**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.

|  |  |  |
| --- | --- | --- |
| Студентка гр. 6382 |  | Фиалковский М.С. |
| Преподаватель |  | Середа В.И. |

Санкт-Петербург

2020

**Цель работы.**

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

**Постановка задачи.**

1. 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, среднеквадратического отклонения, асимметрии и эксцесса).

2. Для системы двух случайных величин (первый признак) и (второй признак):

* Сформировать двумерную выборку, построить корреляционную таблицу и найти – значение выборочного коэффициента корреляции. Для расчета использовать формулу:

Где , – количество интервалов в интервальных рядах для выборок значений случайных величин и соответственно; , – условные варианты, значения которых соответствуют выборочным (интервальным) значениям случайных величин и соответственно; – общий объем выборки.

* Построить доверительный интервал для коэффициента корреляции;
* Осуществить проверку статистической гипотезы о равенстве коэффициента корреляции нулю.

**Ход работы.**

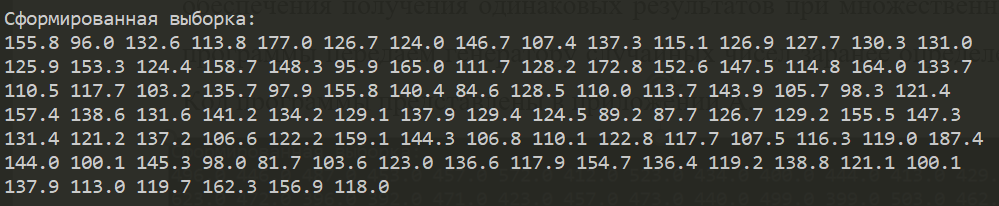
Произведём выборку из представленной генеральной совокупности экспериментальных данных размером 96 элементов. Результат на рис. 1. Для обеспечения получения одинаковых результатов при множественных запусках программы передаём генератору случайных чисел заранее определенное зерно. Код программы представлены в приложении А.

Рис. 1. Начальная выборка

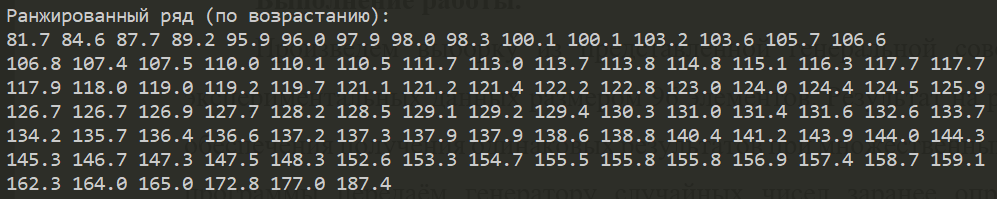
Ранжированный ряд, полученный из выборки, представлен на рис. 2.

Рисунок 2 – Ранжированный ряд.

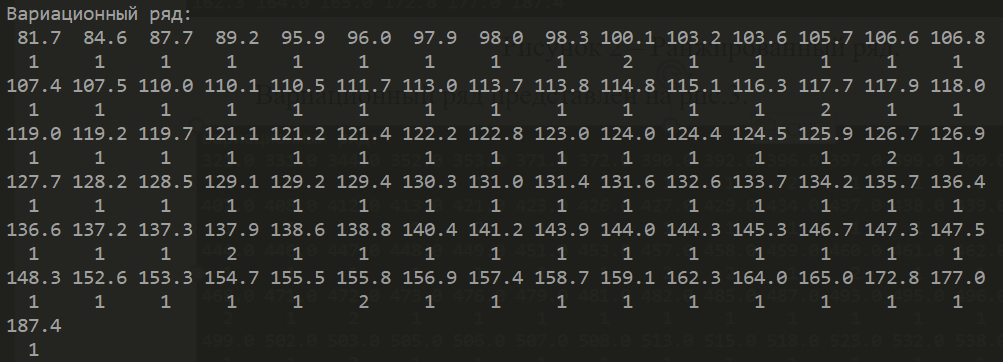
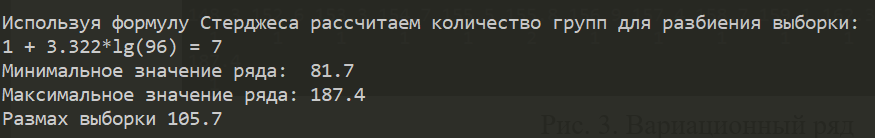
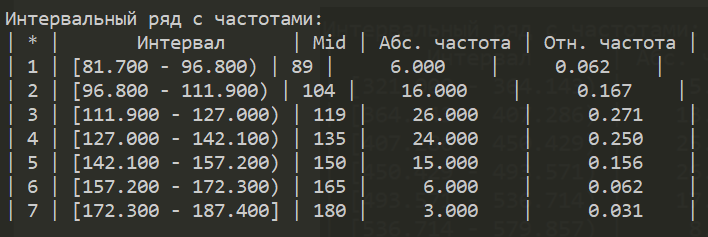
Вариационный ряд представлен на рис.3.

