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	____________________	Бабенко Н.С.
Студент гр. 8383	____________________	Сахаров В.М.
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:

1. Функциональная зависимость:

$$y = g(x)$$

2. Статистическая зависимость:

$$\phi(x/y) = \frac{f(x, y)}{f_2(y)}; \phi(y/x) = \frac{f(x, y)}{f_1(x)}$$

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.

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

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

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

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

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:

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

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

Для пересчёта интервала в доверительный интервал для коэффициента корреляции с тем же значением γ 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 $H_0: r_{xy} = 0$. Гипотеза $H_1: r_{xy} \ne 0$. 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 $k = N - 2$ 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

1. 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.

2. Построить двумерный интервальный вариационный ряд, оформить в виде таблицы.

3. По полученному двумерному интервальному вариационному ряду построить корреляционную таблицу, сделать выводы.

4. Исходя из результатов корреляционной таблицы вычислить статистическую оценку корреляционного момента.

5. Вычислить коэффициент корреляции, сделать выводы.

6. Построить доверительный интервал для коэффициента корреляции при уровне значимости $\gamma \in \{0.95, 0.99\}$, сделать выводы.

7. Осуществить проверку статистической гипотезы о равенстве коэффициента корреляции нулю при заданном уровне значимости $\alpha = 0.05$, сделать выводы.

Выполнение работы

- Статистическая обработка второй выборки

Выборка, сформированная из генеральной совокупности, представлена в таблице 1. Объём выборки: 100.

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 для переменной E представлена в таблице 2.

Таблица 2

i	y_i	i	y_i	i	y_i	i	y_i	i	y_i
1	135.2	***21***	131.4	***41***	159.3	***61***	122.3	***81***	143.6
2	124.7	***22***	103.8	***42***	149.8	***62***	128.9	***82***	144.4
3	147.9	***23***	154.9	***43***	132.5	***63***	122.8	***83***	175.7
4	140.9	***24***	117.7	***44***	94.6	***64***	140.7	***84***	131.3
5	168.5	***25***	145.3	***45***	135.6	***65***	117.7	***85***	112.1
6	147.3	***26***	126.7	***46***	169.6	***66***	112.9	***86***	146.1
7	136.6	***27***	114.8	***47***	142.4	***67***	126.4	***87***	86.1
8	139.6	***28***	109.3	***48***	116.7	***68***	143.2	***88***	132.5
9	84.9	***29***	118.6	***49***	164.0	***69***	125.0	***89***	122.4
10	117.9	***30***	127.7	***50***	142.0	***70***	119.0	***90***	104.7
11	153.5	***31***	117.5	***51***	187.4	***71***	137.2	***91***	111.7
12	122.9	***32***	131.5	***52***	152.6	***72***	140.7	***92***	148.5
13	140.4	***33***	136.8	***53***	126.6	***73***	137.5	***93***	144.0
14	108.6	***34***	98.8	***54***	122.2	***74***	128.4	***94***	115.1
15	157.4	***35***	160.6	***55***	115.9	***75***	136.4	***95***	110.5
16	119.2	***36***	148.2	***56***	110.0	***76***	140.6	***96***	137.7
17	129.4	***37***	122.6	***57***	105.8	***77***	122.3	***97***	139.9
18	138.6	***38***	158.7	***58***	130.3	***78***	195.7	***98***	141.2
19	89.0	***39***	137.9	***59***	169.8	***79***	141.2	***99***	129.2
20	91.4	***40***	148.3	***60***	130.3	***80***	119.7	***100***	145.4

В таблице 3 представлено преобразование выборки в ранжированный ряд.

Таблица 3

i	y_i	i	y_i	i	y_i	i	y_i	i	y_i
1	84.9	***21***	117.5	***41***	127.7	***61***	137.9	***81***	147.3
2	86.1	***22***	117.7	***42***	128.4	***62***	138.6	***82***	147.9
3	89.0	***23***	117.7	***43***	128.9	***63***	139.6	***83***	148.2
4	91.4	***24***	117.9	***44***	129.2	***64***	139.9	***84***	148.3
5	94.6	***25***	118.6	***45***	129.4	***65***	140.4	***85***	148.5

