Лабораторная работа № 4
«Корреляционный анализ»

студента Розинко Е.Д. группы Б21-524. Дата сдачи: 17.12.2023
Ведущий преподаватель: Трофимов А.Г. 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	$R(5, 15)$	$R(a, b)$	10	8,3	300
Y	$N(10, 5)$	$N(m, \sigma)$	10	25	

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	9.946	8.229	0.012	0.009	0.005
Y	10.049	26.546			

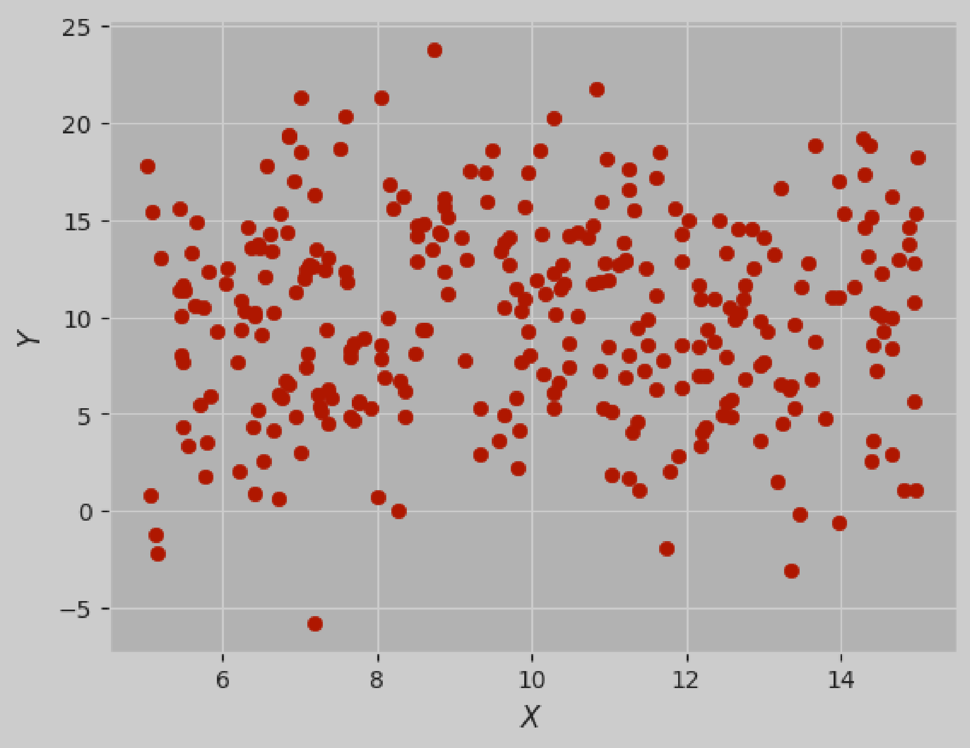
Проверка значимости коэффициентов корреляции:

Статистическая гипотеза, H_0	*p-value*	Статистическое решение при $\alpha = 0.05$	Ошибка стат. решения
H_0: $r_{XY} = 0$	0.824	H_0 принимается	2 рода
H_0: $\rho_{XY} = 0$	0.876	H_0 принимается	2 рода
H_0: $\tau_{XY} = 0$	0.881	H_0 принимается	2 рода

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 таблица сопряженности:

X \ Y	[-5.765; 0.142)	[0.142; 6.050)	[6.050; 11.957)	[11.957; 17.865)	[17.865; 23.772]
Δ_1 = [5.021; 7.016)	2	18	21	19	4
Δ_2 = [7.016; 9.011)	2	12	20	22	4
Δ_3 = [9.011; 11.006)	0	9	24	20	5
Δ_4 = [11.006; 13.001)	1	16	29	19	1
Δ_5 = [13.001; 14.996]	3	10	19	16	4

Теоретическая таблица сопряженности:

X \ Y	[-0.986; 3.905)	[3.905; 8.796)	[8.796; 13.687)	[13.687; 18.578)	[18.578; 23.473]
Δ_1 = [5.010; 7.006)	1.70	13.86	24.10	20.48	3.84
Δ_2 = [7.006; 9.002)	1.6	13	22.6	19.2	3.6
Δ_3 = [9.002; 10.998)	1.54	12.56	21.84	18.56	3.48
Δ_4 = [10.998; 12.994)	1.76	14.3	24.86	21.12	3.96
Δ_5 = [12.994; 14.990]	1.38	11.26	19.58	16.64	3.12