Рис. 3. Вариационный ряд

Для построения интервальный ряда нужно рассчитать количество различных групп. Расчёты и интервальный ряд представлены на рис. 4.

Рис. 4. Расчёты для интервального ряда

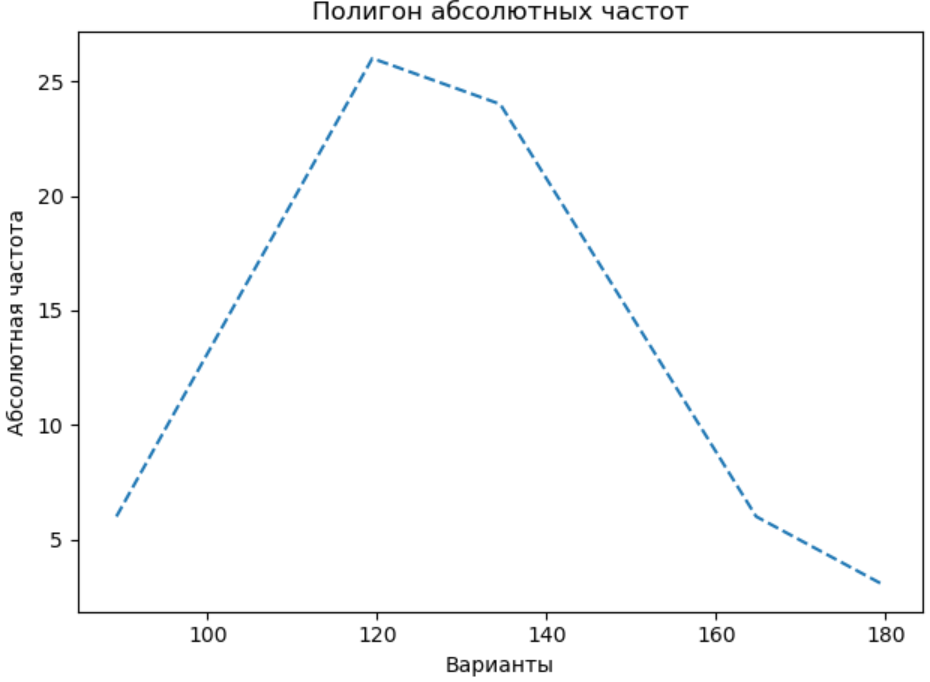
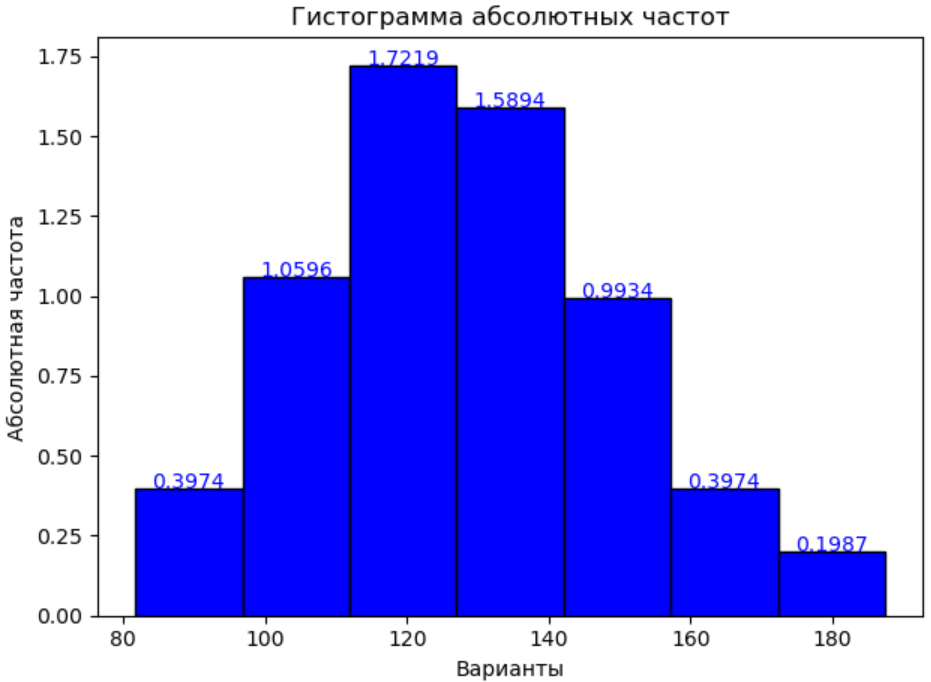
Визуализируем полученный интервальный ряд, построив гистограмму и полигон абсолютных (рис. 5).