6	98.8	***26***	119.0	***46***	130.3	***66***	140.6	***86***	149.8
7	103.8	***27***	119.2	***47***	130.3	***67***	140.7	***87***	152.6
8	104.7	***28***	119.7	***48***	131.3	***68***	140.7	***88***	153.5
9	105.8	***29***	122.2	***49***	131.4	***69***	140.9	***89***	154.9
10	108.6	***30***	122.3	***50***	131.5	***70***	141.2	***90***	157.4
11	109.3	***31***	122.3	***51***	132.5	***71***	141.2	***91***	158.7
12	110.0	***32***	122.4	***52***	132.5	***72***	142.0	***92***	159.3
13	110.5	***33***	122.6	***53***	135.2	***73***	142.4	***93***	160.6
14	111.7	***34***	122.8	***54***	135.6	***74***	143.2	***94***	164.0
15	112.1	***35***	122.9	***55***	136.4	***75***	143.6	***95***	168.5
16	112.9	***36***	124.7	***56***	136.6	***76***	144.0	***96***	169.6
17	114.8	***37***	125.0	***57***	136.8	***77***	144.4	***97***	169.8
18	115.1	***38***	126.4	***58***	137.2	***78***	145.3	***98***	175.7
19	115.9	***39***	126.6	***59***	137.5	***79***	145.4	***99***	187.4
20	116.7	***40***	126.7	***60***	137.7	***80***	146.1	***100***	195.7

Видно, что $y_{min} = 84.9$, а $y_{max} = 195.7$

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

Таблица 4

i	y_i	n_i	$\overline{n_i}$	i	y_i	n_i	$\overline{n_i}$	i	y_i	n_i	$\overline{n_i}$	i	y_i	n_i	$\overline{n_i}$
1	84.9	1	0.01	***26***	119.2	1	0.01	***51***	136.4	1	0.01	***76***	147.9	1	0.01
2	86.1	1	0.01	***27***	119.7	1	0.01	***52***	136.6	1	0.01	***77***	148.2	1	0.01
3	89.0	1	0.01	***28***	122.2	1	0.01	***53***	136.8	1	0.01	***78***	148.3	1	0.01
4	91.4	1	0.01	***29***	122.3	2	0.02	***54***	137.2	1	0.01	***79***	148.5	1	0.01
5	94.6	1	0.01	***30***	122.4	1	0.01	***55***	137.5	1	0.01	***80***	149.8	1	0.01
6	98.8	1	0.01	***31***	122.6	1	0.01	***56***	137.7	1	0.01	***81***	152.6	1	0.01
7	103.8	1	0.01	***32***	122.8	1	0.01	***57***	137.9	1	0.01	***82***	153.5	1	0.01
8	104.7	1	0.01	***33***	122.9	1	0.01	***58***	138.6	1	0.01	***83***	154.9	1	0.01
9	105.8	1	0.01	***34***	124.7	1	0.01	***59***	139.6	1	0.01	***84***	157.4	1	0.01
10	108.6	1	0.01	***35***	125.0	1	0.01	***60***	139.9	1	0.01	***85***	158.7	1	0.01
11	109.3	1	0.01	***36***	126.4	1	0.01	***61***	140.4	1	0.01	***86***	159.3	1	0.01
12	110.0	1	0.01	***37***	126.6	1	0.01	***62***	140.6	1	0.01	***87***	160.6	1	0.01
13	110.5	1	0.01	***38***	126.7	1	0.01	***63***	140.7	2	0.02	***88***	164.0	1	0.01
14	111.7	1	0.01	***39***	127.7	1	0.01	***64***	140.9	1	0.01	***89***	168.5	1	0.01

15	112.1	1	0.01	***40***	128.4	1	0.01	***65***	141.2	2	0.02	***90***	169.6	1	0.01
16	112.9	1	0.01	***41***	128.9	1	0.01	***66***	142.0	1	0.01	***91***	169.8	1	0.01
17	114.8	1	0.01	***42***	129.2	1	0.01	***67***	142.4	1	0.01	***92***	175.7	1	0.01
18	115.1	1	0.01	***43***	129.4	1	0.01	***68***	143.2	1	0.01	***93***	187.4	1	0.01
19	115.9	1	0.01	***44***	130.3	2	0.02	***69***	143.6	1	0.01	***94***	195.7	1	0.01
20	116.7	1	0.01	***45***	131.3	1	0.01	***70***	144.0	1	0.01				
21	117.5	1	0.01	***46***	131.4	1	0.01	***71***	144.4	1	0.01				
22	117.7	2	0.02	***47***	131.5	1	0.01	***72***	145.3	1	0.01				
23	117.9	1	0.01	***48***	132.5	2	0.02	***73***	145.4	1	0.01				
24	118.6	1	0.01	***49***	135.2	1	0.01	***74***	146.1	1	0.01				
25	119.0	1	0.01	***50***	135.6	1	0.01	***75***	147.3	1	0.01				

Количество интервалов разбиения вычислено с помощью формулы Стерджесса:

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

Ширина интервала:

$$h = \frac{y_{max} - y_{min}}{k} = \frac{195.7 - 84.9}{7} = 16$$

В таблице 5 представлен полученный интервальный ряд.

