Лабораторная работа № 4
«Корреляционный анализ»

студента Шамаева Сергея группы Б21-514 . Дата сдачи:__________
Ведущий преподаватель: _______________ оценка: ______ подпись:_______

Вариант №_4_

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	X ~ R(5, 15)	a = 5 b = 15	10	8.33	300
Y	Y ~ N(10, 5)	m = 10 std = 5	10	25	

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

Выборочные характеристики:

СВ	Среднее, $\overline{x}_i$	Оценка дисперсии, s_i^2	КК по Пирсону, $\tilde{r}_{XY}$	КК по Спирмену, $\tilde{\rho}_{XY}$	КК по Кендаллу, $\tilde{\tau}_{XY}$
X	9.808	7.890	0.002	0.015	0.010
Y	10.334	24.963			

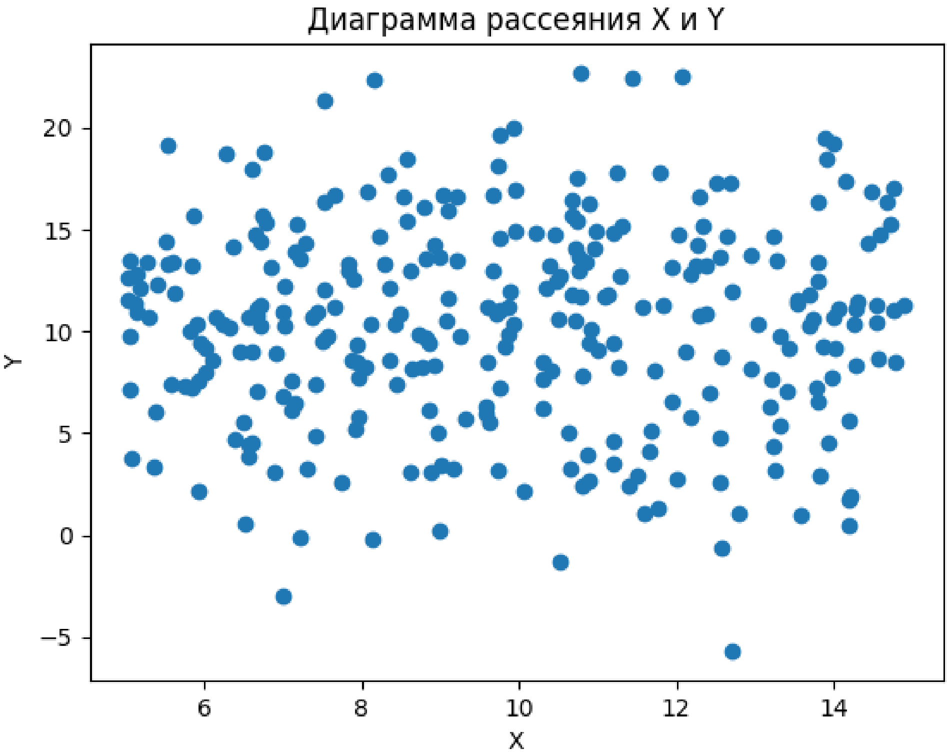
Проверка значимости коэффициентов корреляции:

Статистическая гипотеза, H_0	*p-value*	Статистическое решение при $\alpha = 0.05$	Ошибка стат. решения
H_0: $r_{XY} = 0$	0.975	H0 принимается	нет
H_0: $\rho_{XY} = 0$	0.793	H0 принимается	нет
H_0: $\tau_{XY} = 0$	0.796	H0 принимается	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	[−5.689; −1.6e−02)	[−1.6e−02; 5.667)	[5.657; 11.330)	[11.330; 17.003)	[17.003; 22.676]
Δ_1 = [5.036; 7.008)	0	10	29	20	4
Δ_2 = [7.008; 8.980)	3	7	28	21	4
Δ_3 = [8.980; 10.953)	1	11	23	28	5

Δ_4 = [10.952; 12.925])	2	12	13	18	6
Δ_5 = [12.925; 14.897]	0	10	25	16	4

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

Выборочное значение статистики критерия	*p-value*	Статистическое решение при α = ______	Ошибка стат. решения
16.132	0.443	H0 принимается	нет

