Лабораторная работа № 4
«Корреляционный анализ»

студента Смирнова Даниила группы Б18-501. Дата сдачи: 4.12.20
Ведущий преподаватель: Трофимов А Г 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	chi2	2	2	4	150
Y	N	(3,1)	3	1	

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	2.03	4.23	0.026	0.045	0.03
Y	3	1.025			

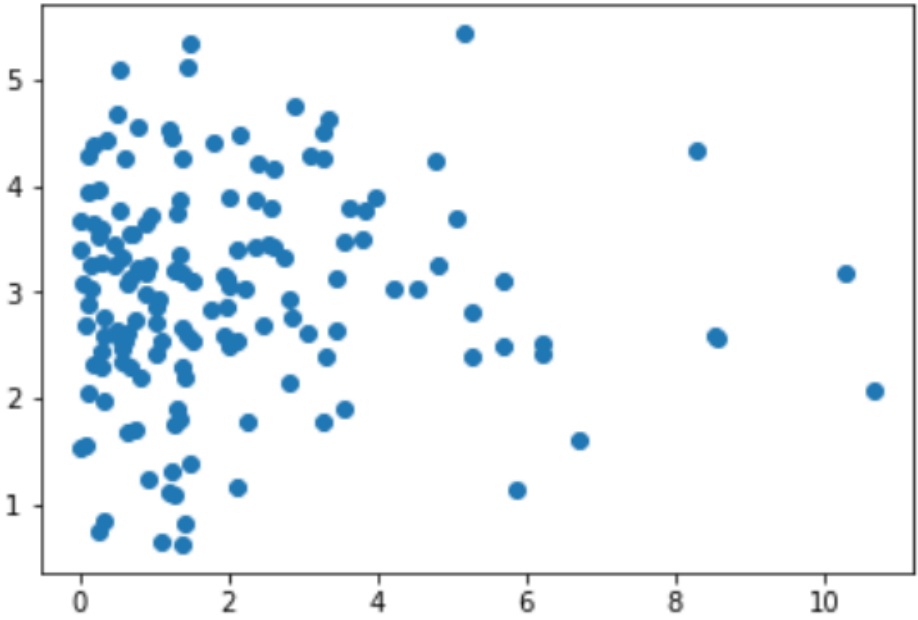
Проверка значимости коэффициентов корреляции:

Статистическая гипотеза, H_0	*p-value*	Статистическое решение при $\alpha = 0.05$	Ошибка стат. решения
H_0: $r_{XY} = 0$	0.75	Гипотеза верна	Гипотеза H0 принимается (Решение верно
H_0: $\rho_{XY} = 0$	0.58	Гипотеза верна	Гипотеза H0 принимается (Решение верно
H_0: $\tau_{XY} = 0$	0.59	Гипотеза верна	Гипотеза H0 принимается (Решение верно

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	[0.614; 1.58)	[1.58; 2.54)	[2.54; 3.51)	[3.51; 4.47)	[4.47; 5.44]
Δ_1 = [0.015; 2.146)	13 / 9.33	22 / 22	39 / 40	20 / 21.33	6 / 7.33
Δ_2 = [2.146; 4.278)	0 / 2.98	5 / 7.04	14 / 12.8	9 / 6.82	41 / 2.35
Δ_3 = [4.278; 6.40)	1 / 1.12	4 / 2.64	4 / 4.8	2 / 2.56	1 / 0.88

Δ_4 = [6.40; 8.54)	0 / 0.28	1 / 0.66	1 / 1.2	1 / 0.64	0 / 0.22
Δ_5 = [8.54; 10.67]	0 / 0.28	1 / 0.66	2 / 1.2	0 / 0.64	0 / 0.22

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

Выборочное значение статистики критерия	*p-value*	Статистическое решение при $\alpha = 0.05$	Ошибка стат. решения
11.08	0.80	Гипотеза верна	Гипотеза H0 принимается (Решение верно

