Лабораторная работа № 4
«Корреляционный анализ»

студента Моисеенко Олеси Игоревны группы Б20-514. Дата сдачи: _____
Ведущий преподаватель: Сорока А.А. оценка: ____ подпись:_____

Вариант №15

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	$\chi^2(15)$	k = 15	15	30	100
Y	$R(5, 25)$	a = 5 b = 25	15	33.333333	

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	16.109734	38.986900	0.04711073807833272	0.0219021902190219	0.01696969696969697
Y	15.114881	28.101614			

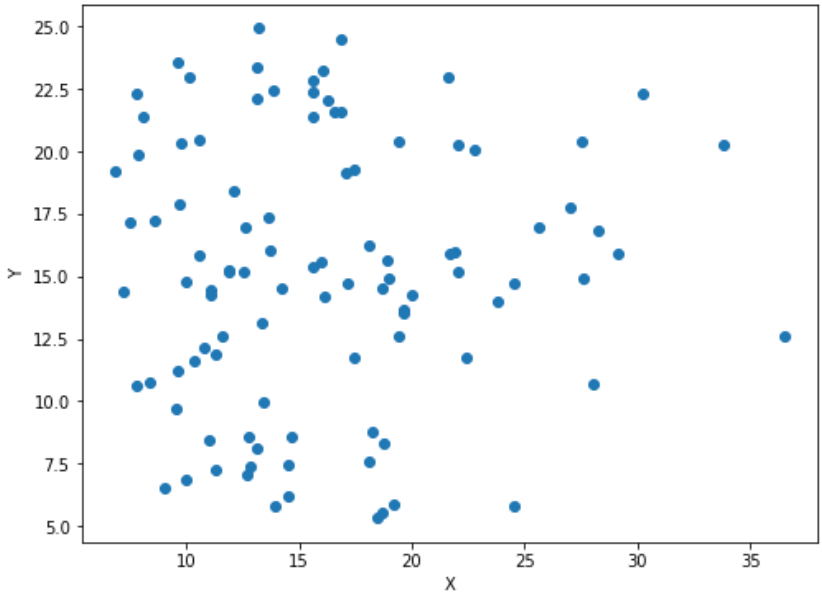
Проверка значимости коэффициентов корреляции:

Статистическая гипотеза, H_0	*p-value*	Статистическое решение при α = 0.05	Ошибка стат. решения
H_0: $r_{XY} = 0$	0.6416142714105674	H_0 принимается	нет
H_0: $\rho_{XY} = 0$	0.8287584399128479	H_0 принимается	нет
H_0: $\tau_{XY} = 0$	0.8024622175039021	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: $H_0 : F_Y(y \mid X \in \Delta_1) = ... = F_Y(y \mid X \in \Delta_k) = F_Y(y)$

Эмпирическая/теоретическая таблицы сопряженности:

X \ Y	[5.33825547; 9.25430786)	[9.25430786; 13.17036024)	[13.17036024; 17.08641263)	[17.08641263; 21.00246501)	[21.00246501; 24.9185174]
Δ_1 = [6.84296756; 12.77148684)	6 6.65	8 5.25	9 10.85	8 5.95	4 6.3
Δ_2 = [12.77148684; 18.70000611)	10 6.84	3 5.4	8 11.16	3 6.12	12 6.48
Δ_3 = [18.70000611; 24.62852538)	3 3.61	2 2.85	10 5.89	3 3.23	1 3.42
Δ_4 = [24.62852538; 30.55704466)	0 1.52	1 1.2	4 2.48	2 1.36	1 1.44
Δ_5 = [30.55704466; 36.48556393]	0 0.38	1 0.3	0 0.62	1 0.34	0 0.36

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

Выборочное значение статистики критерия	*p-value*	Статистическое решение при α = 0.05	Ошибка стат. решения
25.227924646163828	0.06590682272076571	H_0 принимается	нет

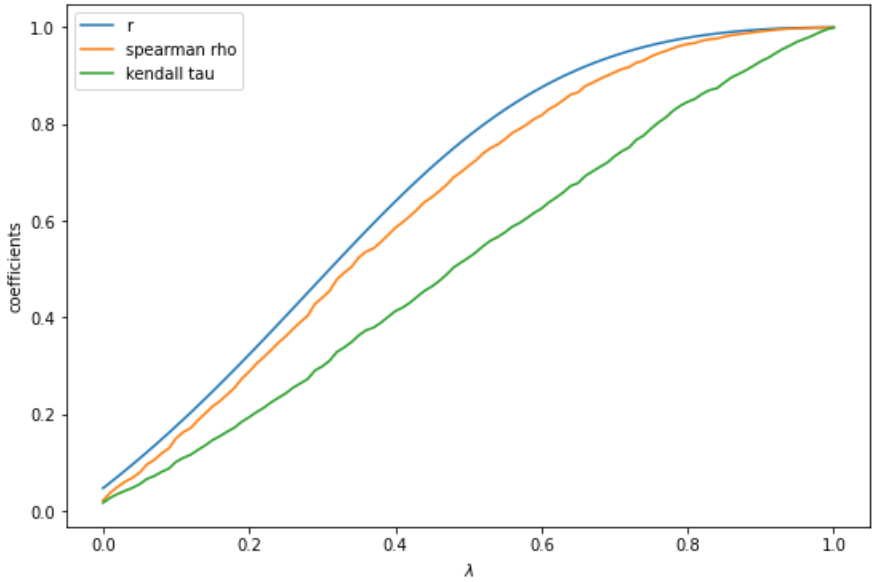
Примечание: для проверки гипотезы использовать функцию **crosstab (scipy.stats.chi2_contingency)**