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

Выборочное значение статистики критерия	*p-value*	Статистическое решение при $\alpha = 0.05$	Ошибка стат. решения
12.227	0.728	H_0 принимается	нет

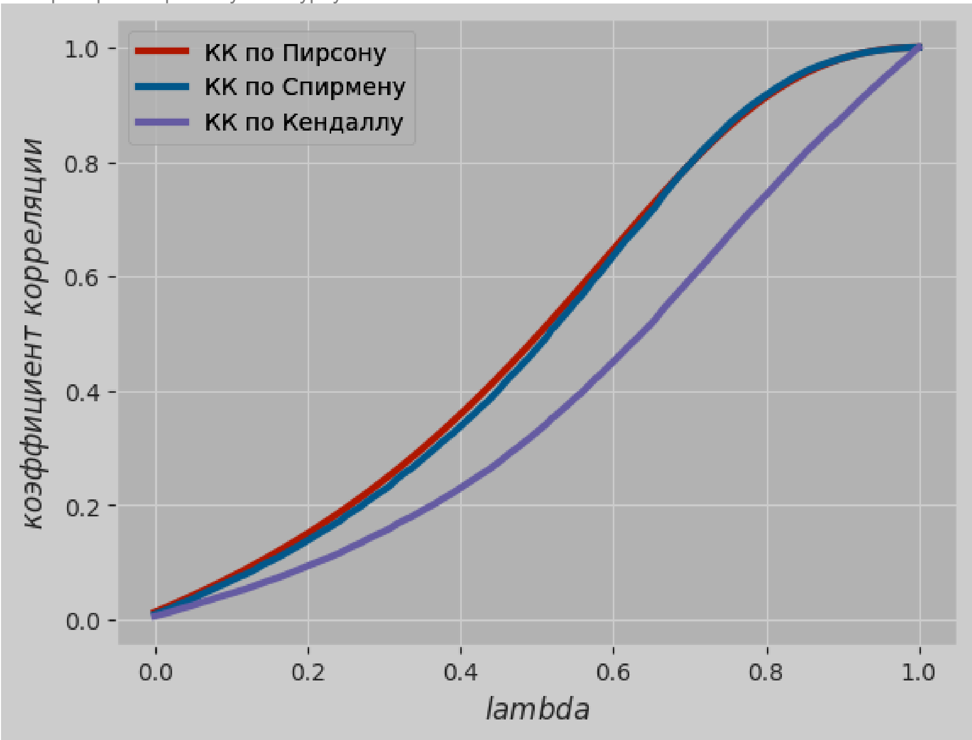
Примечание: для проверки гипотезы использовать функцию **crosstab (scipy.stats.chi2_contingency)**

4. Исследование корреляционной связи

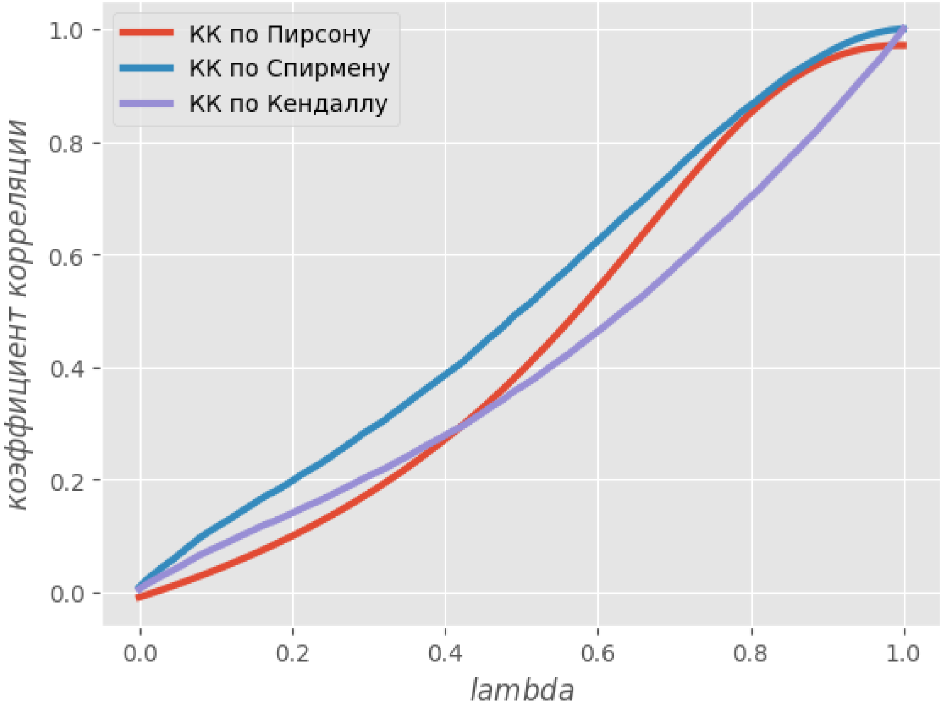
Случайная величина $U = \lambda X + (1-\lambda)Y$, $\lambda \in [0; 1]$

