Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчет по лабораторной работе № 5
по дисциплине: Математическая статика.

Выполнила студентка:
Заболотских Екатерина Дмитриевна
группа: 3630102/70301

Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2020 г.

Оглавление

Список иллюстраций

Список таблиц

Постановка задачи

Сгенерировать двумерные выборки размера 20, 60, 100 для нормального двумерного распределения $\mathrm{N}(x, y, 0, 0, 1, 1, \rho)$. Коэффициент корреляции ρ взять равным 0, 0.5, 0.9. Каждую выборку сгенерировать 1000 раз и вычислить: среднее значение, среднее значение квадрата, дисперсию коэффициентов корреляции Пирсона, Спирмена и квадратного коэффициента корреляции.

Повторить все вычисления для смеси нормальных распределений:

$$f(x, y) = 0.9 \cdot \mathrm{N}(x, y, 0, 0, 1, 1, 0.9) + \ 0.1 \ \cdot \ \mathrm{N}(x, y, 0, 0, 10, 10, -0.9).$$

Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

Теория
Двумерное нормальное распределение

Двумерная случайная величина (X, Y) называется распределенной нормально, если её плотность вероятности определена формулой:

$$\mathrm{N}(x, y, m_1, m_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right) \quad (1)$$

В свою очередь компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $m_X = m_1, m_Y = m_2$ и среднеквадратичными отклонениями $\sigma_X = \sigma_1, \sigma_Y = \sigma_2$. В свою очередь, параметр ρ – коэффициент корреляции.

Ковариация и коэффициент корреляции

Ковариацией двух случайных величин X и Y называется величина:

$$K_{XY} = M\ [\ (X - m_X)(Y - m_Y)] \quad (2)$$

В свою очередь коэффициентом корреляции называется:

$$\rho_{XY} = \frac{K_{XY}}{\sigma_X\sigma_Y} \quad (3)$$

Коэффициент корреляции характеризует зависимость между случайными величинами X и Y. Именно его мы задаем в двумерном нормальном распределении как ρ. Если случайные величины X и Y независимы, то $\rho_{XY} = 0$ т.к. в этом случае очевидно $K_{XY} = 0$.

Выборочные коэффициенты корреляции
Пирсона

Пусть по выборке значений $\{x_i, y_i\}_{i=1}^n$ двумерной случайной величины (X, Y). Естественной оценкой для ρ_{XY} служит выборочный коэффициент корреляции (Пирсона):

$$r = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \quad \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

Важное свойство: при данной оценке гипотеза $\rho_{XY} \neq 0$ может быть принята на уровне значимости 0.05 если выполнено:

$$|r|\sqrt{n-1} > 2.5 \quad (5)$$

Квадратный

Выборочным квадрантным коэффициентом корреляции называется величина:

$$r_Q = \frac{(n_1 + n_3)(n_2 - n_4)}{n} \quad (6)$$

где n_1, n_2, n_3, n_4- количества элементов выборки попавших соответственно в I, II, III и IV квадранты декартовой системы координат с центром в $(med\ x, med\ y)$ и осями

$x_1 = x - med\ x, y_1 = y - med\ y$, где med – выборочная медиана.

Формулу (6) можно переписать эквивалентным образом:

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} sign(x_i - med\ x) sign(y_i - med\ y) \quad (7)$$

Спирмена

На практике нередко требуется оценить степень взаимодействия между качественными признаками изучаемого объекта. Качественным называется признак, который нельзя измерить точно, но который позволяет сравнивать изучаемые объекты между собой и располагать их в порядке убывания или возрастания их качества. Для этого объекты выстраиваются в определённом порядке в соответствии с рассматриваемым признаком. Процесс упорядочения называется ранжированием, и каждому члену упорядоченной последовательности объектов присваивается ранг, или порядковый номер.

Например, объекту с наименьшим значением признака присваивается ранг 1, следующему за ним объекту − ранг 2, и т.д. Таким образом, происходит сравнение каждого объекта со всеми объектами изучаемой выборки. Если объект обладает не одним, а двумя качественными признаками − переменными X и Y, то для исследования их взаимосвязи используют выборочный коэффициент корреляции между двумя последовательностями рангов этих признаков.