Таблица 5

Границы интервалов	*Середины интервалов*	*Абсолютная частота*	*Относительная частота*
[84.9, 100.9)	92.9	6	0.06
[100.9, 116.9)	108.9	14	0.14
[116.9, 132.9)	124.9	32	0.32
[132.9, 148.9)	140.9	33	0.33
[148.9, 164.9)	156.9	9	0.09
[164.9, 180.9)	172.9	4	0.04
[180.9, 195.7)	188.3	2	0.02

Далее для интервального ряда абсолютных частот были построены полигон и гистограмма. Полигон представлен на рис. 1.

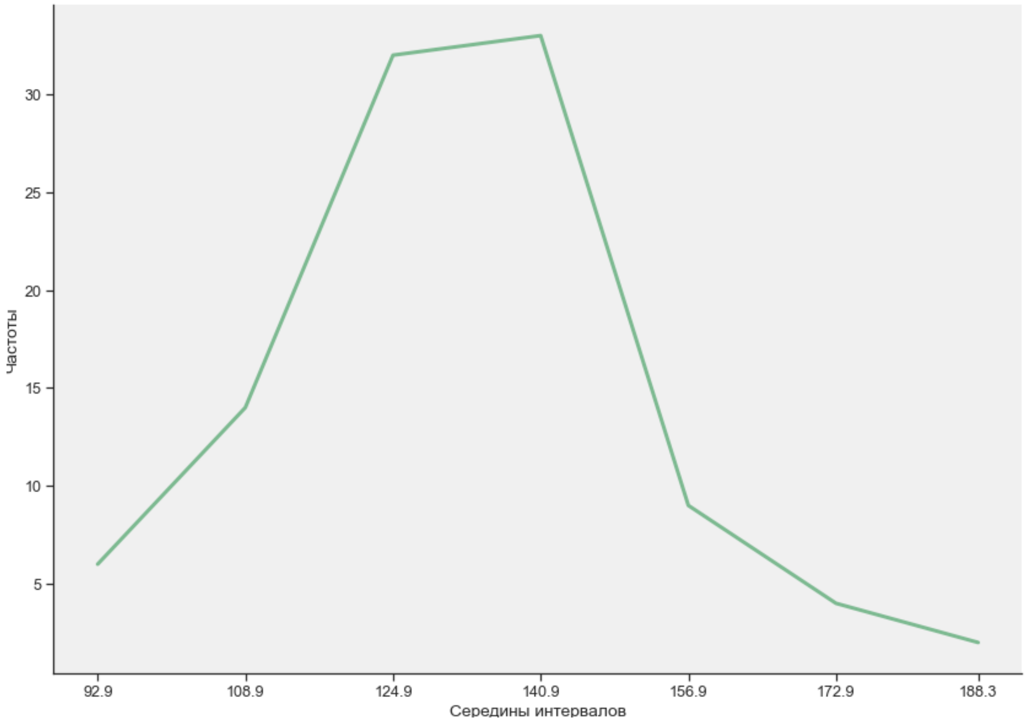

Рисунок 1 – Полигон для абсолютных частот

Полигон представляет собой ломаную, соединяющую точки, соответствующие срединным значениям интервалов и абсолютным частотам этих интервалов. Гистограмма, представлена на рис. 2.