Рис. 5. Гистограмма и полигон абсолютных частот

Далее для оценки параметров распределения построим таблицу из второй лабораторной работы, используя построенный только что интервальный ряд. Условные варианты вычислены по формуле , где , .

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

Правильность составления таблицы проверим по этой формуле:

Равенство выполняется, поэтому таблица построена верно.

По таблице найдем точечные оценки параметров распределения, опустив п выкладки вычислений выборочных и центральных выборочных моментов.

* Оценка математического ожидания
* Смещенная оценка дисперсии
* Оценка среднеквадратического отклонения
* Исправленная оценка дисперсии
* Исправленная оценка СКО
* Оценка асимметрии
* Оценка эксцесса

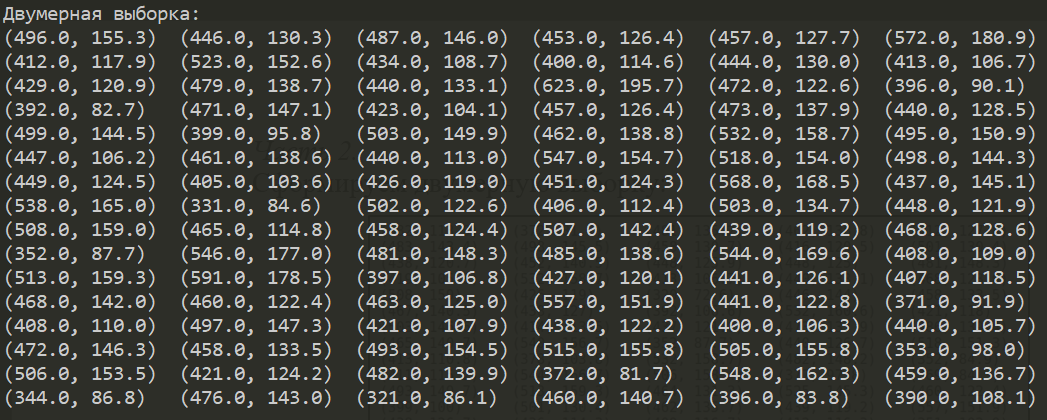
 *Часть 2.*

Рис. 6. Полученная двумерная выборка.

Построим корреляционную таблицу:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

Зафиксируем результаты промежуточных вычислений. Оценки математического ожидания условных вариант:

Оценки стандартных отклонений уловных вариант:

Оценка коэффициента корреляции:

Теперь построим доверительный интервал для коэффициента корреляции. Для этого построим интервал для :

После расчётов получим:

Воспользуемся обратным преобразованием для перехода к доверительному интервалу для коэффициента корреляции:

Итого после расчётов получим доверительный интервал, покрывающий истинное значение параметра с 95% доверительной вероятностью:

Теперь сформулируем нулевую и альтернативную гипотезы о корреляции исследуемых значений (уровень значимости ):

: Коэффициент корреляции равен нулю.

: Коэффициент корреляции не равен нулю.

Найдем значение статистики критерия:

*=>*

Найдем в таблице критическое значение , используя тот факт, что статистика критерия имеет распределение Стьюдента с степенями свободы.

Получаем, что отвергаем гипотезу в пользу .

**Выводы.**

В ходе выполнения лабораторной работы была обработана выборка по второму признаку с помощью разработанных программных средств для лабораторных работ 1-3 для получения нужных данных об исследовании корреляции признаков.

Для этого исследования была сформирована двумерная выборка, построена корреляционная таблица для условных вариант и вычислена точечная оценка коэффициента корреляции Пирсона, равная с доверительным интервалом при уровне доверия 95%. Можно сделать первичный вывод, что корреляционная зависимость между исследуемыми величинами существует.

В заключении была проверена гипотеза о равенстве коэффициента корреляции нулю. Для этого были рассчитаны выборочное значение статистики критерия , и критическое значение для уровня значимости . На основании сравнения этих двух величин можно отклонить гипотезу о том, что коэффициент корреляции равен нулю.

Таким образом, можно утверждать, что корреляционная зависимость между исследуемыми параметрами действительно существует.

