Лабораторная работа № 4
«Корреляционный анализ»

студента Грачева Игоря группы Б23-504. Дата сдачи:__________
Ведущий преподаватель: _______________ оценка: ______ подпись:_______

Вариант №5

Цель работы: изучение функций Statistics and Machine Learning Toolbox™ MATLAB / Python SciPy.stats для проведения корреляционного анализа данных.

1. Исходные данные

Характеристики наблюдаемых случайных величин:

СВ	Распределение	Параметры	Математическое ожидание, m_i	Дисперсия, σ_i^2	Объем выборки, *n*
X	N(5, 2)	m = 5, σ = 2	5	4	100
Y	N(5, 2)	m = 5, σ = 2	5	4	

Примечание: для генерации случайных чисел использовать функции **rand, randn, chi2rnd (scipy.stats: uniform.rvs, norm.rvs, chi2.rvs)**

Выборочные характеристики:

СВ	Среднее, $\bar{x}_i$	Оценка дисперсии, s_i^2	КК по Пирсону, $\tilde{r}_{XY}$	КК по Спирмену, $\tilde{\rho}_{XY}$	КК по Кендаллу, $\tilde{\tau}_{XY}$
X	4.5520	3.1480	-0.0727	-0.0777	-0.0501
Y	5.1883	4.1221			

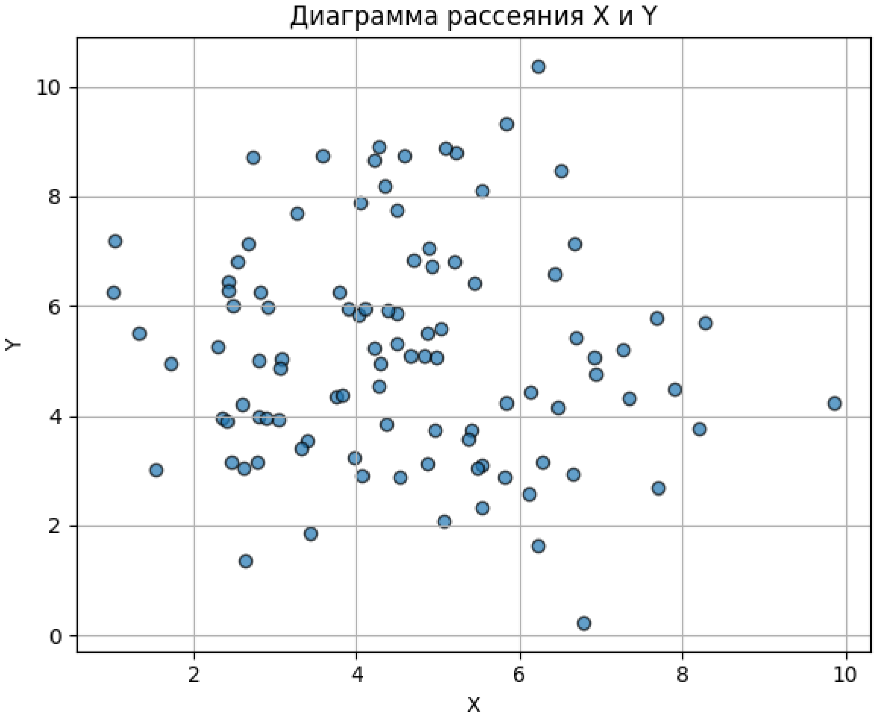
Проверка значимости коэффициентов корреляции:

Статистическая гипотеза, H_0	*p-value*	Статистическое решение при $\alpha = 0.05$	Ошибка стат. решения
H_0: $r_{XY} = 0$	0.4723	Не отвергаем	Нет
H_0: $\rho_{XY} = 0$	0.4424	Не отвергаем	Нет
H_0: $\tau_{XY} = 0$	0.4602	Не отвергаем	Нет

Примечание: для проверки гипотез использовать функцию **corr (scipy.stats.pearsonr)**

2. Визуальное представление двумерной выборки

Диаграмма рассеяния случайных величин *X* и *Y*:



Примечание: для построения диаграммы использовать функции **plot, scatter (matplotlib.pyplot.scatter)**