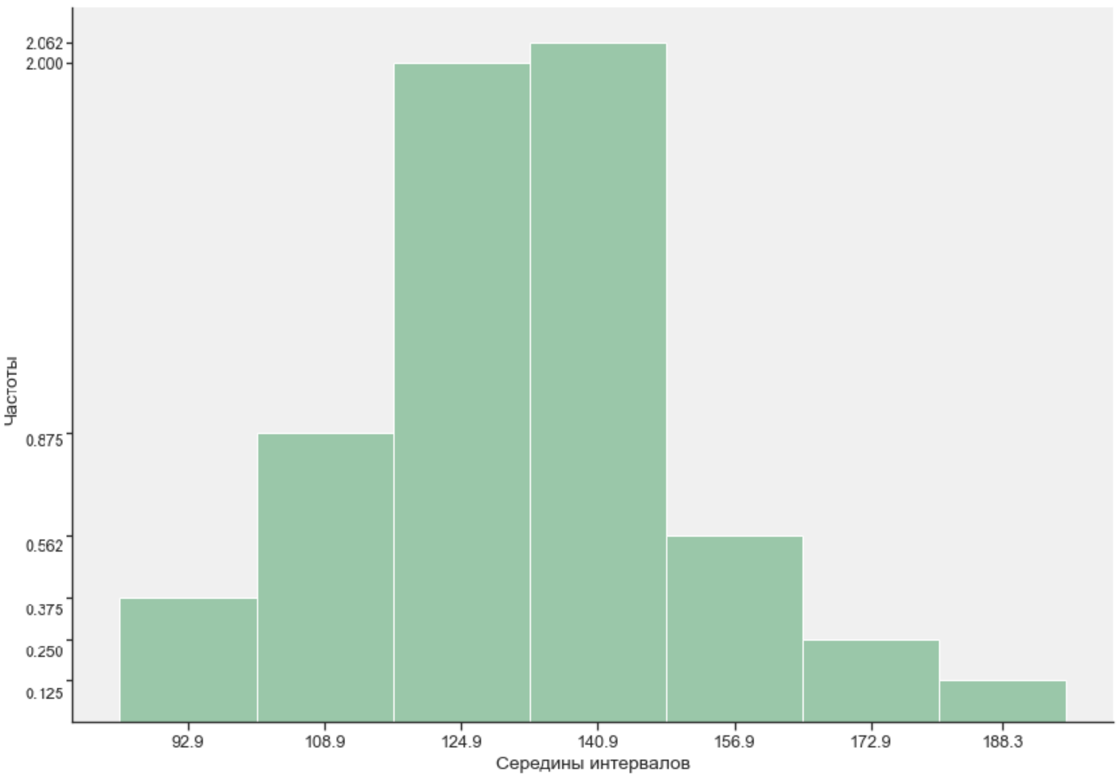

Рисунок 2 – Гистограмма для абсолютных частот

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. Эмпирическая функция распределения, построенная применительно к интервальному ряду для относительных частот представлен на рис. 3.

Функция распределения:

$$F(92.9) = 0$$
$$F(108.9) = 0.06$$
$$F(124.9) = 0.20$$
$$F(140.9) = 0.52$$
$$F(156.9) = 0.85$$
$$F(172.9) = 0.94$$
$$F(188.3) = 0.98$$

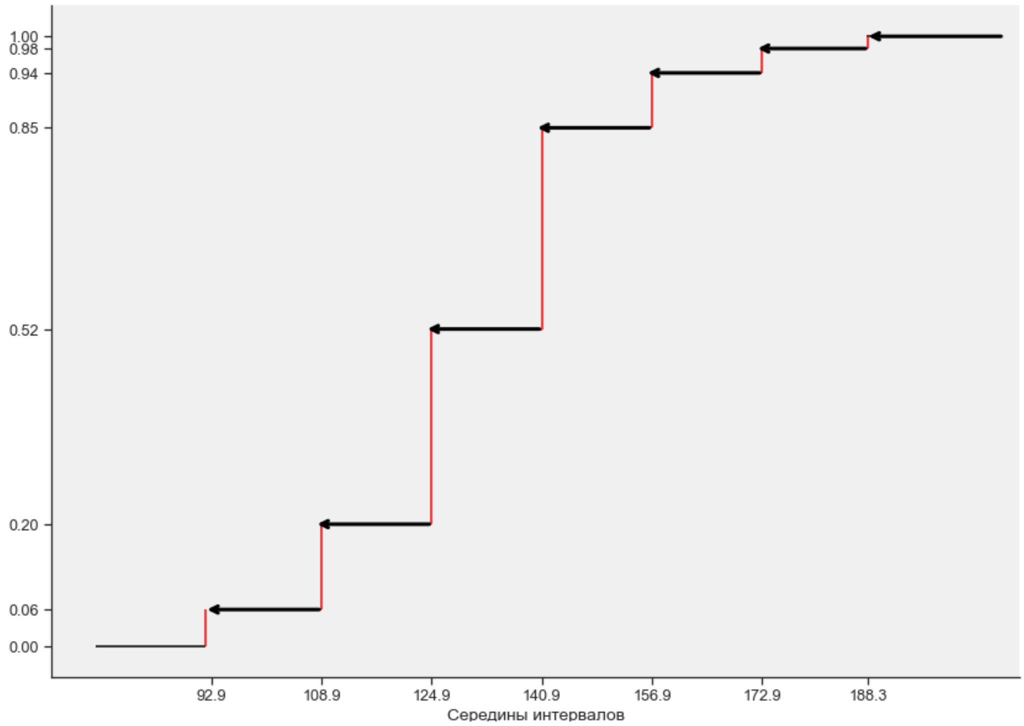

Рисунок 3 – График эмпирической функции распределения

Полигон для относительных частот представлен на рис. 4.