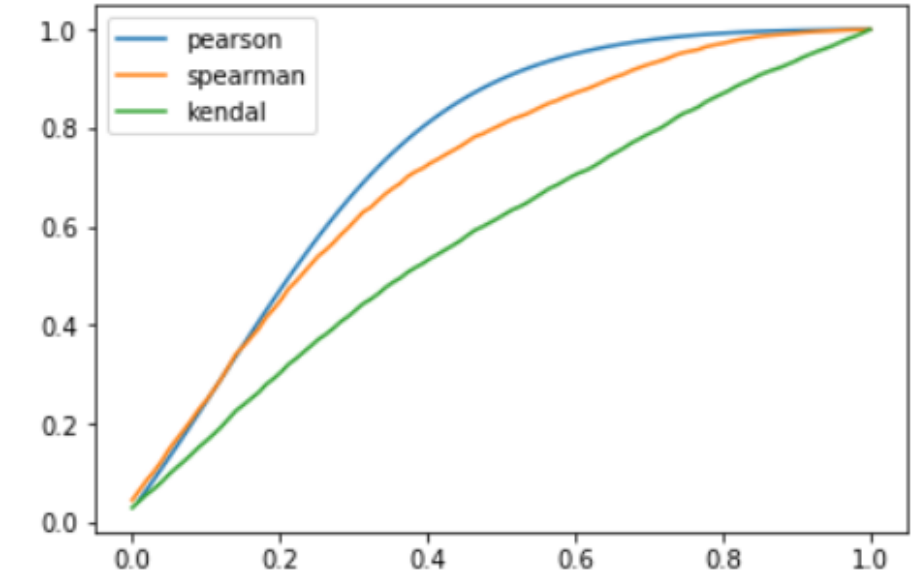
Примечание: для проверки гипотезы использовать функцию **crosstab (scipy.stats.chi2_contingency)**

4. Исследование корреляционной связи

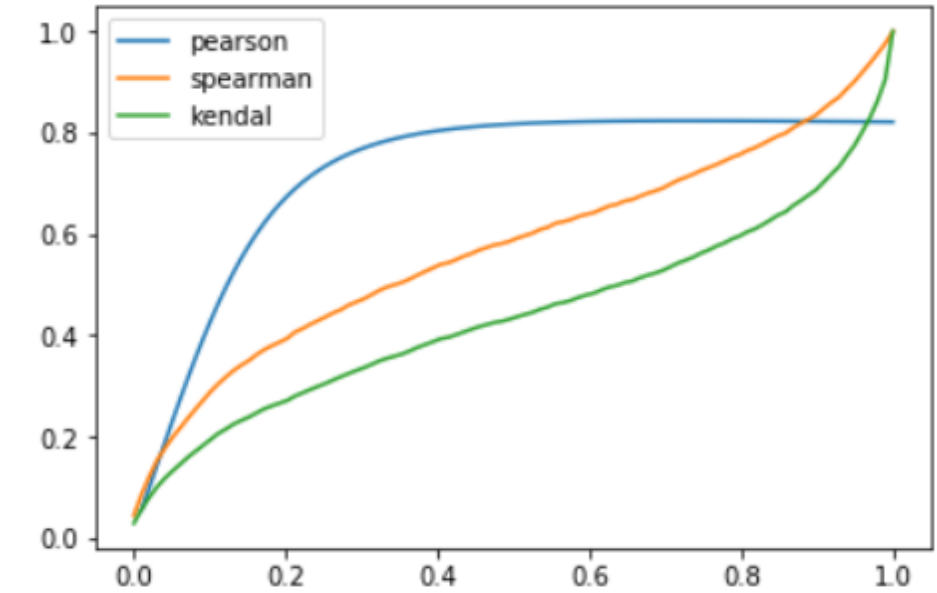
Случайная величина $U = \lambda X + (1-\lambda)Y$, $\lambda \in [0; 1]$