4. Исследование корреляционной связи

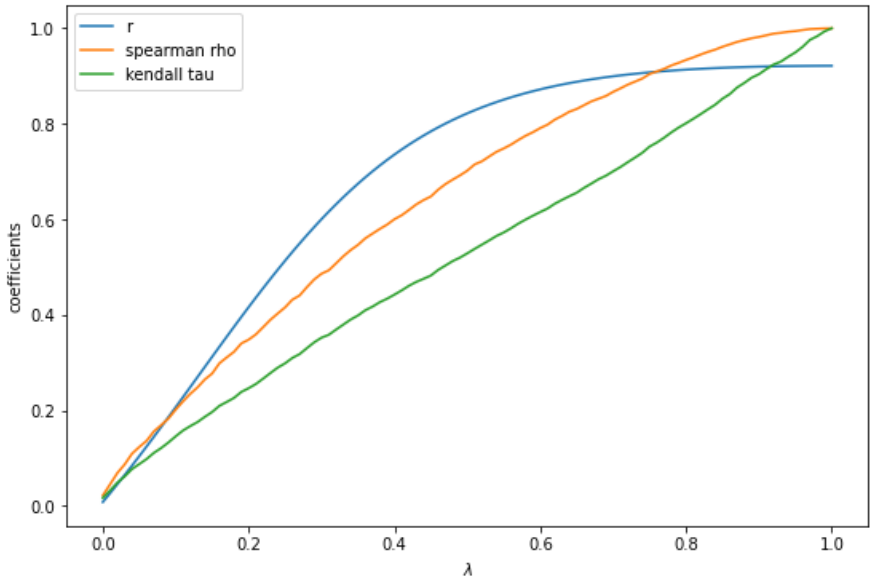
Случайная величина $U = \lambda X + (1–\lambda)Y$, $\lambda \in [0; 1]$

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 λ = 0:

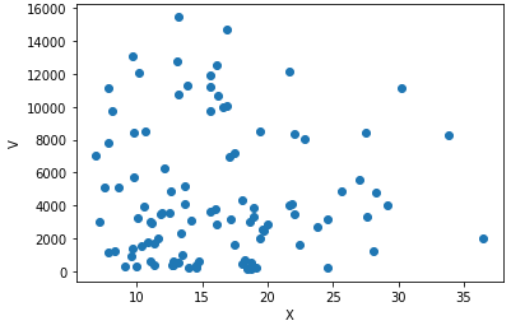


Диаграмма рассеяния **рангов** случайных величин *X* и *V* при λ = 0:

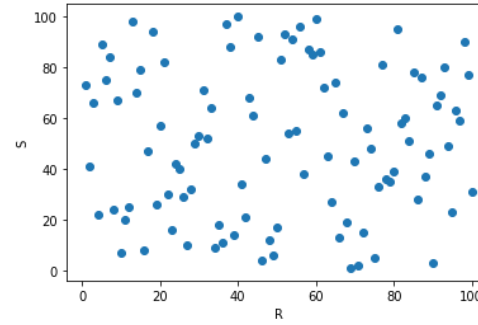


Диаграмма рассеяния случайных величин *X* и *V* при λ = 1:

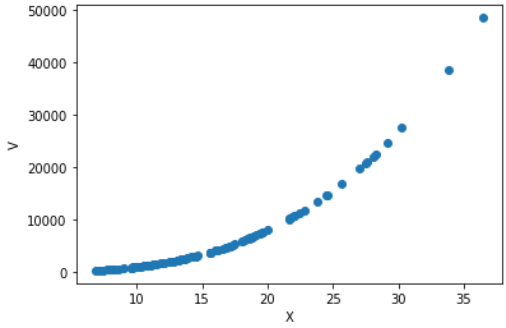
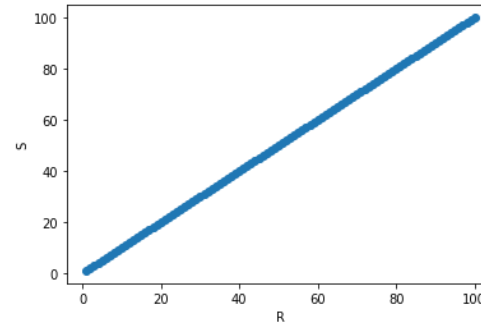


Диаграмма рассеяния **рангов** случайных величин *X* и *V* при λ = 1:



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

Выводы: из диаграммы рассеяния случайных величин X и V 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.