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. 8382 _____ Звегинцева Е.Н.

Студент гр. 8382 _____ Мирончик П.Д.

Преподаватель _____ Середа А.-В.И.

Санкт-Петербург
2022

Цель работы.

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

Основные теоретические положения.

Рассмотрим систему двух случайных величин $\{X; Y\}$. Эти случайные величины могут быть независимыми, если плотность их совместного распределения равна произведению плотностей распределения каждой из величин:

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

Иначе между величинами существует *статистическая зависимость*, частным случаем которой является *корреляционная зависимость*. Корреляционной называют статистическую зависимость, при которой изменение значения одной величины приводит к изменению математического ожидания другой величины. Функции регрессии:

$$M\left(\frac{X}{y}\right) = q_1(y)$$

$$M\left(\frac{Y}{x}\right) = q_2(x)$$

Корреляционный момент:

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

Линейный коэффициент корреляции:

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

Для коэффициента корреляции справедливо соотношение:

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

Если r_{xy}=0, величины X и Y некоррелированы. Иначе они коррелированы. Из того, что X и Y коррелированы следует, что они зависимы, но не наоборот: зависимые величины могут быть как коррелированными, так и некоррелированными. При этом независимые величины всегда некоррелированы.

Выборочный коэффициент корреляции можно вычислить по формуле:

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

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:

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

В результате доверительный интервал для r_{xy} генеральной совокупности с доверительной вероятностью γ определяется по следующей схеме:

1. По формуле (1) вычисляется выборочное значение $\bar{z}$.
2. По формуле (2) вычисляется значение $\bar{\sigma}_z$.
3. Интервал для генерального значения представляется в виде:

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

где значение $\lambda(\gamma)$ должно удовлетворять условию:

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

4. Для пересчёта интервала в доверительных интервал для коэффициента корреляции с тем же значением γ необходимо воспользоваться обратным преобразованием Фишера:

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

Так как r_{xy} – случайная величина, то, что r¯xy≠0, не означает, что коэффициент корреляции rxy для генеральной совокупности также не равен нулю. Необходимо выполнить проверку статистической гипотезы. Нулевая и альтернативная гипотезы формулируются следующим образом:

- $H_0: r_{xy} = 0$
- $H_1: r_{xy} \neq 0$

В качестве критерия проверки статистической гипотезы о значимости выборочного коэффициента корреляции можно принять случайную величину:

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

имеющую распределение по закону Стьюдента с $k = N - 2$ степенями свободы. Критическая область для данного критерия двусторонняя.

Проверка гипотезы осуществляется по стандартной схеме:

1. По формуле (3) вычисляется значение $T_{\text{набл}}$.
2. По заданному уровню значимости α и значению k из таблицы определяется значение $t_{\text{крит}}(\alpha, k)$.
3. 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.

Выполнение работы.

- **Проведем статистическую обработку второй выборки в объеме лабораторных работ №1 и №2**

Имеющаяся генеральная совокупность экспериментальных данных представлена в табл. 1. Объём выборки: 114.

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			

Полученные данные при обработке выборки величины *E*, для удобства представлены в табл.2 ниже, без дублирования вычислений.

Таблица 2 - Точечные статистические оценки параметров распределения

Математическое ожидание	129.8438
Дисперсия	439.691
Среднеквадратичное отклонение	20.9688
Исправленное среднеквадратичное отклонение	21.06138
Асимметрия	-0.10855
Эксцесс	-1.22854
Мода	124.175
Медиана	129.5941

- **Построим двумерный интервальный вариационный ряд (рис.1)**

Область значений, принимаемых каждой из величин, разбивается на 8 интервалов, вычисляются середины этих интервалов.

В таблице j-ая строка соответствует j-ому интервалу величины ν, i-ый столбец – i-ому интервалу величины E. В ячейки таблицы записывается абсолютная частота одновременного попадания значения ν двумерной случайной величины (ν,E) в j-ый интервал и значения E этой величины – в i-ый интервал.

Y X	[84.3;95.3)	[95.3;106.3)	[106.3;117.3)	[117.3;128.3)	[128.3;139.3)	[139.3;150.3)	[150.3;161.3)	[161.3;172.3)
[321;350)	2	0	0	0	0	0	0	0
[350;379)	5	4	1	0	0	0	0	0
[379;408)	1	3	6	0	0	0	0	0
[408;437)	0	3	2	6	1	1	0	0
[437;466)	0	0	2	19	11	4	0	0
[466;495)	0	0	0	1	4	12	2	0
[495;524)	0	0	0	0	1	4	4	4
[524;553)	0	0	0	0	0	3	3	5

Рисунок 1 – таблица двумерного интервального ряда.

•По полученному двумерному интервальному вариационному ряду построим корреляционную таблицу.

Основываясь на двумерном интервальном вариационном ряде построим корреляционную таблицу. Для этого внесем суммарные абсолютные частоты для каждого из интервалов величин (значения совпали, с представленными в лабораторной работе 1, следовательно вычисления верны). Результат представлен на рис. 2.

Y / X	[84.3;95.3)	[95.3;106.3)	[106.3;117.3)	[117.3;128.3)	[128.3;139.3)	[139.3;150.3)	[150.3;161.3)	[161.3;172.3)	N
[321;350)	2	0	0	0	0	0	0	0	2
[350;379)	5	4	1	0	0	0	0	0	10
[379;408)	1	3	6	0	0	0	0	0	10
[408;437)	0	3	2	6	1	1	0	0	13
[437;466)	0	0	2	19	11	4	0	0	36
[466;495)	0	0	0	1	4	12	2	0	19
[495;524)	0	0	0	0	1	4	4	4	13
[524;553)	0	0	0	0	0	3	3	5	11
N	8	10	11	26	17	24	9	9	114

Рисунок 2 – корреляционная таблица двумерного интервального ряда.