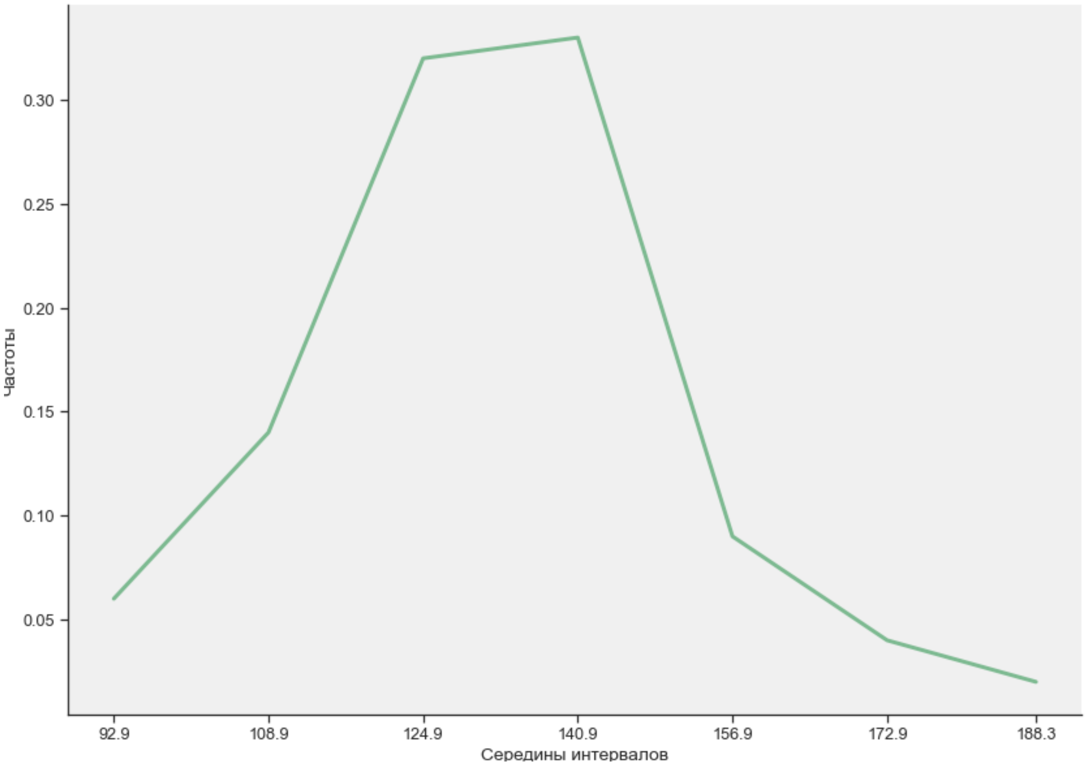

Рисунок 4 – Полигон для относительных частот

Гистограмма для относительных частот, представлена на рис. 5.

