Лабораторная работа № 4
«Корреляционный анализ»

студента Мельниковой М.Н. Дата сдачи:__________
Ведущий преподаватель: Трофимов А.Г. оценка: ______ подпись:_______

Вариант №9

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	$N(10, 2)$	m = 10 σ = 2	10	4	200
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, $\overline{x}_i$	Оценка дисперсии, s_i^2	КК по Пирсону, $\tilde{r}_{XY}$	КК по Спирмену, $\tilde{\rho}_{XY}$	КК по Кендаллу, $\tilde{\tau}_{XY}$
X	9.887208	5.331838	-0.064485	-0.088461	-0.062311
Y	4.039111	4.226291			

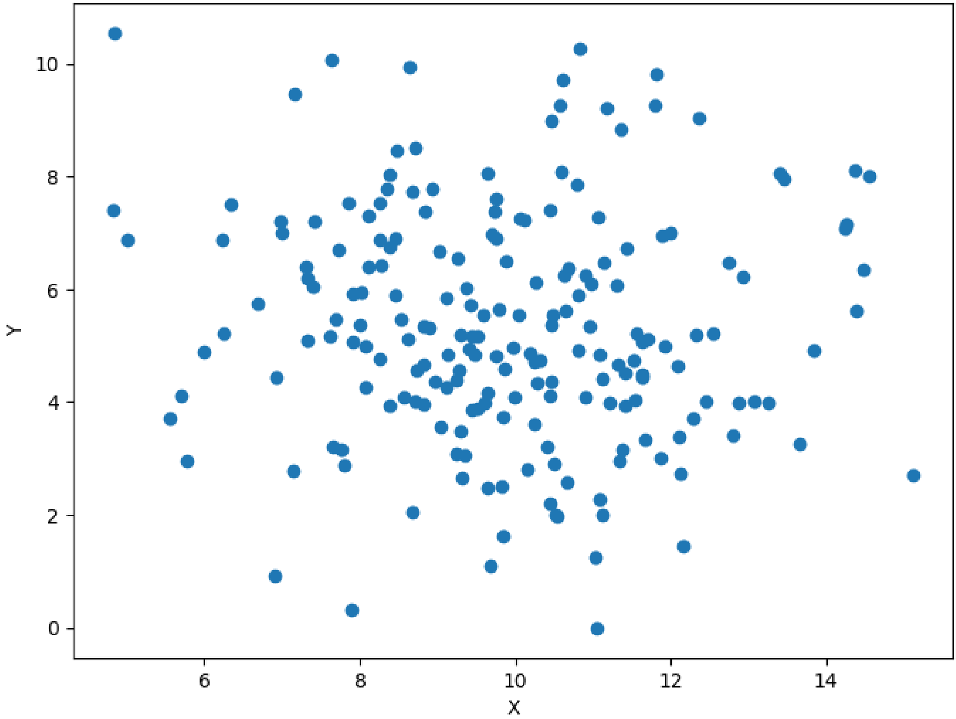
Проверка значимости коэффициентов корреляции:

Статистическая гипотеза, H_0	*p-value*	Статистическое решение при $\alpha = 0{,}05$	Ошибка стат. решения
H_0: $r_{XY} = 0$	0.3643074413164177	H_0 принимается	нет
H_0: $\rho_{XY} = 0$	0.2128967786706946	H_0 принимается	нет
H_0: $\tau_{XY} = 0$	0.1900758510657695	H_0 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	[-4.7215e-03; 2.1058e+00)	[2.1058e+00; 4.2164e+00)	[4.2164e+00; 6.3270e+00)	[6.3270e+00; 8.4376e+00)	[8.4376e+00; 1.0548e+01]
Δ_1 = [4.82618929; 6.88403027)	0 / 0.605	3 / 2.75	3 / 4.125	4 / 2.695	1 / 0.825
Δ_2 = [6.88403027; 8.94187124)	3 / 2.97	8 / 13.5	20 / 20.25	18 / 13.23	5 / 4.05
Δ_3 = [8.94187124; 10.99971221)	4 / 4.235	21 / 19.25	34 / 28.875	14 / 18.865	4 / 5.775
Δ_4 = [10.99971221; 13.05755318)	4 / 2.475	14 / 11.25	16 / 16.875	6 / 11.025	5 / 3.375
Δ_5 = [13.05755318; 15.11539416]	0 / 0.715	4 / 3.25	2 / 4.875	7 / 3.185	0 / 0.975

