Лабораторная работа № 4
«Корреляционный анализ»

студента __Васильев__ группы_Б21-524_. Дата сдачи:__16.12.2023__
Ведущий преподаватель: _______________ оценка: ______ подпись:_______

Вариант №_6_

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	$\chi^2(2)$	$\boldsymbol{\chi^2(n)}$	2	4	150
Y	$N(3, 1)$	$\boldsymbol{N(m, \sigma)}$	3	1	

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	2.012	4.197	-0.002	-0.038	-0.027
Y	2.890	0.896			

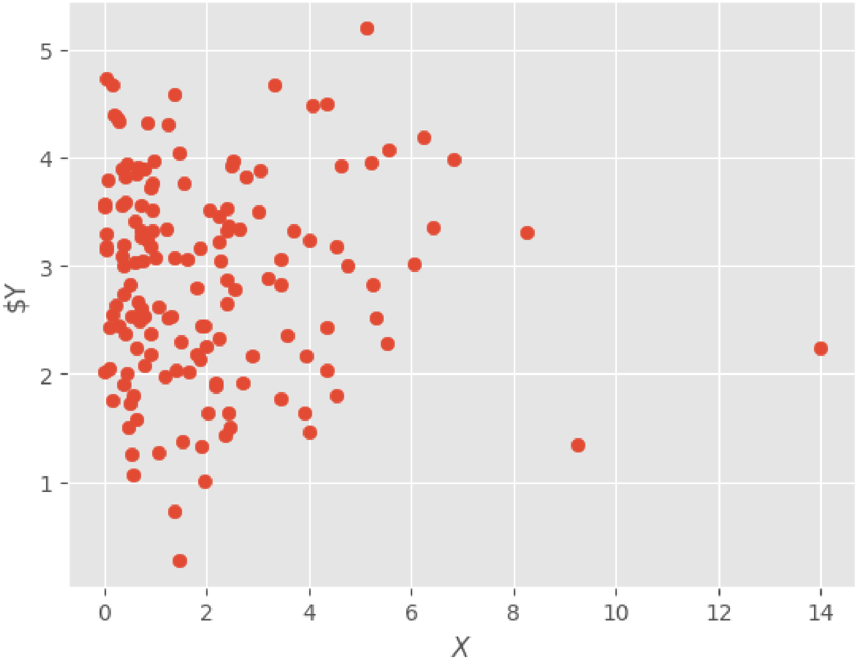
Проверка значимости коэффициентов корреляции:

Статистическая гипотеза, H_0	*p-value*	Статистическое решение при α = _0.05_	Ошибка стат. решения
H_0: $r_{XY} = 0$	0.976	H_0 принимается	2 рода
H_0: $\rho_{XY} = 0$	0.645	H_0 принимается	2 рода
H_0: $\tau_{XY} = 0$	0.627	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	[0.397; 3.296)	[3.296; 6.196)	[6.196; 9.096)	[9.096; 11.995)	[11.995; 14.895)
$\Delta_1 = [0.296; 4.455)$	5	27	39	36	8
$\Delta_2 = [4.455; 8.613)$	0	7	11	6	4
$\Delta_3 = [8.613; 12.771)$	0	0	1	4	0
$\Delta_4 = [12.771; 16.929)$	0	1	0	0	0
$\Delta_5 = [16.929; 21.086]$	0	1	0	0	0

<u>Теоретическая таблицы сопряженности:</u>

X \ Y	[0.397; 3.296)	[3.296; 6.196)	[6.196; 9.096)	[9.096; 11.995)	[11.995; 14.895)
Δ_1 = [0.296;4.455)	3.833	27.60	39.10	35.27	9.200
Δ_2 = [4.455;8.613)	0.933	6.720	9.520	8.586	2.240
Δ_3 = [8.613; 12.771)	0.166	1.200	1.700	1.533	0.400
Δ_4 = [12.771; 16.929)	0.033	0.240	0.340	0.306	0.080
Δ_5 = [16.929; 21.086]	0.033	0.240	0.340	0.306	0.080

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

Выборочное значение статистики критерия	*p-value*	Статистическое решение при α = _0.05_	Ошибка стат. решения
16.233	0.436	H_0 *принимается*	нет