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



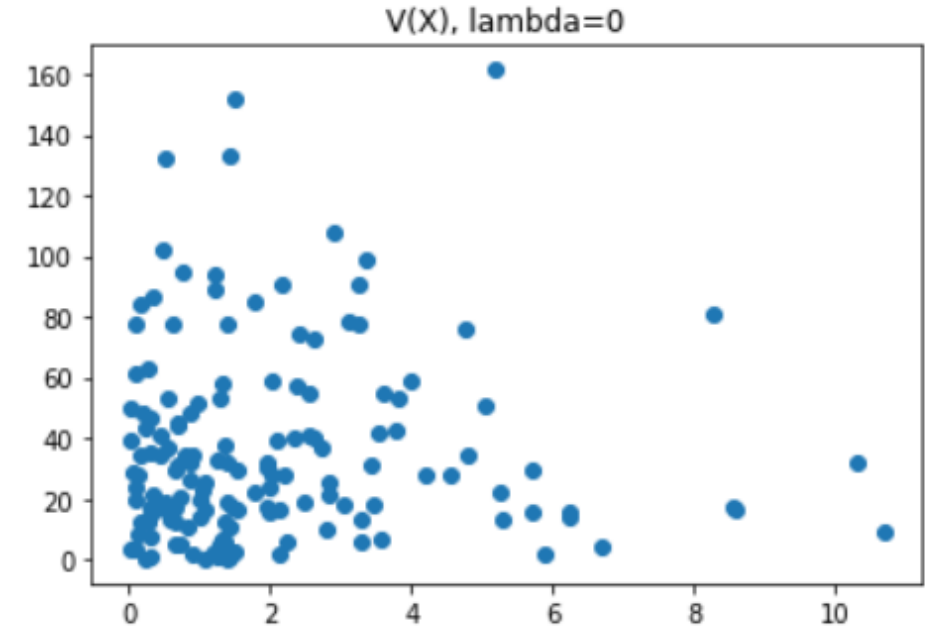
Графики зависимостей $\tilde{r}_{xv}(\lambda)$, $\tilde{\rho}_{xv}(\lambda)$, $\tilde{\tau}_{xv}(\lambda)$



Выводы: при увеличении лямбды увеличиваются коэффициенты корреляции. Коэффициент корреляции Пирсона не достигает единицы во втором случае, тк он равен единице при линейной зависимости, а тут кубическая зависимость