Обозначим ранги, соответствующие значениям переменной X, через u, а ранги, соответствующие значениям переменной Y, − через v. Выборочный коэффициент ранговой корреляции Спирмена определяется как выборочный коэффициент корреляции Пирсона между рангами u, v переменных X, Y:

$$r_S = \frac{\frac{1}{n}\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum(u_i - \bar{u})^2 \frac{1}{n}\sum(v_i - \bar{v})^2}} \quad (8)$$

где $\bar{u} = \bar{v} = \frac{1+2+\cdots+n}{n} = \frac{n+1}{2}$ – среднее значение рангов.

Эллипсы рассеивания

Рассмотрим поверхность распределения, изображающую функцию (1). Она имеет вид холма, вершина которого находится над точкой $(\bar{x}, \bar{y})$.

В сечении поверхности распределения плоскостями, параллельными оси $N(x, y, x, y, \sigma x, \sigma y, \rho)$, получаются кривые, подобные нормальным кривым распределения. В сечении поверхности распределения плоскостями, параллельными плоскости xOy, получаются эллипсы. Напишем уравнение проекции такого эллипса на плоскость xOy:

$$\frac{(x-\bar{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} = const \qquad (9)$$

Уравнение эллипса (9) можно проанализировать обычными методами аналитической геометрии. Применяя их, убеждаемся, что центр эллипса (9) находится в точке с координатами ($\bar{x}$, $\bar{y}$); что касается направления осей симметрии эллипса, то они составляют с осью Ox углы, определяемые уравнением:

$$tg2\alpha = \frac{2p\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \qquad (10)$$

Это уравнение дает два значения углов: α и $\alpha 1$, различающиеся на $\frac{\pi}{2}$.

Таким образом, ориентация эллипса (9)**Error! Reference source not found.** относительно координатных осей находится в прямой зависимости от коэффициента корреляции ρ системы (X, Y); если величины не коррелированы (т.е. в данном случае и независимы), то оси симметрии эллипса параллельны координатным осям; в противном случае они составляют с координатными осями некоторый угол. Пересекая поверхность распределения плоскостями, параллельными плоскости xOy, и проектируя сечения на плоскость xOy мы получим целое семейство подобных и одинаково расположенных эллипсов с общим центром (x, y). Во всех точках каждого из таких эллипсов плотность распределения N (x, y, x, y, σx, σy, ρ) постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче эллипсами рассеивания. Общие оси всех эллипсов рассеивания называются главными осями рассеивания.

Реализация

Код программы, реализующий данную задачу, был написан на языке Python в интегрированной среде разработке PyCharm.

Были использованы библиотеки:

- **Numpy** – библиотека для работы с данными.
- **Matplotlib** – вывод графиков.

Результаты
Коэффициенты корреляции

n = 20	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.0	0.0	0.0
$E(z^2)$	0.05	0.05	0.05
$D(z)$	0.050394	0.050121	0.051036
n = 60	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.01	-0.01	-0.01
$E(z^2)$	0.016	0.016	0.018
$D(z)$	0.016349	0.015956	0.017728
n = 100	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.01	-0.0	0.0
$E(z^2)$	0.0101	0.0108	0.0102
$D(z)$	0.01010	0.010811	0.010161

Таблица 1: p = 0

n = 20	r (4)	r_s(8)	r_Q(7)
$E(z)$	0.5	0.5	0.3
$E(z^2)$	0.27	0.25	0.15
$D(z)$	0.035037	0.035054	0.044572
n = 60	r (4)	r_s(8)	r_Q(7)
$E(z)$	0.5	0.47	0.33
$E(z^2)$	0.26	0.23	0.12
$D(z)$	0.00981	0.01102	0.014527
n = 100	r (4)	r_s(8)	r_Q(7)
$E(z)$	0.5	0.48	0.33
$E(z^2)$	0.26	0.24	0.12
$D(z)$	0.005412	0.005892	0.008509

Таблица 2: p = 0.5

n = 20	r (4)	r_s(8)	r_Q(7)
$E(z)$	0.897	0.86	0.69
$E(z^2)$	0.81	0.75	0.5
$D(z)$	0.002323	0.004809	0.029723
n = 60	r (4)	r_s(8)	r_Q(7)
$E(z)$	0.899	0.881	0.71
$E(z^2)$	0.808	0.78	0.51
$D(z)$	0.00063	0.001155	0.008627
n = 100	r (4)	r_s(8)	r_Q(7)
$E(z)$	0.899	0.886	0.71
$E(z^2)$	0.809	0.786	0.51
$D(z)$	0.000407	0.000588	0.004774