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

Выборочное значение статистики критерия	*p-value*	Статистическое решение при $\alpha = 0{,}05$	Ошибка стат. решения
21.530446465511403	0.1590100708861939	H_0 принимается	нет

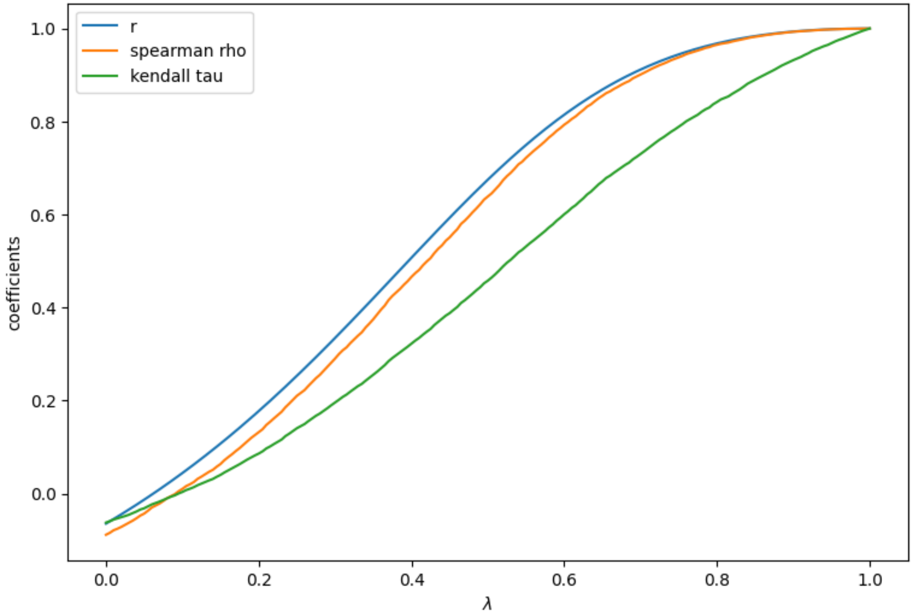
Примечание: для проверки гипотезы использовать функцию **crosstab (scipy.stats.chi2_contingency)**

4. Исследование корреляционной связи

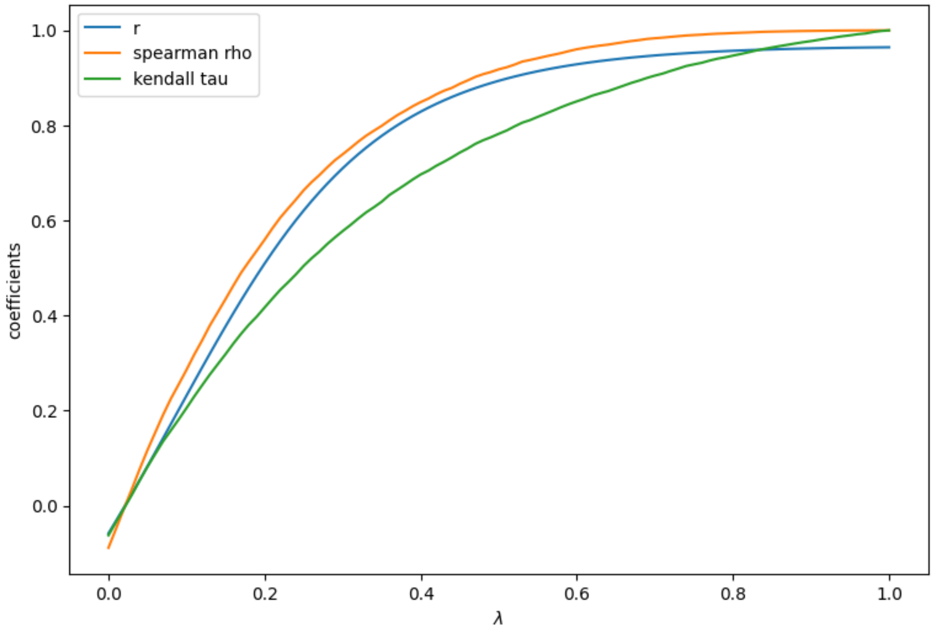
Случайная величина $U = \lambda X + (1-\lambda)Y$, $\lambda \in [0; 1]$

