y_iX_i$			
	338.5			375.5			412.5			449.5			486.5			523.5			559						
72.55		338.5			375.5																	714	51800.7		
		1			1			-			-			-			-			-					
	72.55			72.55																					
88.65		1015.5			751																	1428	156600.225		
		3			2			-			-			-			-			-					
	265.95			177.3																					
104.75		338.5			1877.5			3300			449.5											5965.5	624886.125		
		1			5			8			1			-			-			-					
	104.75			523.75			838			104.75															
120.85								5775			4944.5			973								11692.5	1413038.625		
		-			-			14			11			2			-			-					
							1691.9			1329.35			241.7												
136.95								412.5			5394			5838			1047					12691.5	1738100.925		
		-			-			1			12			12			2			-					
							136.95			1643.4			1643.4			273.9									
153.05											449.5			4865			4188			1118		10620.5	1625467.525		
		-			-			-			1			10			8			2					
											153.05			1530.5			1224.4			306.1					
169.05																		2617.5			1118		3735.5	631486.275	
		-			-			-			-			-				5			2				
																	845.25			338.1					
$Y_j$	443.25			773.6			2666.85			3230.55			3415.6			2343.55			644.2					6241380.4	
$x_jY_j$	15040.125			290486.8			1100075.625			1452132.225			1661689.4			1226848.425			360107.8					6241380.4	

$$\sum_{i=1}^{K_y} \sum_{j=1}^{K_x} n_{ij} y_i x_j = 6241380.4$$

Исходя из результатов корреляционной таблицы был вычислен выборочный коэффициент корреляции.

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y} \sum_{j=1}^{K_x} n_{ij} y_i x_j - N \bar{x}_B \bar{y}_B}{NS_x S_y} = \frac{6241380.4 - 104 * 453.71 * 129.98}{104 * 53.79 * 22.06} = 0.8765$$

Выборочный коэффициент корреляции отличен от нуля, следовательно  $X$  и  $Y$  коррелированы. Если случайные величины  $X$  и  $Y$  коррелированы, то они зависимы. Также  $\bar{r}_{xy}$  – положительный, следовательно можно сказать о положительной корреляционной зависимости, то есть, если  $X$  возрастает, то и  $Y$  возрастает.

Также выборочный коэффициент корреляции был посчитан с помощью условных вариантов. Вычисления представлены в таблице 11.

Таблица 11

Y	X								$X_i$	$y_i X_i$
	-3	-2	-1	0	1	2	3			
-3	-3	-2						-5	15	
	1	1	-	-	-	-	-			
	-3	-3								
-2	-9	-4						-13	26	
	3	2	-	-	-	-	-			
	-6	-4								
-1	-3	-10	-8	0				-21	21	
	1	5	8	1	-	-	-			
	-1	-5	-8	-1						
0			-14	0	2			-12	0	
	-	-	14	11	2	-	-			
			0	0	0					
1			1	0	12	4		15	15	
	-	-	1	12	12	2	-			
			1	12	12	2				
2				0	10	16	6	32	64	
	-	-	-	1	10	8	2			
				2	20	16	4			
3						10	6	16	48	
	-	-	-	-	-	5	2			
						15	6			
$Y_j$	-10	-12	-7	13	32	33	10		189	
$x_j Y_j$	30	24	7	0	32	66	30	189		

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y} \sum_{j=1}^{K_x} n_{ij} u_i v_j - N \bar{u}_B \bar{v}_B}{N S_u S_v} = \frac{189 - 104 * 0.115 * 0.567}{104 * 1.45 * 1.364} = 0.8758, \text{ где}$$

$u_B = 0.115, v_B = 0.567$  – условные средние для условных вариантов,

$S_u = 1.45, S_v = 1.364$  – несмещенные СКО условных вариантов

Коэффициенты корреляции, рассчитанные двумя способами, совпадают с точностью до сотых.

В случае нормального распределения системы случайных величин  $\{X; Y\}$  для оценки значения  $r_{xy}$ , если  $\bar{r}_{xy}$  – значим, можно использовать соотношение:

$$\bar{r}_{xy} - 3 \frac{1 - \bar{r}_{xy}^2}{\sqrt{N}} \leq r_{xy} \leq \bar{r}_{xy} + 3 \frac{1 + \bar{r}_{xy}^2}{\sqrt{N}}$$
$$0.8765 - 3 \frac{1 - 0.8765^2}{\sqrt{104}} \leq r_{xy} \leq 0.8765 + 3 \frac{1 + 0.8765^2}{\sqrt{104}}$$
$$0.8083 \leq r_{xy} \leq 1$$

### *Доверительный интервал для коэффициента корреляции*

С помощью преобразования Фишера перейдем к случайной величине  $z$ :

$$\bar{z} = 0.5 \ln \frac{1 + \bar{r}_{xy}}{1 - \bar{r}_{xy}} = 0.5 \ln \frac{1 + 0.8765}{1 - 0.8765} = 1.36$$