**ПРИЛОЖЕНИЕ А**

**ИСХОДНЫЙ КОД**

from math import sqrt, log, exp

import lab1

def print\_beauty(sample: list, size):

    for i in range(1, len(sample)+1):

        print(sample[i-1], end='\t')

        if i % size == 0:

            print(end="\n")

def check(value\_1, value\_2, border\_1, border\_2):

    is\_1\_in\_interval  = value\_1 >= border\_1[0] and value\_1 < border\_1[1]

    is\_2\_in\_interval = value\_2 >= border\_2[0] and value\_2 < border\_2[1]

    return is\_1\_in\_interval and is\_2\_in\_interval

def build\_corr\_table(sample\_2D, borders\_1, borders\_2):

    table = []

    for i in range(len(borders\_1)):

        table.append([])

        for j in range(len(borders\_2)):

            tmp = map(lambda x: check(x[0], x[1], borders\_1[i], borders\_2[j]), sample\_2D)

            table[i].append(sum(tmp))

    return table

n = lab1.selection\_size

general\_population = lab1.read\_data(filename=lab1.data\_file\_name)

sample\_density = lab1.get\_sample\_first(general\_population, n)

sample\_elastic = lab1.get\_sample\_second(general\_population, n)

borders\_1, buckets\_1 = lab1.get\_interval\_sample(sample\_density)

borders\_2, buckets\_2 = lab1.get\_interval\_sample(sample\_elastic)

freqs\_1 = [len(array) for array in buckets\_1]

freqs\_2 = [len(array) for array in buckets\_2]

# Расчёты прогоняем, меняя источник данных в лабе 2

# Часть 2

sample\_2D = list(zip(sample\_density, sample\_elastic))

print("Двумерная выборка:")

print\_beauty(sample\_2D, size=6)

print(end="\n")

table = build\_corr\_table(sample\_2D, borders\_1, borders\_2)

print("Данные для таблицы:")

for row in table:

    print(row)

print(end="\n")

mid\_borders\_1 = [(border[0] + border[1])/2 for border in borders\_1]

mid\_borders\_2 = [(border[0] + border[1])/2 for border in borders\_2]

step\_size\_1 = borders\_1[0][1] - borders\_1[0][0]

step\_size\_2 = borders\_2[0][1] - borders\_2[0][0]

C\_1 = 472 # из прошлых работ

C\_2 = 122

# Условные варианты

v = [(elem - C\_1) / step\_size\_1 for elem in mid\_borders\_1]

u = [(elem - C\_2) / step\_size\_2 for elem in mid\_borders\_2]

# средние для условных вариант

mean\_v = sum([x[0] \* x[1] for x in zip(v, freqs\_1)]) / n

mean\_u = sum([x[0] \* x[1] for x in zip(u, freqs\_2)]) / n

# СКО для условных вариант

S\_v = sqrt(sum([x[0]\*\*2 \* x[1] for x in zip(v, freqs\_1)]) / n - mean\_v\*\*2)

S\_u = sqrt(sum([x[0]\*\*2 \* x[1] for x in zip(u, freqs\_2)]) / n - mean\_u\*\*2)

print('u: Среднее: {0:.3f}, СКО: {1:.3f}'.format(mean\_u, S\_u))

print('v: Среднее: {0:.3f}, СКО: {1:.3f}'.format(mean\_v, S\_v))

print(end="\n")

sum\_from\_table = 0

for i in range(7):

    for j in range(7):

        sum\_from\_table += table[i][j] \* v[i] \* u[j]

print(sum\_from\_table)

# коэффициент корреляции Пирсона

r = (sum\_from\_table + n \* mean\_v \* mean\_u) / (n \* S\_v \* S\_u)

print('Коэффициент корреляции П ирсона {0:.4f}'.format(r))

# нахождение доверительного интервала

z = 0.5 \* log((1 + r)/(1 - r))

# доверительный интервал для z

left  = z - 1.96 / sqrt(n - 3)

right = z + 1.96 / sqrt(n - 3)

print('Доверительный интервал для Z: ({0:.3f}, {1:.3f})'.format(left, right))

# обратное преобразование

left  = (exp(2 \* left) - 1) / (exp(2 \* left) + 1)

right = (exp(2 \* right) - 1) / (exp(2 \* right) + 1)

print('Доверительный интервал: ({0:.3f}, {1:.3f})'.format(left, right))

# проверка гипотезы

t\_sample = r \* sqrt(n-2) / sqrt(1 - r\*\*2)

t\_cr = 1.98

print(t\_cr)

print('t выборочное  = {0:.3f}'.format(t\_sample))

print('t критическое = {0:.3f}'.format(t\_cr))

if t\_sample <= t\_cr:

    print('Не достаточно отснований, чтобы отклонить H0')

else:

    print('Отвергаем гипотезу H0')