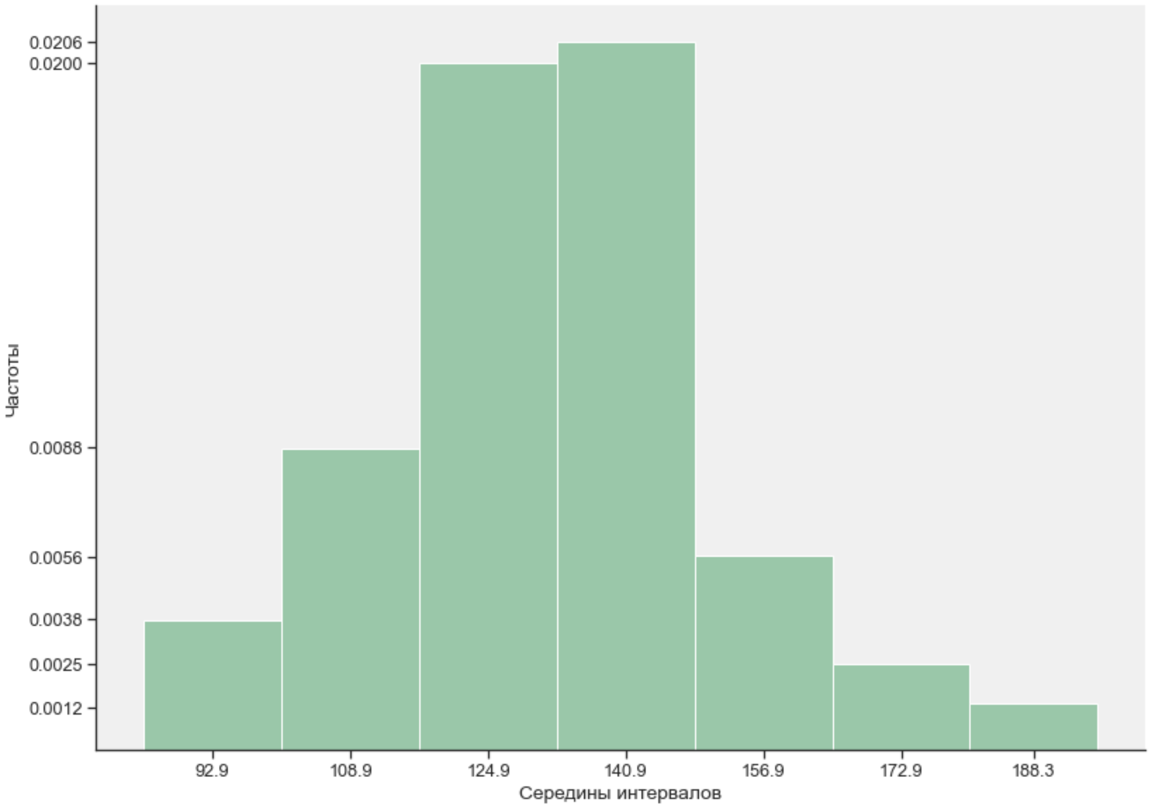

Рисунок 5 – Гистограмма для относительных частот

Интервальный ряд для переменной E и с посчитанными накопленными частотами представлен в таблице 6.

Таблица 6

Границы интервалов	*Середины интервалов*	*Абсолютная частота*	*Относительная частота*	*Накопленная частота*
[84.9, 100.9)	92.9	6	0.06	0.06
[100.9, 116.9)	108.9	14	0.14	0.2
[116.9, 132.9)	124.9	32	0.32	0.52
[132.9, 148.9)	140.9	33	0.33	0.85
[148.9, 164.9)	156.9	9	0.09	0.94
[164.9, 180.9)	172.9	4	0.04	0.98
[180.9, 195.7)	188.3	2	0.02	1

Результаты вычислений условных моментов представлены в табл. 7.

Таблица 7

$\boldsymbol{v}$	$\boldsymbol{n}$	$\boldsymbol{u}$	$\boldsymbol{n*u}$	$\boldsymbol{n*u^2}$	$\boldsymbol{n*u^3}$	$\boldsymbol{n*u^4}$	$\boldsymbol{n*(u+1)^4}$
92.9	0.06	-3	-0.18	0.54	-1.62	4.86	0.96
108.9	0.14	-2	-0.28	0.56	-1.12	2.24	0.14
124.9	0.32	-1	-0.32	0.32	-0.32	0.32	0.0
140.9	0.33	0	0.0	0.0	0.0	0.0	0.33
156.9	0.09	1	0.09	0.09	0.09	0.09	1.44
172.9	0.04	2	0.08	0.16	0.32	0.64	3.24
188.3	0.02	3	0.06	0.18	0.54	1.62	5.12
$\sum$	1	–	-0.55	1.85	-2.11	9.77	11.23

Проверим вычисления с помощью последнего столбца:

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

$$= 9.77 + 4 * -2.11 + 6 * 1.85 + 4 * -0.55 + 1 = 11.23$$

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

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

$$\overline{x_{\text{в}}} = \overline{M_1} = \overline{M_1^*}h + C = 132.1$$

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

$$D_{\text{в}} = \overline{m_2} = \left(\overline{M_2^*} - \left(\overline{M_1^*}\right)^2\right)h^2 = 395.16$$

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

$$\overline{x_{\text{в}}} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = 132.09$$

$$D_{\text{в}} = \frac{1}{N}\sum_{i=1}^{k} (x_i - \overline{x_{\text{в}}})^2\, n_i = 394.8$$

Исправленная оценка дисперсии:

$$s^2 = \frac{N}{N-1}D_{\text{в}} = \frac{100}{99} * 394.8 = 398.79$$

Были найдены статистические оценки СКО:

$$\sigma_{\text{в}} = \sqrt{D_{\text{В}}} = \sqrt{394.8} = 19.87$$

$$s = \sqrt{s^2} = \sqrt{398.79} = 19.97$$

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

Были найдены статистические оценки коэффициентов асимметрии и эксцесса:

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

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

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

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

Статистическая оценка коэффициента асимметрии:

$$\overline{A_s} = \frac{\overline{m_3}}{s^3} = 0.000039$$

Статистическая оценка коэффициента эксцесса:

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

Коэффициент асимметрии положительный – это правосторонняя асимметрия. Коэффициент эксцесса отрицательный – пик распределения около математического ожидания гладкий.