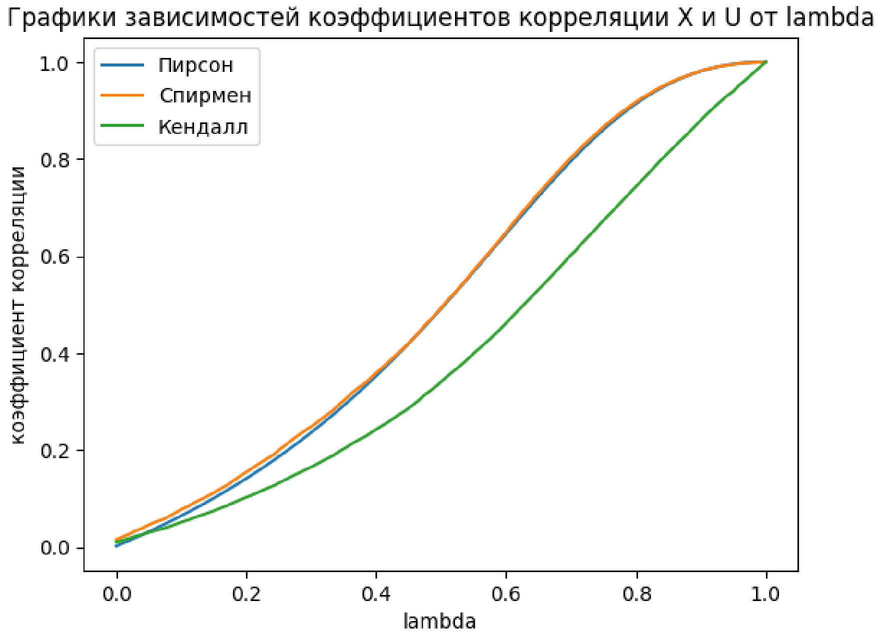
Примечание: для проверки гипотезы использовать функцию **crosstab (scipy.stats.chi2_contingency)**

4. Исследование корреляционной связи

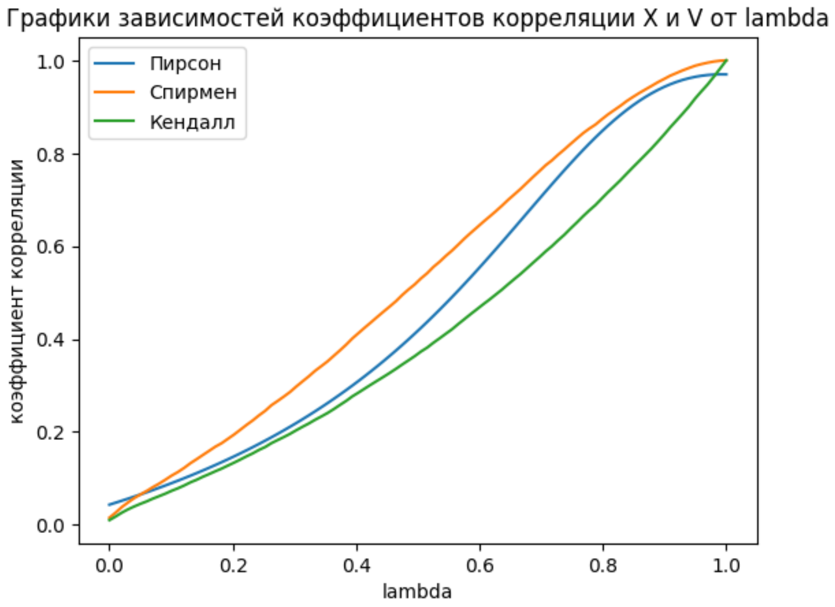
Случайная величина $U = \lambda X + (1-\lambda)Y$, $\lambda \in [0; 1]$