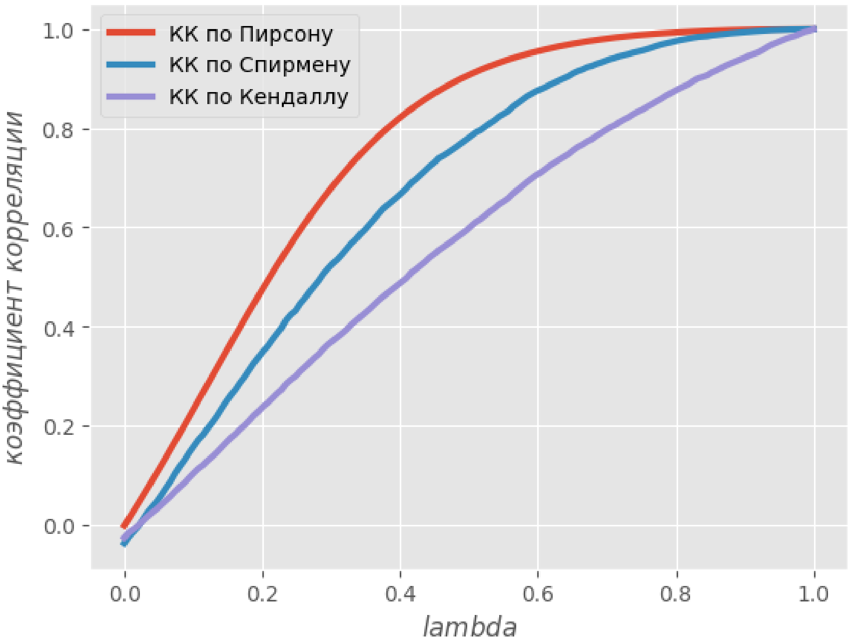
Примечание: для проверки гипотезы использовать функцию **crosstab (scipy.stats.chi2_contingency)**

<u>4. Исследование корреляционной связи</u>

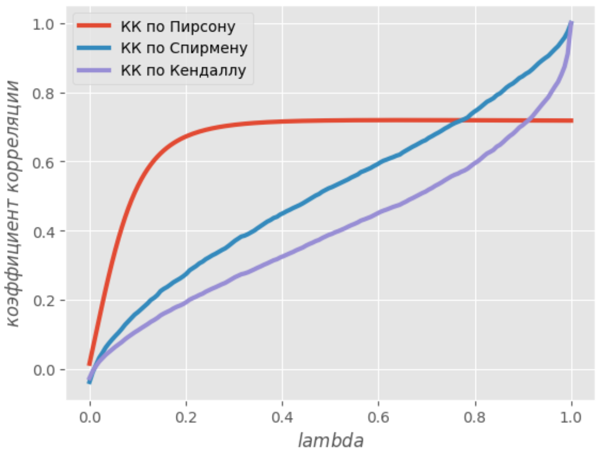
Случайная величина $U = \lambda X + (1-\lambda)Y$, $\lambda \in [0; 1]$

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 $r_{XU}(\lambda)$, ранговый коэффициент корреляции по Спирмену $\rho_{XU}(\lambda)$ и по Кендаллу $\tau_{XU}(\lambda)$ стремятся к единице. При $\lambda = 0$ коэффициенты корреляции равны нулю и статистическая связь между случайными величинами отсутствует, а при увеличении значения λ теснота статистической свзи между случайными величинами увеличивается, и при $\lambda = 1$ между случайными величинами имеется линейная функциональная связь.

Диаграмма рассеяния случайных величин *X* и *V* при λ = 0:

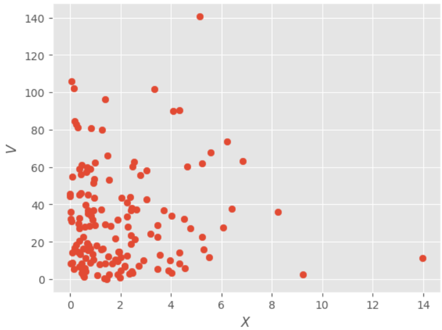


Диаграмма рассеяния **рангов** случайных величин *X* и *V* при λ = 0:

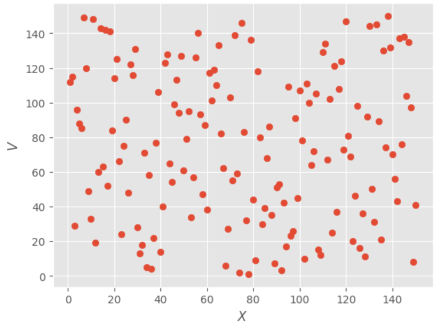


Диаграмма рассеяния случайных величин *X* и *V* при λ = 1:

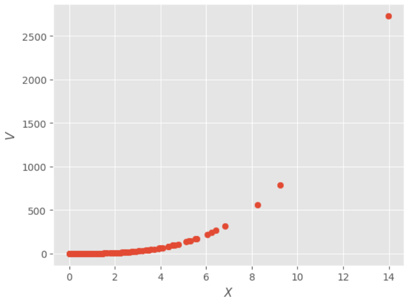
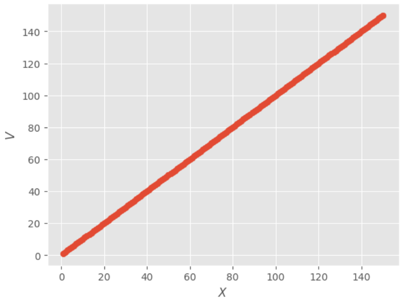


Диаграмма рассеяния **рангов** случайных величин *X* и *V* при λ = 1:



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.