- Двумерный интервальный вариационный ряд

В таблице 8 представлен построенный двумерный интервальный вариационный ряд (корреляционная таблица).

Таблица 8

Y	**X**							
	343	387	431	475	519	563	604	n_y
92.9	3	3	0	0	0	0	0	6
108.9	1	5	6	2	0	0	0	14
124.9	0	1	18	12	1	0	0	32
140.9	0	0	3	20	9	1	0	33
156.9	0	0	0	1	7	1	0	9
172.9	0	0	0	0	0	4	0	4
188.3	0	0	0	0	0	0	2	2
$\mathbf{n_x}$	4	9	27	35	17	6	2	100

Как видно из таблицы суммы частот по столбцам совпадают с абсолютными частотами интервального вариационного ряда по признаку nu, то же самое можно сказать и для строк (переменная E), таблица составлена корректно.

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

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

Чтобы удобно посчитать двойную сумму, можно воспользоваться преобразованием ниже, данные вычисления представлены в таблице 9.

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

Таблица 9

Y	***X***								
	343	*387*	*431*	*475*	*519*	*563*	*604*	X_i	y_iX_i
92.9	1029 **3** 278.7	1161 **3** 278.7						2190	203451
108.9	343 **1** 108.9	1935 **5** 544.5	2586 **6** 653.4	950 **2** 217.8				5814	633144.6
124.9		387 **1** 124.9	7758 **18** 2248.2	5700 **12** 1498.8	519 **1** 124.9			14364	1794063.6
140.9			1293 **3** 422.7	9500 **20** 2818	4671 **9** 1268.1	563 **1** 140.9		16027	2258204.3
156.9				475 **1** 156.9	3633 **7** 1098.3	563 **1** 156.9		4671	732879.9
172.9						2252 **4** 691.6		2252	389370.8
188.3							1208 **2** 376.6	1208	227466.4
$\boldsymbol{Y_j}$	387.6	948.1	3324.3	4691.5	2491.3	989.4	376.6	**6238580.6**	
$\boldsymbol{x_jY_j}$	132946.8	366914.7	1432773.3	2228462.5	1292984.7	557032.2	227466.4		

Вычислен выборочный коэффициент корреляции:

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}y_ix_j - N\bar{x}_{\text{в}}\bar{y}_{\text{в}}}{NS_xS_y} = \frac{6238580.6 - 100 * 465.26 * 132.09}{100 * 54.57 * 19.97} = 0.853$$

Выборочный коэффициент корреляции не равен нулю и положителен, значит X и Y коррелированы и зависимы, а также это положительная корреляционная зависимость.

Также по аналогии было посчитано значение выборочного коэффициента корреляции с помощью условных вариант.

$$\bar{r}_{xy} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}u_i v_j - N\overline{u_{\text{в}}}\overline{v_{\text{в}}}}{NS_u S_v} = \frac{145 - 100 * -0.2115 * -0.529}{100 * 1.214 * 1.224} = 0.8525$$

Коэффициенты корреляции, рассчитанные с помощью основной формулы и условных вариант совпали.

Оценим значение r_{xy} в случае нормального распределения:

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

$$0.853 - 3\frac{1 - 0.853^2}{\sqrt{100}} \le r_{xy} \le 0.853 + 3\frac{1 + 0.853^2}{\sqrt{100}}$$

$$0.7713 \le r_{xy} \le 1$$

- Доверительный интервал для коэффициента корреляции

  Построим доверительный интервал для коэффициента корреляции. Перейдём к случайной величине z:

$$\bar{z} = 0.5\ln\frac{1 + \bar{r}_{xy}}{1 - \bar{r}_{xy}} = 0.5\ln\frac{1 + 0.853}{1 - 0.853} = 1.267$$

Среднеквадратическое отклонение:

$$\bar{\sigma}_z = \frac{1}{\sqrt{N - 3}} = \frac{1}{\sqrt{100 - 3}} = 0.1015$$

Доверительный интервал:

$$(\bar{z} - \lambda(\gamma)\bar{\sigma}_z; \bar{z} + \lambda(\gamma)\bar{\sigma}_z), \Phi\big(\lambda(\gamma)\big) = \frac{\gamma}{2}$$

При уровне значимости $\gamma = 0.99$:

$$\Phi\big(\lambda(\gamma)\big) = 0.495 \Rightarrow \lambda(\gamma) = 2.58$$

$$(1.267 - 2.58 * 0.1015; 1.267 + 2.58 * 0.1015)$$

$$(1.0051; 1.5289)$$

Для построения доверительного интервала для коэффициента корреляции воспользуемся обратным преобразованием Фишера:

$$r_{xy} \in \left(\frac{e^{2z_l}-1}{e^{2z_l}+1}; \frac{e^{2z_r}-1}{e^{2z_r}+1}\right)$$

$$\frac{e^{2z_l}-1}{e^{2z_l}+1} = 0.7637; \frac{e^{2z_r}-1}{e^{2z_r}+1} = 0.9102$$

Доверительный интервал $(0.7637; 0.9102)$ покрывает истинное значение коэффициента корреляции с надежностью $\gamma = 0.99$.