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

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


V(X), lambda=0
160
140
120
100
80
60
40
20
0
0
2
4
6
8
10

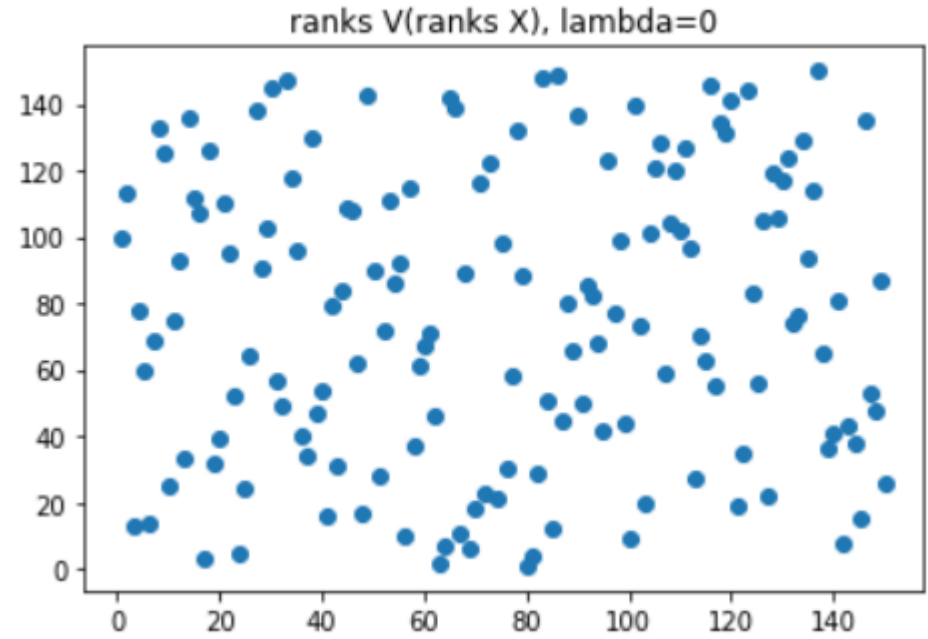
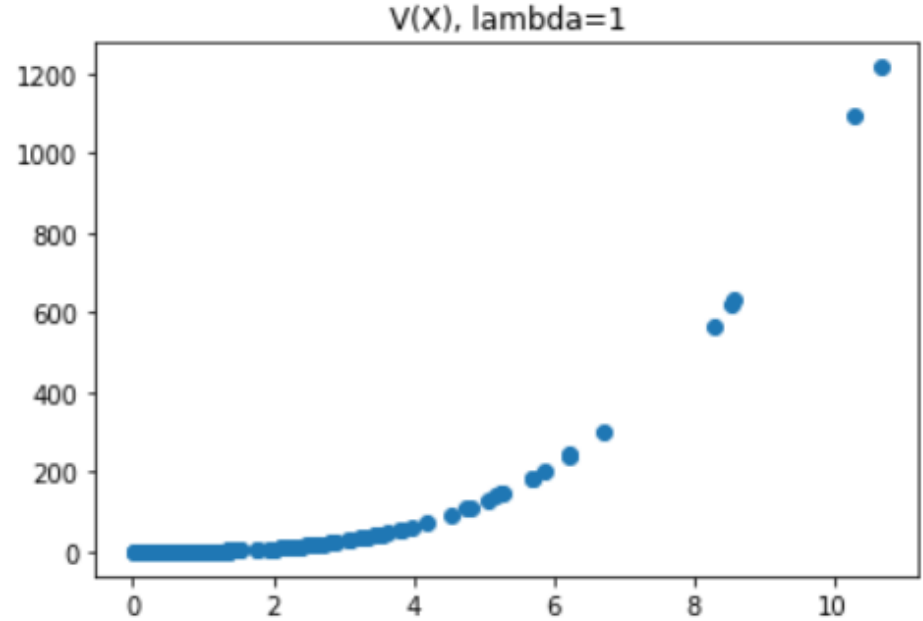
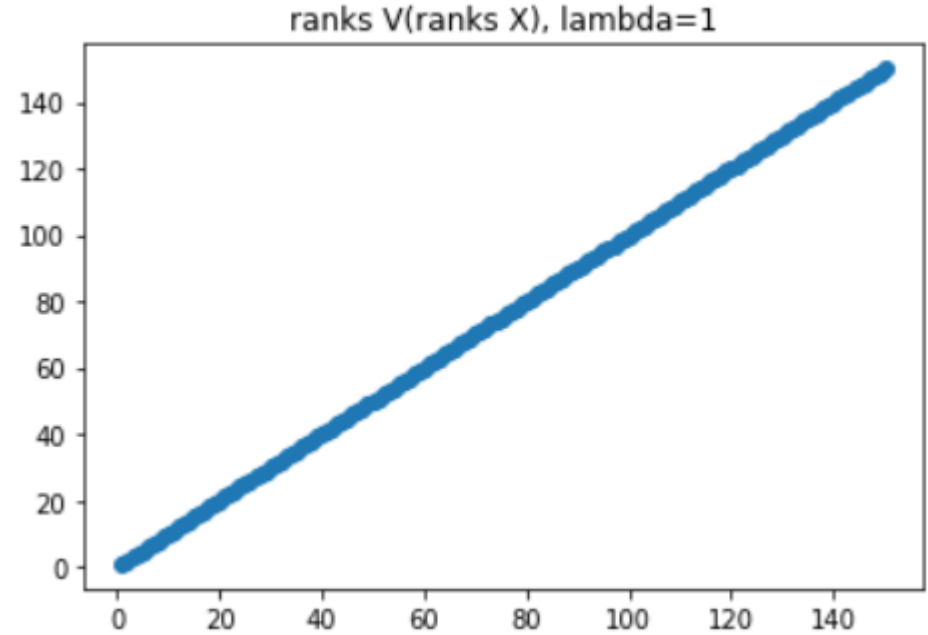

ranks V(ranks X), lambda=0
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140

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

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


V(X), lambda=1
1200
1000
800
600
400
200
0
0
2
4
6
8
10


ranks V(ranks X), lambda=1
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140

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

Выводы: При лямбда равном 0 V не зависит от X, тк ранги распределены равномерно по всей области. При лямбда равном единице есть функциональная зависимость между V и X. Также есть зависимость между рангами этих величин