СКО распределения  $z$ :

$$\bar{\sigma}_z = \frac{1}{\sqrt{N - 3}} = \frac{1}{\sqrt{104 - 3}} = 0.0995$$

Доверительный интервал для генерального значения:

$$(\bar{z} - \lambda(\gamma)\bar{\sigma}_z; \bar{z} + \lambda(\gamma)\bar{\sigma}_z), \text{ где}$$

$$\Phi(\lambda(\gamma)) = \frac{\gamma}{2}$$

Тогда при уровне значимости  $\gamma = 0.95$ :

$$\Phi(\lambda(\gamma)) = 0.475$$

$$\lambda(\gamma) = 1.96$$

$$z \in (1.36 - 1.96 * 0.0995; 1.36 + 1.96 * 0.0995)$$

$$z \in (1.165; 1.555)$$

Для пересчёта интервала в доверительный интервал для коэффициента корреляции с тем же значением  $\gamma$  необходимо воспользоваться обратным преобразованием Фишера:

$$r = th(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z} - 1}{e^{2z} + 1}$$

$$\frac{e^{2z_1} - 1}{e^{2z_1} + 1} = 0.8227$$

$$\frac{e^{2z_2} - 1}{e^{2z_2} + 1} = 0.9146$$

Можно сделать вывод, что интервал (0.8227; 0.9146) с вероятностью (надежностью)  $\gamma = 0.95$  содержит в себе истинное значение коэффициента корреляции.

### ***Гипотеза о значимости выборочного коэффициента корреляции***

Поскольку  $\bar{r}_{xy}$  является случайной величиной, то это еще не значит, что  $r_{xy}$  тоже отличен от нуля. Проверим гипотезу  $H_0: r_{xy} = 0$ . Альтернативой будет гипотеза  $H_1: r_{xy} \neq 0$ .

В качестве критерия проверки гипотезы можно принять случайную величину:

$$T = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}}$$

При справедливости нулевой гипотезы случайная величина  $T$  распределена по закону Стьюдента с  $k = K - 2 = 5$  степенями свободы.

Найдено  $T_{\text{набл}}$  по формуле выше:

$$T_{\text{набл}} = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}} = \frac{0.8765 * \sqrt{102}}{\sqrt{1-0.8765^2}} = 18.388$$

По заданному уровню значимости  $\alpha = 0.05$  и значению  $k = N - 2 = 102$  из таблицы было определено значение  $t_{\text{крит}} = 1.985$

$$T_{\text{набл}} = 18.388$$

$$t_{\text{крит}} = 1.985$$

$|T_{\text{набл}}| > t_{\text{крит}}(\alpha, k)$  – основная гипотеза  $H_0$  должна быть отвергнута, это означает, что выборочный коэффициент корреляции  $\bar{r}_{xy}$  значительно отличается от нуля (значим).

### Выводы

В ходе выполнения лабораторной работы была проведена статистическая обработка выборки по второму признаку – были определены точечные оценки параметров распределения второго признака.

Построен двумерный интервальный вариационный ряд. С его помощью была построена корреляционная таблица, где была вычислена двойная сумма для выборочного коэффициента корреляции. Исходя из результатов корреляционной таблицы был вычислен выборочный коэффициент корреляции  $\bar{r}_{xy} = 0.8765$ . Выборочный коэффициент корреляции отличен от нуля, следовательно  $X$  и  $Y$  коррелированы и зависимы. Также  $\bar{r}_{xy}$  – положительный, следовательно можно сказать о положительной корреляционной зависимости.

Построен доверительный интервал для коэффициента корреляции при уровне значимости  $\gamma = 0.95$ . Можно сделать вывод, что интервал  $(0.8227; 0.9146)$  с вероятностью (надежностью)  $\gamma = 0.95$  содержит в себе истинное значение коэффициента корреляции.

Проведена проверка статистической гипотезы о равенстве коэффициента корреляции нулю при уровне значимости  $\alpha = 0.05$ . Было выяснено, что  $|T_{\text{набл}}| > t_{\text{крит}}(\alpha, k)$ , то есть основная гипотеза  $H_0$  должна быть отвергнута, это означает, что выборочный коэффициент корреляции  $\bar{r}_{xy}$  значительно отличается от нуля (значим).