- Гипотеза о равенстве коэффициента корреляции нулю

Проверим гипотезу $H_0: r_{xy} = 0$; $H_1: r_{xy} \neq 0$.

В качестве критерия проверки гипотезы примем случайную величину:

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

Найдём $T_{\text{набл}}$ по формуле:

$$T_{\text{набл}} = \frac{\bar{r}_{xy}\sqrt{N-2}}{\sqrt{1-\bar{r}_{xy}^2}} = \frac{0.853 * \sqrt{98}}{\sqrt{1-0.853^2}} = 16.18$$

Для уровня значимости $\alpha = 0.05$ и $k = 102$ было определено $t_{\text{крит}} = 1.986$.

Определено, что $|T_{\text{набл}}| > t_{\text{крит}}$, то есть основная гипотеза H_0 должна быть отвергнута, это означает, что выборочный коэффициент корреляции $\bar{r}_{xy}$ значим.

Выводы

В ходе выполнения лабораторной работы был построен двумерный интервальный вариационный ряд (корреляционная таблица). На основании результатов корреляционной таблицы был вычислен выборочный коэффициент корреляции $\overline{r_{xy}} = 0.853$. Выборочный коэффициент корреляции не равен нулю и положителен, значит X и Y коррелированы и зависимы, а также это положительная корреляционная зависимость. Также было посчитано значение выборочного коэффициента корреляции с помощью условных вариант. Коэффициенты корреляции, рассчитанные с помощью основной формулы и условных вариант совпали. С помощью выборочного коэффициента корреляции было оценено значение r_{xy} в случае нормального распределения.

Построен доверительный интервал для коэффициента корреляции при уровне значимости $\gamma = 0.99$. Определено, что доверительный интервал $(0.7637; 0.9102)$ покрывает истинное значение коэффициента корреляции с надежностью $\gamma = 0.99$.

Осуществлена проверка статистической гипотезы о равенстве коэффициента корреляции нулю при заданном уровне значимости $\alpha = 0.05$. Найдены значения $T_{\text{набл}} = 16.18$ и $t_{\text{крит}} = 1.986$. Определено, что $|T_{\text{набл}}| > t_{\text{крит}}$, то есть основная гипотеза H_0 должна быть отвергнута, это означает, что выборочный коэффициент корреляции $\bar{r}_{xy}$ значим.

ПРИЛОЖЕНИЕ А
ИСХОДНЫЙ КОД

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('data.csv')
X = df['nu']
Y = df['E']

h1, h2 = 44, 16
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))))

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)

df_int.iloc[77, 2:6] = df_int.iloc[50, 2:6]
# 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()

N = 100
xv = 465.26
sx = 54.57
yv = 132.09
sy = 19.97
```

```
df_kor = pd.DataFrame(col-
umns=['yi','x1','x2','x3','x4','x5','x6','x7','Xi','yX'])
df_kor['yi'] =
[np.NaN,92.9,108.9,124.9,140.9,156.9,172.9,188.3,np.NaN,np.NaN]
df_kor['x1'] = [343,3,1,0,0,0,0,0,np.NaN,np.NaN]
df_kor['x2'] = [387,3,5,1,0,0,0,0,np.NaN,np.NaN]
df_kor['x3'] = [431,0,6,18,3,0,0,0,np.NaN,np.NaN]
df_kor['x4'] = [475,0,2,12,20,1,0,0,0,np.NaN]
df_kor['x5'] = [519,0,0,1,9,7,0,0,np.NaN,np.NaN]
df_kor['x6'] = [563,0,0,0,1,1,4,0,np.NaN,np.NaN]
df_kor['x7'] = [604,0,0,0,0,0,0,2,np.NaN,np.NaN]

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()
df_curr2
df_kor

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

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

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

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

gamma = 0.99
F = gamma/2
l = 2.58
z1 = (z-l*sz).round(4)
z2 = (z+l*sz).round(4)
(z1,z2)

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)

K = 7
Tn = ((r*np.sqrt(N-2))/np.sqrt(1-r**2)).round(3)
Tn
tk = 1.986

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