3. Проверка независимости методом таблиц сопряженности

Статистическая гипотеза: $H_0 : F_Y(y \mid X \in \Delta_1) = \ldots = F_Y(y \mid X \in \Delta_k) = F_Y(y)$

Эмпирическая/теоретическая таблицы сопряженности:

X \ *Y*	[0.22; 2.25)	[2.25; 4.29)	[4.29; 6.32)	[6.32; 8.35)	[8.35; 10.39]
Δ_1 = [1.00; 2.78)	1 / 0.9	6 / 5.76	6 / 6.66	4 / 2.88	1 / 1.80

Δ_2 = [2.78; 4.55)	1 / 1.75	10 / 11.20	17 / 12.95	4 / 5.60	3 / 3.50
Δ_3 = [4.55; 6.32)	2 / 1.50	11 / 9.60	6 / 11.10	6 / 4.80	5 / 3.00
Δ_4 = [6.32; 8.09)	1 / 0.7	3 / 4.48	7 / 5.18	2 / 2.24	1 / 1.40
Δ_5 = [8.09; 9.86]	0 / 0.15	2 / 0.96	1 / 1.11	0 / 0.48	0 / 0.30

Примечание: для группировки использовать функцию **hist3 (matplotlib.pyplot.hist2d)**

Выборочное значение статистики критерия	*p-value*	Статистическое решение при α = 0.05	Ошибка стат. решения
10.9347	0.8135	Не отвергаем	Нет

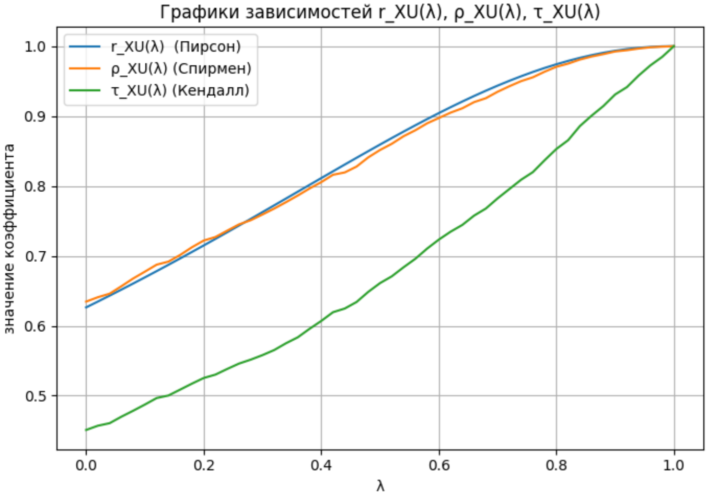
Примечание: для проверки гипотезы использовать функцию **crosstab (scipy.stats.chi2_contingency)**

4. Исследование корреляционной связи

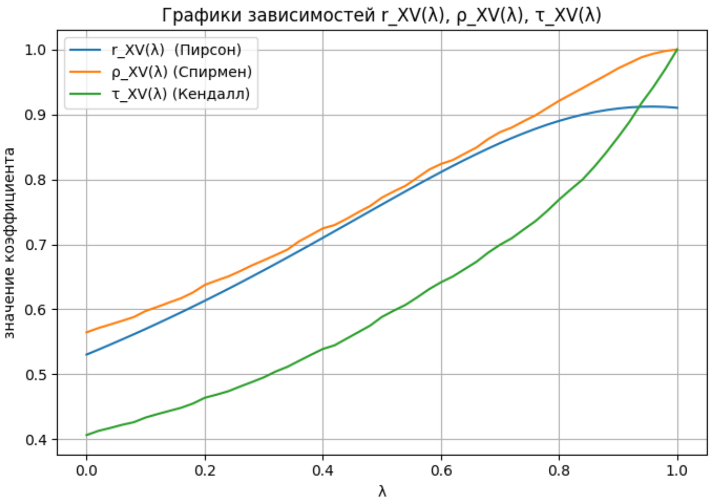
Случайная величина $U = \lambda X + (1-\lambda)Y$, $\lambda \in [0; 1]$