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

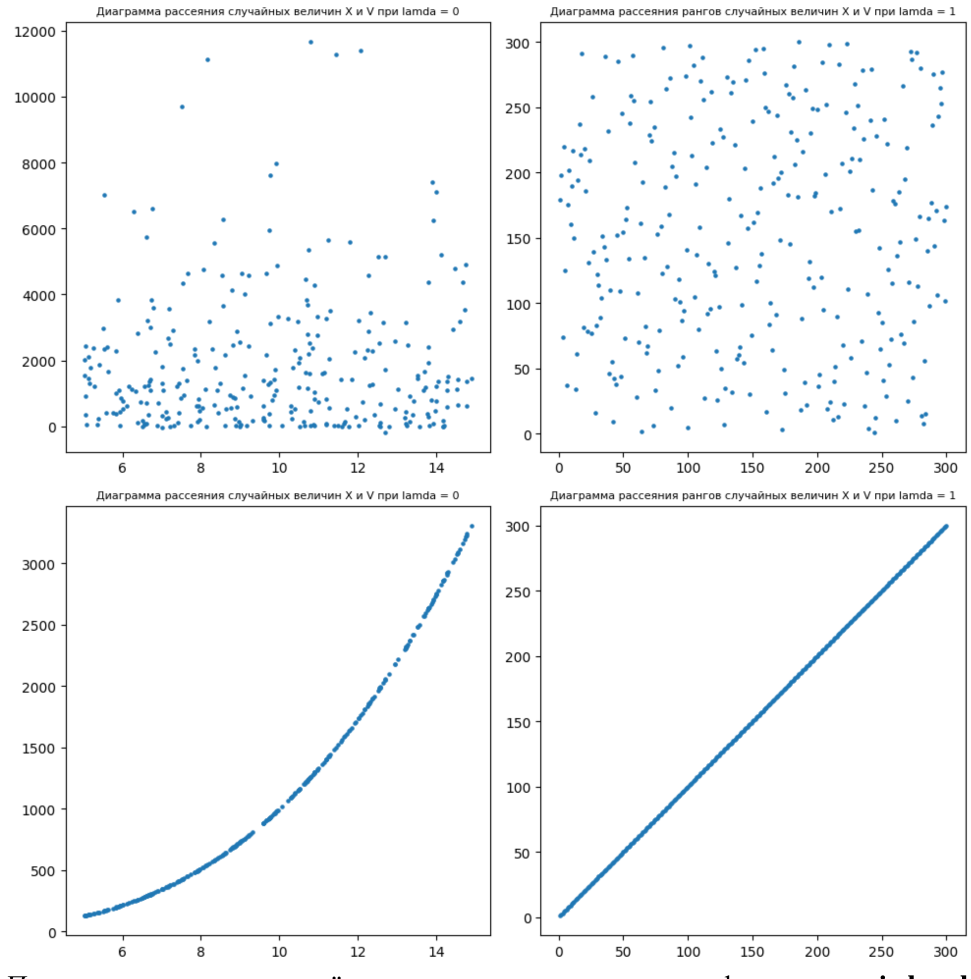
Графики зависимостей коэффициента корреляции $\tilde{r}_{XU}(\lambda)$, рангового коэффициента корреляции по Спирмену $\tilde{\rho}_{XU}(\lambda)$, по Кендаллу $\tilde{\tau}_{XU}(\lambda)$



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



Выводы: При приближении lambda к единице, коэфф. корреляции приближаются к 1, так как в этот случае ситуация стремится к зависимости X от X. Также мы можем наблюдать, что все графики – строго возрастающие функции. На графике 1 все кэфф. достигли единицы, так при приближении к 1 есть линейная зависимость. Однако, на рисунке 2 мы можем наблюдать, что коэфф. корреляции Пирсона не достигает 1, так кубическая связь не является линейной. Коэффициент корреляции по Спирмену и по Кендаллу достигает 1, так как они рассматривают монотонность функций (для X и X^3 монотонность есть).



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

Выводы: На первых двух графиках отсутствует какая-либо видимая взаимосвязь, т.к. там рассматриваются (X и Y^3). Причем, мы можем также говорить и об отсутствии какой-либо монотонности, исходя из рисунка 2 (нет связи рангов). На графике 3 мы наблюдаем чёткую функциональную зависимость (кубическую), что согласуется с нашими данными (X и X^3). Рисунок 4 также показывает монотонность этой связи (чёткая линейная функция).