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: из графиков видно, что по мере возрастания λ увеличиваются и корреляционные коэффициенты по Пирсону, Спирмену и Кендаллу, что при отсутствии функциональной связи попарно между величинами X и V, X и U, то есть при λ=0, значения коэффициентов будут равняться нулю. А при λ=1 на первом графике наблюдаем равенство всех коэффициентов единице, что свидетельствует о линейной функциональной связи между величинами X и U, на втором графике – корреляционный коэффициент по Пирсону не достигает значения единицы, но очень близок к нему, что отвечает наличию нелинейной функциональной связи между X и V(при КД = 1), в то время как значение τ = 1 свидетельствует о монотонно возрастающей зависимости между X и V.

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

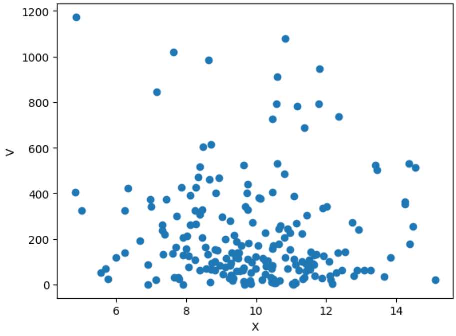


Диаграмма рассеяния **рангов** случайных величин X и V при $\lambda = 0$:

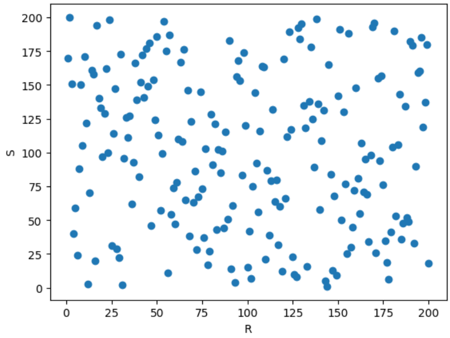


Диаграмма рассеяния случайных величин X и V при $\lambda = 1$:

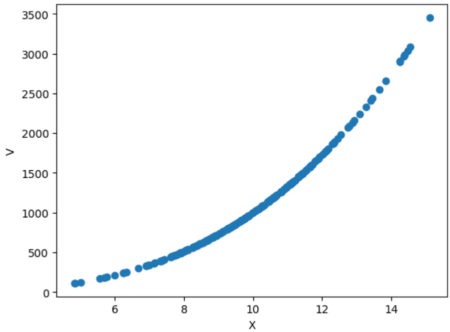
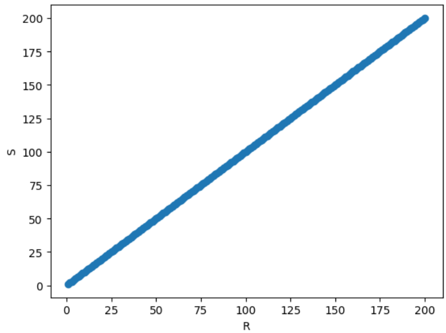


Диаграмма рассеяния **рангов** случайных величин X и V при $\lambda = 1$:



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

Выводы: из диаграммы рассеяния случайных величин X и V и диаграммы рассеяния рангов случайных величин X и V при $\lambda = 0$ видно, что зависимость между случайными величинами X и V отсутствует, для независимых случайных величин характерно практически равномерное рассеяние выборочных рангов. А для случая $\lambda = 1$ на диаграмме рассеяния случайных величин X и V наблюдается монотонно возрастающая зависимость, «выпрямляющаяся» на диаграмме рассеяния рангов величин X и V.