Случайная величина $V = \lambda X^3 + (1-\lambda)Y^3$ $\lambda \in [0; 1]$

Графики зависимостей коэффициента корреляции $\tilde{r}_{XU}(\lambda)$, рангового коэффициента корреляции по Спирмену $\tilde{\rho}_{XU}(\lambda)$, по Кендаллу $\tilde{\tau}_{XU}(\lambda)$



Графики зависимостей $\tilde{r}_{XV}(\lambda)$, $\tilde{\rho}_{XV}(\lambda)$, $\tilde{\tau}_{XV}(\lambda)$



Выводы: все коэффициенты монотонно растут от λ = 0 к λ = 1. Коэффициенты Спирмена и Кендалла почти везде чуть выше Пирсона. Для пары (X, U) зависимости почти линейные и растут почти параллельно друг другу, для пары (X, V) кривизна графиков сильнее.

Диаграмма рассеяния случайных величин X и V при $\lambda = 0$:

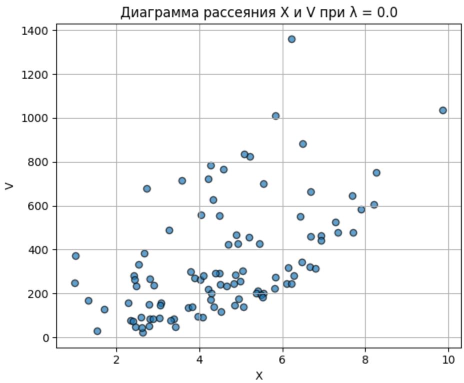


Диаграмма рассеяния **рангов** случайных величин X и V при $\lambda = 0$:

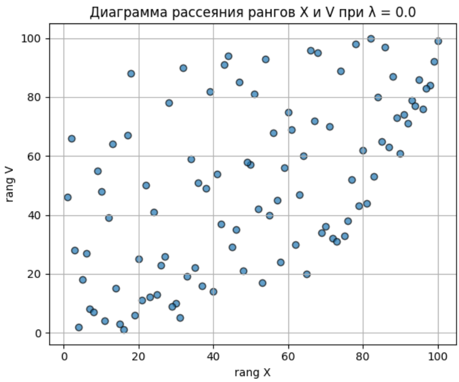


Диаграмма рассеяния случайных величин X и V при $\lambda = 1$:

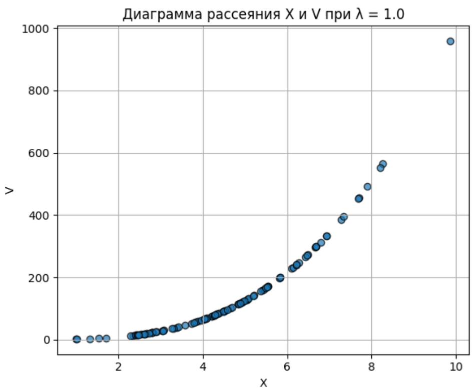
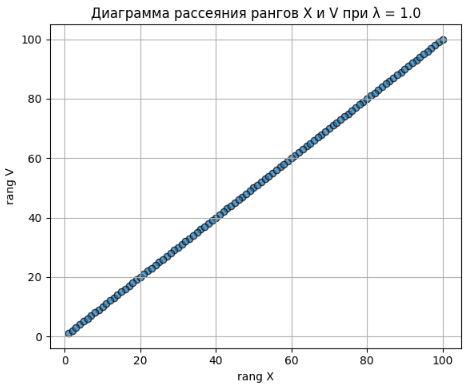


Диаграмма рассеяния **рангов** случайных величин X и V при $\lambda = 1$:



Примечание: для расчёта рангов использовать функцию **tiedrank (scipy.stats.rankdata)**

Выводы: при $\lambda = 0$ видно, что при больших X больше и V, их связь возрастает, поэтому линейная корреляция Пирсона есть. Связь между рангами X и V монотонная, но не идеальная, значит коэффициенты Спирмена и Кендалла высокие, но меньше 1.
При $\lambda = 1$ видно, что связь между X и V строго монотонная, поэтому корреляция Пирсона близка к 1. Диаграмма рангов образует почти ровную прямую, что означает почти идеальную ранговую зависимость, то есть коэффициенты Спирмена и Кендалла близки к 1.