Таблица 3: p = 0.9

n = 20	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.0	0.5	0.5
$E(z^2)$	0.6	0.3	0.3
$D(z)$	0.457086	0.080878	0.038778
n = 60	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.6	0.47	0.56
$E(z^2)$	0.5	0.25	0.32
$D(z)$	0.083955	0.029063	0.010997
n = 100	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.7	0.48	0.56
$E(z^2)$	0.5	0.25	0.32
$D(z)$	0.031831	0.016868	0.006436

Таблица 4: Смесь нормальных распределений

Эллипсы равновероятности

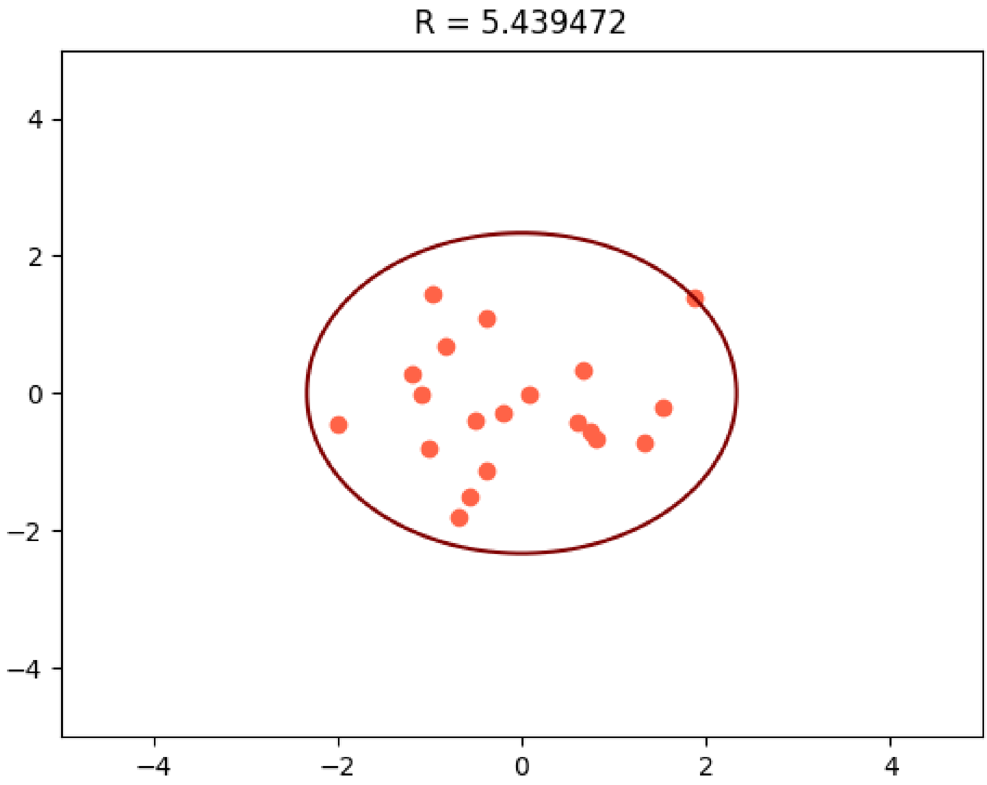

Рисунок 1: p = 0; n = 20

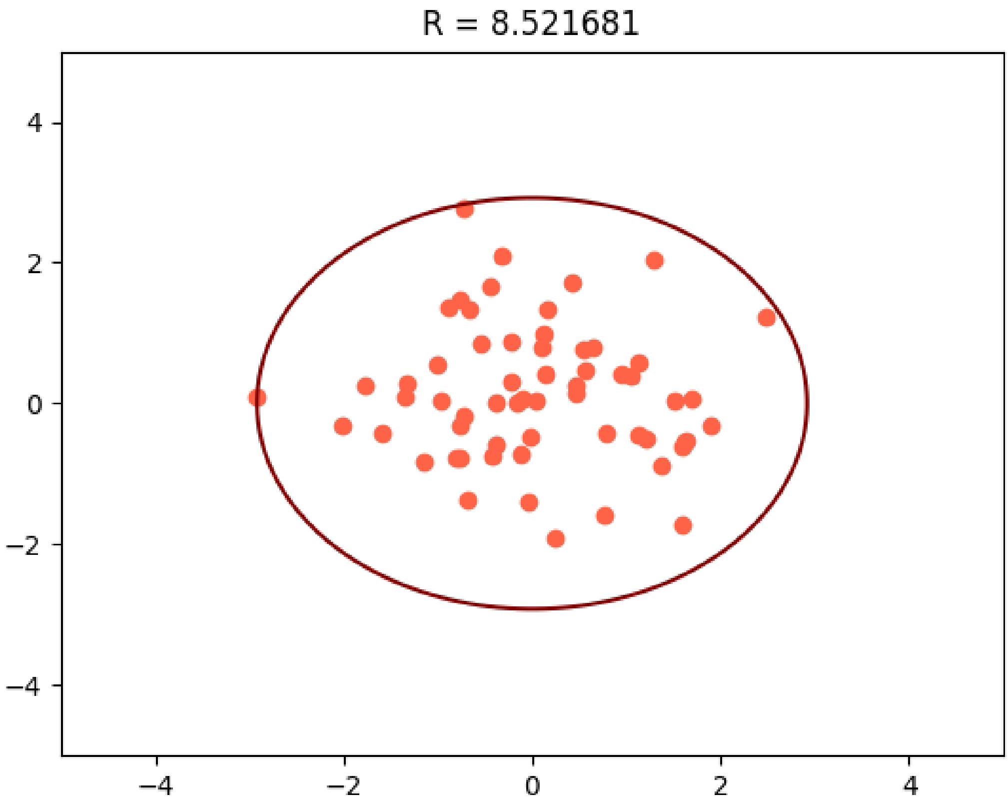