## ПРИЛОЖЕНИЕ А

### ИСХОДНЫЙ КОД

```
#!/usr/bin/env python
```

```
# coding: utf-8
```

```
# In[1]:
```

```
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)
```

```
# In[2]:
```

```
df = pd.read_csv('c:/Users/gandh/dev/unv/smoed/me/data/main_data.csv')
```

```
X = df['nu']
```

```
Y = df['E']
```

```
# In[3]:
```

```
h1, h2 = 37, 16.1
```

```
ivs_X = np.hstack((np.arange(min(X), max(X), h1), np.array(max(X))))
```

```
ivs_Y = np.hstack((np.arange(min(Y), max(Y), h2), np.array(max(Y))))
```

```
# ## Двумерный интервальный ряд
```

```
# In[4]:
```

```
df_int = df.copy()
df_int['intX'] = pd.cut(df_int['nu'], bins=ivs_X, right=False)
df_int['intXl'] = pd.cut(df_int['nu'], bins=ivs_X,
                        labels=[1,2,3,4,5,6,7], right=False)
df_int['intY'] = pd.cut(df_int['E'], bins=ivs_Y, right=False)
df_int['intYl'] = pd.cut(df_int['E'], bins=ivs_Y,
                        labels=[1,2,3,4,5,6,7], right=False)
```

```
# In[5]:
```

```
df_int.iloc[64, 4] = df_int.iloc[63, 4]
df_int.iloc[64, 5] = df_int.iloc[63, 5]
df_int.iloc[97, 2] = df_int.iloc[99, 2]
df_int.iloc[97, 3] = df_int.iloc[99, 3]
# df_int['intXl'].value_counts().sort_index()
# df_int['intYl'].value_counts().sort_index()
# df_int.sort_values(by=['nu'], ignore_index = True).head()
# df_int.value_counts(['intYl', 'intXl']).sort_index()
```

```
# ## Корреляционная таблица
```

```
# In[6]:
```

```
N = 104
xv = 453.71
```

```

sx = 53.79
yv = 129.98
sy = 22.06

```

```

# In[9]:

```

```

df_kor = pd.DataFrame(columns=['yi', 'x1', 'x2', 'x3', 'x4', 'x5', 'x6', 'x7', 'Xi', 'yX'])
df_kor['yi'] =
[ np.NaN, 72.55, 88.65, 104.75, 120.85, 136.95, 153.05, 169.05, np.NaN, np.NaN]
df_kor['x1'] = [ 338.5, 1, 3, 1, 0, 0, 0, 0, np.NaN, np.NaN]
df_kor['x2'] = [ 375.5, 1, 2, 5, 0, 0, 0, 0, np.NaN, np.NaN]
df_kor['x3'] = [ 412.5, 0, 0, 8, 14, 1, 0, 0, np.NaN, np.NaN]
df_kor['x4'] = [ 449.5, 0, 0, 1, 11, 12, 1, 0, 0, np.NaN]
df_kor['x5'] = [ 486.5, 0, 0, 0, 2, 12, 10, 0, np.NaN, np.NaN]
df_kor['x6'] = [ 523.5, 0, 0, 0, 0, 2, 8, 5, np.NaN, np.NaN]
df_kor['x7'] = [ 559, 0, 0, 0, 0, 0, 2, 2, np.NaN, np.NaN]
# df_kor['yi']*df_kor['x1']
df_curr1 = pd.DataFrame()
df_curr2 = pd.DataFrame()
for i in range(7):
    df_curr1[i] = df_kor.iloc[0,1:8]*df_kor.iloc[i+1,1:8]
    df_kor.loc[i+1, 'Xi'] =
np.dot(df_kor.iloc[0,1:8], df_kor.iloc[i+1,1:8])
    df_curr2[i] = df_kor.iloc[1:8,0]*df_kor.iloc[1:8,i+1]
    df_kor.iloc[8,i+1] = np.dot(df_kor.iloc[1:8,0], df_kor.iloc[1:8,i+1])

df_kor['yX'] = df_kor['yi']*df_kor['Xi']
df_kor.iloc[9,:] = df_kor.iloc[0,:]*df_kor.iloc[8,:]
df_kor.loc[8, 'yX'] = df_kor['yX'].sum()
df_kor.loc[9, 'Xi'] = df_kor.iloc[9,:].sum()

```



```
df_curr1.transpose() # желт
df_curr2 # зелен
df_kor
```

```
# ### Коэффициент корреляции
```

```
# In[14]:
```

```
r = ((df_kor.loc[8, 'yX'] - N*xv*yv)/(N*sx*sy)).round(4)
r
```

```
# ### Оценка кк
```

```
# In[20]:
```

```
((r-3*((1-r**2)/np.sqrt(N))).round(4),
(r+3*((1+r**2)/np.sqrt(N))).round(4))
```

```
# ## Доверительный интервал для кк
```

```
# In[31]:
```

```
z = (0.5*np.log((1+r)/(1-r))).round(3)
z
```

```
# In[30]:
```

```
sz = (1/np.sqrt(N-3)).round(4)
```

```
sz
```

```
# In[35]:
```

```
gamma = 0.95
```

```
F = gamma/2
```

```
l = 1.96
```

```
z1 = (z-l*sz).round(4)
```

```
z2 = (z+l*sz).round(4)
```

```
(z1,z2)
```

```
# In[38]:
```

```
r1 = ((np.exp(2*z1)-1)/(np.exp(2*z1)+1)).round(4)
```

```
r2 = ((np.exp(2*z2)-1)/(np.exp(2*z2)+1)).round(4)
```

```
(r1, r2)
```

```
# ## Гипотеза о значимости выборочного коэффициента корреляции
```

```
# In[50]:
```

```
K = 7
```

```
Tn = ((r*np.sqrt(N-2))/np.sqrt(1-r**2)).round(3)
```

```
tk = 1.985
```

```
# In[51]:
```

```
'True' if np.abs(Tn) <= tk else 'False'
```

```
# In[ ]:
```