Далее преобразуем середины интервалов в условные варианты, для упрощения вычислений. В качестве условного нуля примем середин интервала с наибольшей частотой. На рис.3 представлена таблица, с интервалами, замененными на условные варианты

v u	-4	-3	-2	-1	0	1	2	3	n_v
-4	2	0	0	0	0	0	0	0	2
-3	5	4	1	0	0	0	0	0	10
-2	1	3	6	0	0	0	0	0	10
-1	0	3	2	6	1	1	0	0	13
0	0	0	2	19	11	4	0	0	36
1	0	0	0	1	4	12	2	0	19
2	0	0	0	0	1	4	4	4	13
3	0	0	0	0	0	3	3	5	11
n_u	8	10	11	26	17	24	9	9	114

Рисунок 3 – таблица двумерного интервального ряда с условными вариантами.

- **Вычислим выборочный коэффициент корреляции при помощи условных вариант**

Значение выборочного коэффициента корреляции вычисляется по формуле:

$$\overline{r_{xy}} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}u_i v_j - N\overline{v}_{\text{в}}\overline{u}_{\text{в}}}{NS_vS_u},$$

где $\overline{v}_{\text{в}}$, $\overline{u}_{\text{в}}$ - выборочные средние для условных вариант, S_v, S_u – неисправленные среднеквадратические отклонения условных вариант.

Для вычисления построим вспомогательную таблицу (умножим абсолютные частоты на условные варианты). Результаты представлены на рис.4.

v u	-4	-3	-2	-1	0	1	2	3	n_v
-4	32	0	0	0	0	0	0	0	32
-3	60	36	6	0	0	0	0	0	102
-2	8	18	24	0	0	0	0	0	50
-1	0	9	4	6	0	-1	0	0	18
0	0	0	0	0	0	0	0	0	0
1	0	0	0	-1	0	12	4	0	15
2	0	0	0	0	0	8	16	24	48
3	0	0	0	0	0	9	18	45	72
n_u	100	63	34	5	0	28	38	69	337

Рисунок 4 – вспомогательная таблица двумерного интервального ряда.

$$\overline{r_{xy}} = 0{,}89$$

•Вычислим выборочный коэффициент корреляции с помощью стандартной формулы

Значение выборочного коэффициента корреляции вычисляется по формуле:

$$\overline{r_{xy}} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}u_i E_i - N\bar{v}\bar{E}}{NS_v S_u},$$

где S_v, S_u – неисправленные среднеквадратические отклонения условных вариант. В начале лабораторной работы приведены значения исправленных СКО s_v, s_E, которые вычисляются как:

$$s_v = \sqrt{\frac{N}{N-1}}S_v,\ s_E = \sqrt{\frac{N}{N-1}}S_E$$

Тогда формула примет вид:

$$\overline{r_{xy}} = \frac{\sum_{i=1}^{K_y}\sum_{j=1}^{K_x} n_{ij}u_i v_j - N\overline{v}_{\text{в}}\overline{u}_{\text{в}}}{(N-1)S_v S_u}$$

$$\overline{r_{xy}} = 0{,}89$$

Значение совпало с предыдущим, значит вычисления проведены верно. Данное значение близко к единице и положительно, значит можно сделать вывод о том, что данные сильно коррелируют и при возрастании одной величины произойдет возрастание и второй.

•Построим доверительный интервал для коэффициента корреляции при уровне значимости $\gamma\epsilon\{0,95;0,99\}$.

Используем Z преобразование Фишера. Где распределение $\bar{z}$:

$$\bar{z} = 0{,}5 \ln \frac{1 + \overline{r_{xy}}}{1 - \overline{r_{xy}}}$$

СКО для распределения z:

$$\overline{\sigma_z} = \frac{1}{\sqrt{N-3}}$$

Для построения доверительного интервала для коэффициента корреляции сделаем обратное преобразование Фишера:

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

Вычислим:

```
Доверительный интервал для доверительной вероятности 0.95:
        1.2395887502906935 < z < 1.6116518477533175,
        0.8453382645322609 < r < 0.9234037598334105
Доверительный интервал для доверительной вероятности 0.99:
        1.1811334011063026 < z < 1.6701071969377084,
        0.8278086643776805 < r < 0.9315658558635433
```

Можно заметить, что при увеличении уровня надежности получили более широкий доверительный интервал.

•Осуществим проверку статистической гипотезы о равенстве коэффициента корреляции нулю при заданном уровне значимости $\alpha = 0,05$

Проверим гипотезу о равенстве нулю коэффициента корреляции. Вычислим $T_{\text{набл}}$ по формуле:

$$T_{\text{набл}} = \frac{\overline{r_{xy}}\sqrt{N-2}}{\sqrt{1-{r_{xy}}^2}} = 20{,}743$$

$$T_{\text{крит}} = 1{,}981(\text{для уровня значимости } \alpha = 0{,}05)$$

Исходя из того, что $T_{\text{набл}} > T_{\text{крит}}$, гипотеза о равенстве нулю коэффициента корреляции отвергается.

Вывод

В ходе выполнения лабораторной работы был проведен корреляционный анализ распределения случайных величин. Был вычислен линейный коэффициент корреляции двумя различными способами. Его близость к 1 дает нам сделать вывод о сильной корреляционной зависимости меду величинами. Также были построены доверительные интервалы с различными вероятностями и проведена проверка, в ходе которой мы отвергли нулевую гипотезу.

Из перечисленного можно сделать вывод, что выборочный коэффициент корреляции значимо отличен от нуля.