Рисунок 2: p = 0; n = 60

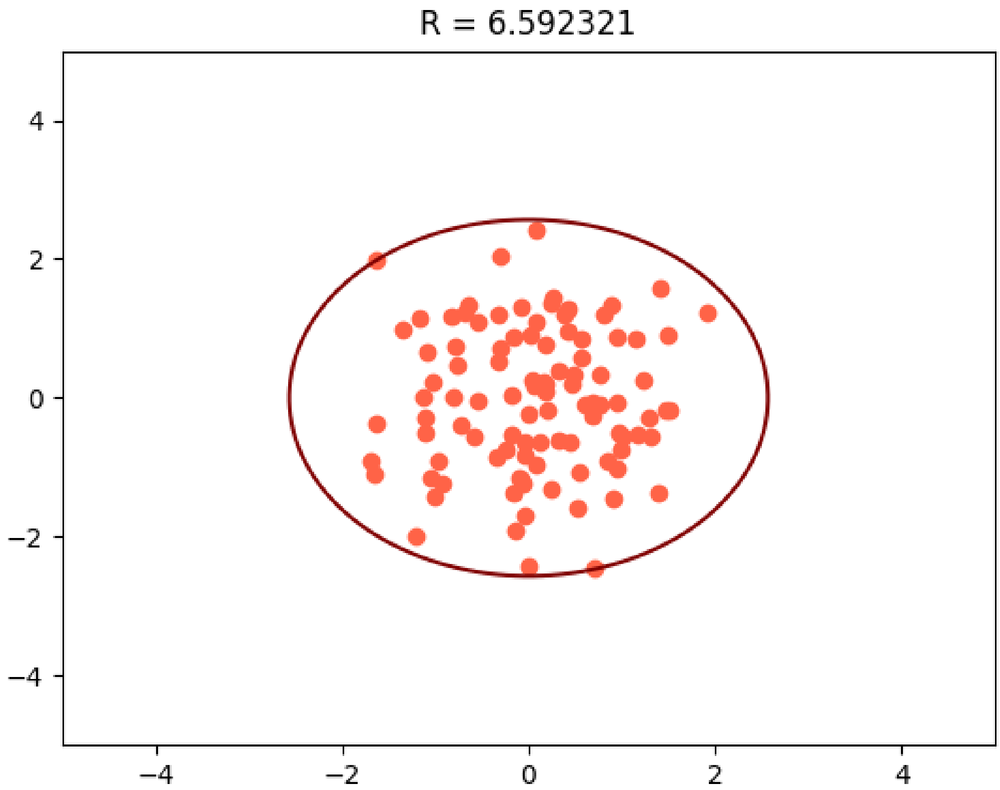

Рисунок 3: p = 0; n = 100

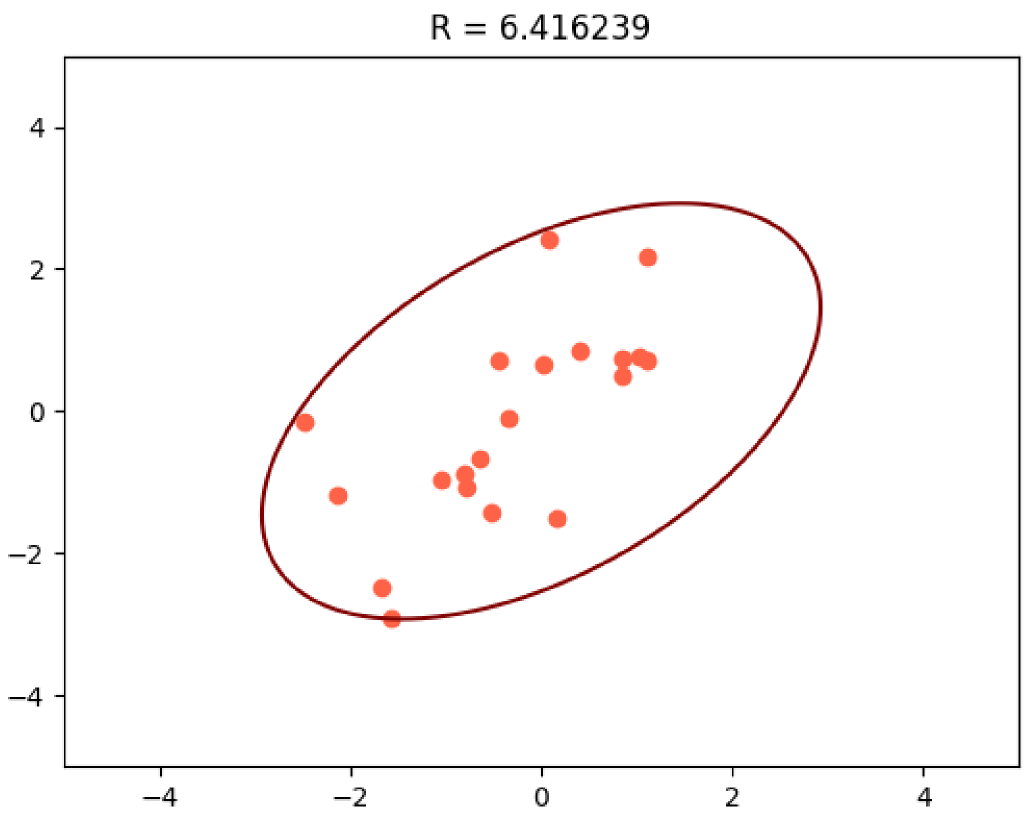

Рисунок 4: p = 0.5; n = 20

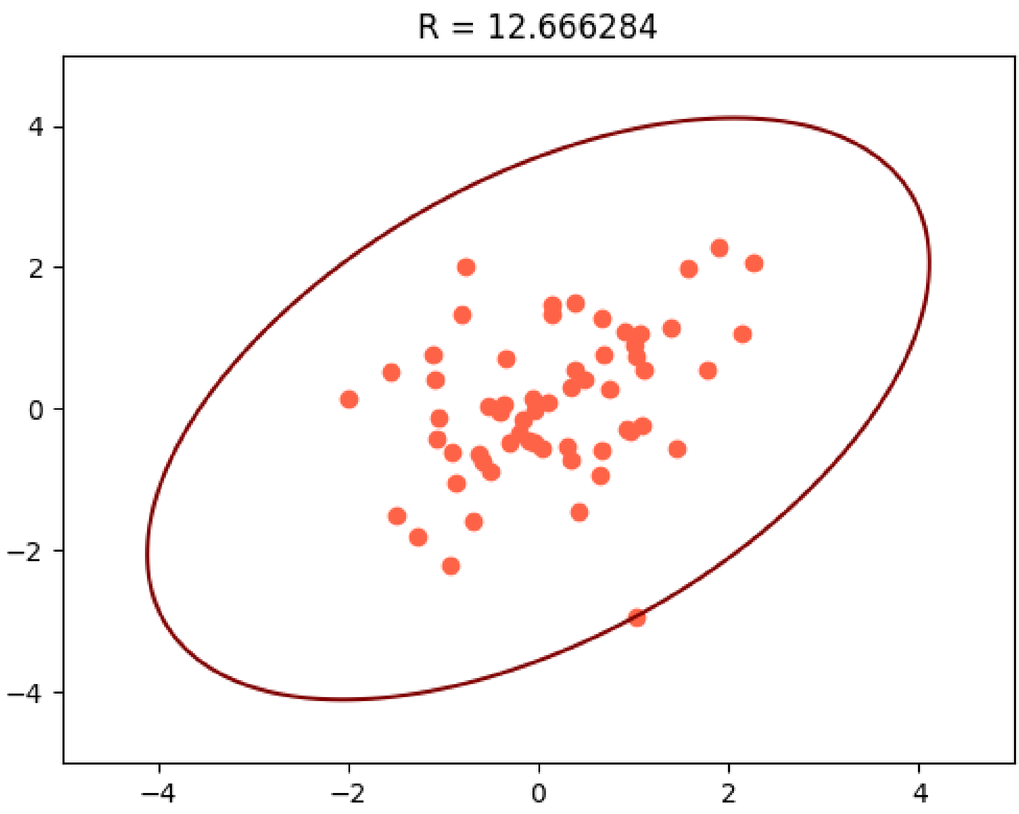

Рисунок 5: p = 0.5; n = 60

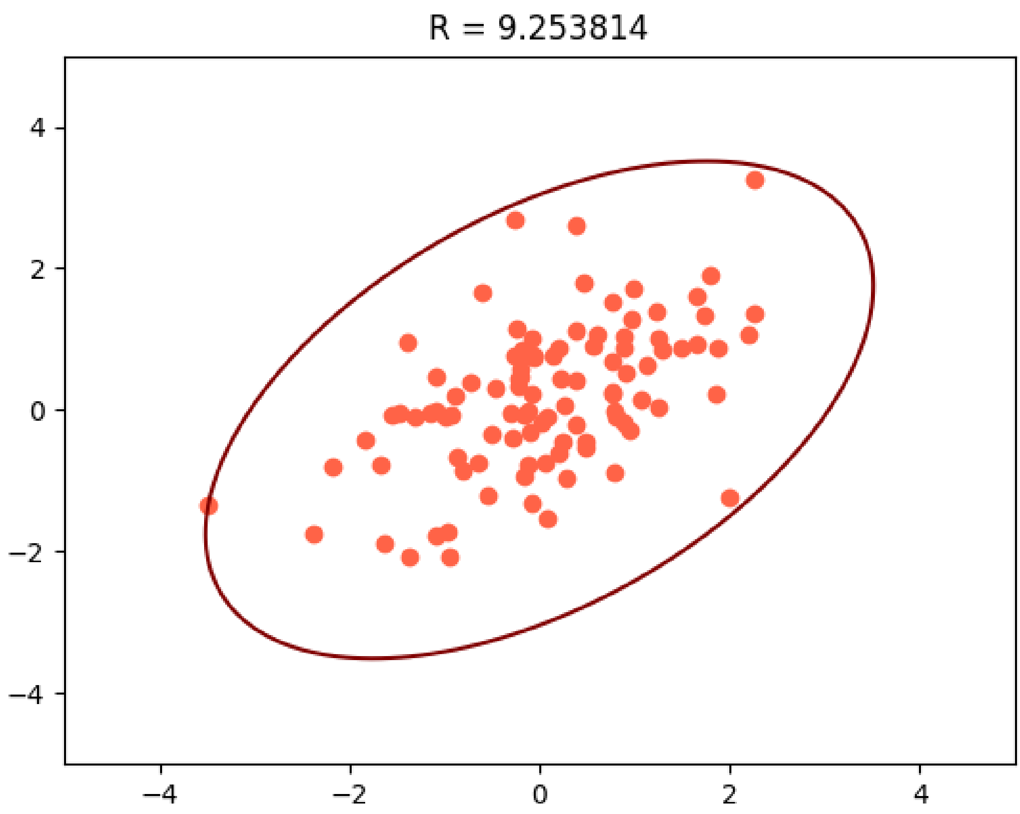