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 $\tilde{\tau}_{XU}(\lambda)$



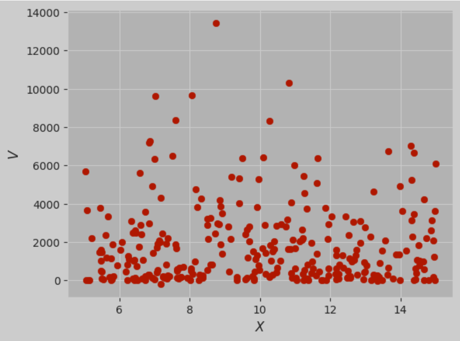
Графики зависимостей $\tilde{r}_{XV}(\lambda)$, $\tilde{\rho}_{XV}(\lambda)$, $\tilde{\tau}_{XV}(\lambda)$



Выводы: С увеличением значения $\lambda \in [0, 1]$ коэффициент корреляции $\tilde{r}_{XU}(\lambda)$ ранговый коэффициент корреляции по Спирмену $\tilde{\rho}_{XU}(\lambda)$ и по Кендаллу $\tilde{\tau}_{XU}(\lambda)$ стремятся к единице. При $\lambda = 0$ коэффициенты корреляции равны нулю и статистическая связь между случайными величинами отсутствует, а при увеличении значения λ теснота статистической связи между случайными величинами увеличивается, и при $\lambda = 1$ между случайными величинами имеется линейная функциональная связь.

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

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



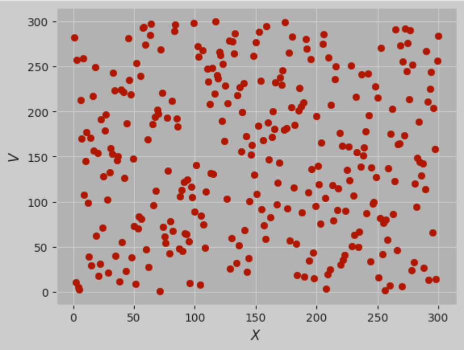
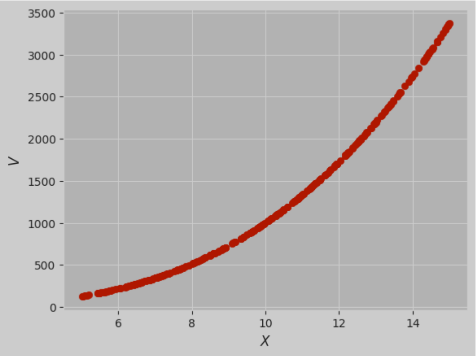
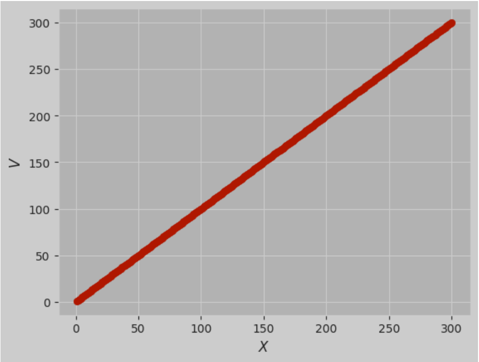


Диаграмма рассеяния случайных величин X и V при $\lambda = 1$:

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





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

Выводы: Из диаграммы рассеяния случайных величин X и V при $\lambda = 0$ видно, что статистическая связь между данными случайными величинами отсутствует, при этом ранги случайных величин X и V при $\lambda = 0$ рассеяны практически равномерно внутри квадрата. На диаграмме рассеяния случайных величин X и V при $\lambda = 1$ прослеживается монотонная зависимость между случайными величинами, при этом переход к рангам выпрямляет данную зависимость.