Рисунок 6: p = 0.5; n = 100

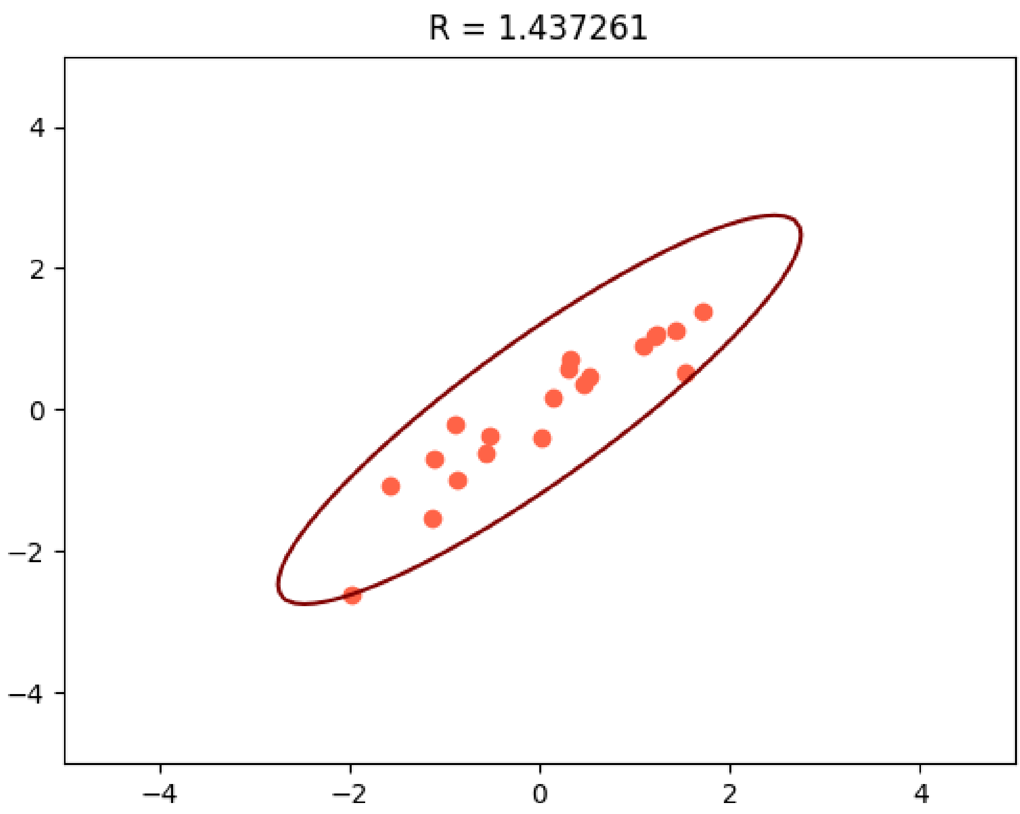

Рисунок 7: p = 0.9; n = 20

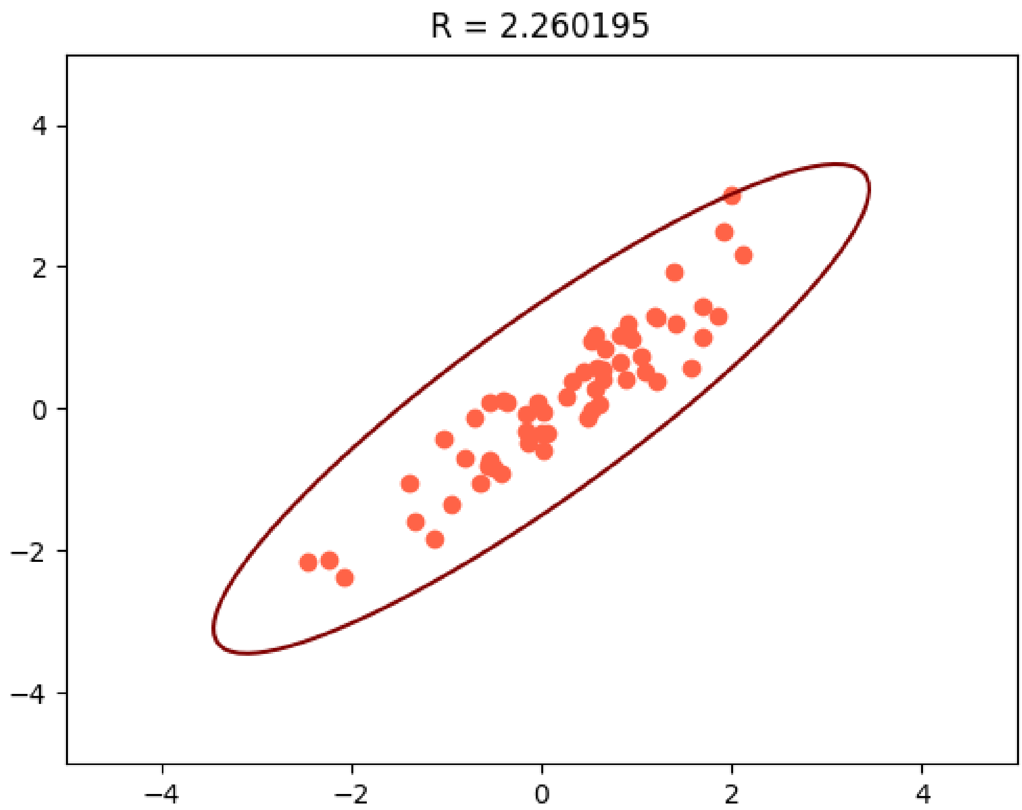

Рисунок 8: p = 0.9; n = 60

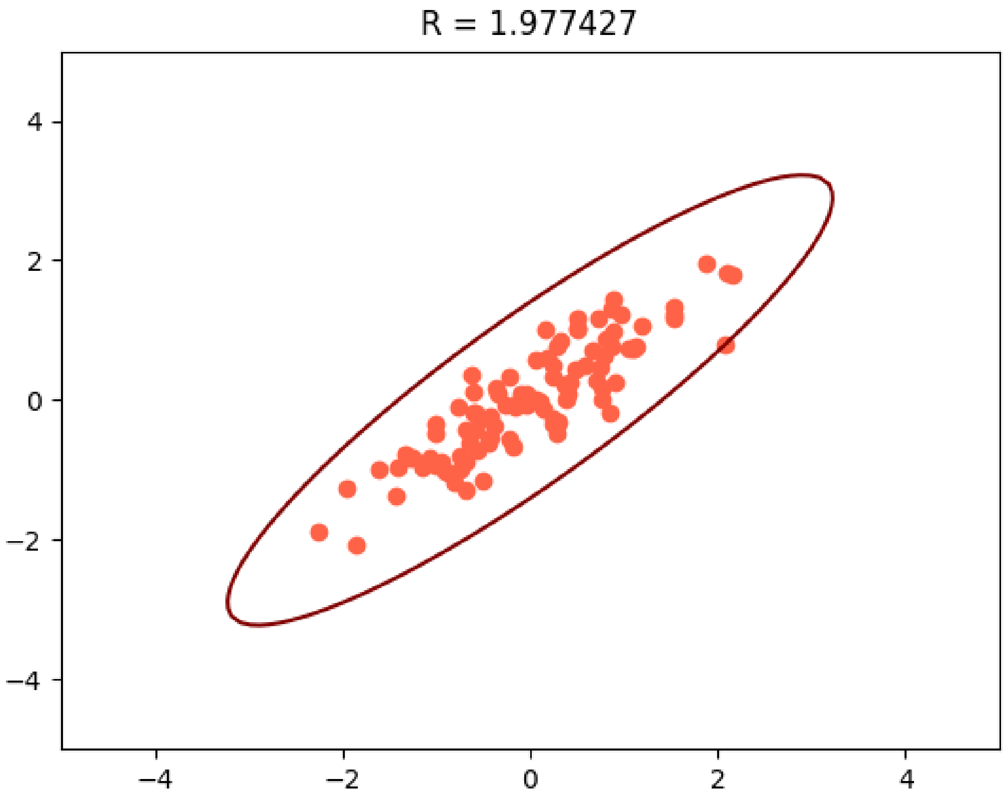

Рисунок 9: p = 0.9; n = 100

Обсуждение

Коэффициенты корреляции

Из таблиц 1, 2 и 3 видно, что r, r_S являются состоятельными оценками ρ_{XY} т.к. они все ближе к нему с ростом n.

Из таблицы 4 видим, что r_Q устойчивая к выбросам оценка. Квадрантный коэффициент корреляции показывает лучшие результаты в устойчивости.

Эллипсы равновероятности

Видно, что чем ближе ρ к 1, тем эллипс равновероятности становится все больше похож на прямую. Т.е. наглядно показано как между с.в. X и Y возникает линейная зависимость.

Список литературы

1. Конспекты лекции
2. Википедия: https://ru.wikipedia.org/wiki

Ссылка на github: https://github.com/KateZabolotskih/MathStatLabs