

Санкт-Петербургский политехнический университет  
Петра Великого

Институт прикладной математики и механики  
Кафедра «Прикладная математика»

**Отчёт**  
**по лабораторной работе №5**  
**по дисциплине**  
**«Математическая статистика»**

Выполнил студент:  
Самутичев Евгений Романович  
группа: 3630102/70201

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

Санкт-Петербург  
2020 г.

# Содержание

<b>1</b>	<b>Постановка задачи</b>	<b>2</b>
<b>2</b>	<b>Теория</b>	<b>3</b>
2.1	Двумерное нормальное распределение . . . . .	3
2.2	Ковариация и коэффициент корреляции . . . . .	3
2.3	Выборочные коэффициенты корреляции . . . . .	3
2.3.1	Пирсона . . . . .	3
2.3.2	Квадрантный . . . . .	3
2.3.3	Спирмена . . . . .	4
2.4	Эллипс равновероятности . . . . .	4
<b>3</b>	<b>Реализация</b>	<b>6</b>
<b>4</b>	<b>Результаты</b>	<b>7</b>
4.1	Коэффициенты корреляции . . . . .	7
4.2	Эллипсы равновероятности . . . . .	9
<b>5</b>	<b>Обсуждение</b>	<b>14</b>
5.1	Коэффициенты корреляции . . . . .	14
5.2	Эллипсы равновероятности . . . . .	14
<b>6</b>	<b>Приложения</b>	<b>15</b>

## Список иллюстраций

1	$\rho = 0.0, n = 20$ . . . . .	9
2	$\rho = 0.0, n = 60$ . . . . .	9
3	$\rho = 0.0, n = 100$ . . . . .	10
4	$\rho = 0.5, n = 20$ . . . . .	10
5	$\rho = 0.5, n = 60$ . . . . .	11
6	$\rho = 0.5, n = 100$ . . . . .	11
7	$\rho = 0.9, n = 20$ . . . . .	12
8	$\rho = 0.9, n = 60$ . . . . .	12
9	$\rho = 0.9, n = 100$ . . . . .	13

# 1 Постановка задачи

Сгенерировать двумерные выборки размера 20, 60, 100 для нормального двумерного распределения  $N(x, y, 0, 0, 1, 1, \rho)$ . Коэффициент корреляции  $\rho$  взять равным 0, 0.5, 0.9. Каждая выборка генерируется 1000 раз и для неё вычисляются: среднее значение, среднее значение квадрата и дисперсия коэффициентов корреляции Пирсона, Спирмена и квадрантного коэффициента корреляции.

Повторить все вычисления для смеси нормальных распределений:

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

. Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

## 2 Теория

### 2.1 Двумерное нормальное распределение

Двумерная случайная величина  $(X, Y)$  называется *распределенной нормально* (или просто *нормальной*) если её плотность вероятности определена формулой

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

Можно показать [1, стр. 133-134] что компоненты  $X, Y$  двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями  $m_X = m_1, m_Y = m_2$  и среднеквадратическими отклонениями  $\sigma_X = \sigma_1, \sigma_Y = \sigma_2$ . В свою очередь параметр  $\rho$  называют *коэффициентом корреляции*. Его значение будет раскрыто далее.

### 2.2 Ковариация и коэффициент корреляции

*Ковариацией* двух случайных величин  $X$  и  $Y$  называется величина:

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

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  $\rho$ . Если случайные величины  $X$  и  $Y$  независимы, то  $\rho_{XY} = 0$  т.к. в этом случае очевидно  $K_{XY} = 0$ .

### 2.3 Выборочные коэффициенты корреляции

#### 2.3.1 Пирсона

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

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

Важным для приложений свойством является то что при данной оценке гипотеза  $\rho_{XY} \neq 0$  (о наличии зависимости между случайными) величинами может быть принята на уровне значимости 0.05 если выполнено:

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

это можно найти к примеру в [1, стр. 538]

#### 2.3.2 Квадрантный

*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  $(\text{med } x, \text{med } y)$  и осями  $x_1 = x - \text{med } x, y_1 = y - \text{med } y$ , где  $\text{med}$  - выборочная медиана.

Формулу (6) можно переписать эквивалентным образом:

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

Важным свойством этой оценки является робастность. Её мы можем проверить используя схему засорения (смесь нормальных распределений).

### 2.3.3 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  $X, Y$ :

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

## 2.4 Эллипс равновероятности

Рассмотрим выражение для плотности двумерного нормального распределения (1) несколько подробнее, а именно найдем линии уровня или что равносильно проекции сечения графика плотности плоскостями параллельными  $xOy$  на плоскость  $xOy$ :

$$N(x, y, m_1, m_2, \sigma_1, \sigma_2, \rho) = \text{const}$$

, или что равносильно:

$$\frac{(x - m_1)^2}{\sigma_1^2} - 2\rho \frac{(x - m_1)(y - m_2)}{\sigma_1 \sigma_2} + \frac{(y - m_2)^2}{\sigma_2^2} = \text{const} \quad (9)$$

Во всех точках каждого из таких эллипсов плотность двумерного нормального распределения  $N(x, y, m_1, m_2, \sigma_1, \sigma_2, \rho)$  постоянна. Поэтому они и называются *эллипсами равновероятности* [2, стр. 44-45]. Отметим что в предельном случае  $\rho = 1$ :

$$\left( \frac{x - m_1}{\sigma_1} - \frac{y - m_2}{\sigma_2} \right)^2 = \text{const}$$

, такое уравнение задает семейство прямых параллельных прямой:

$$\frac{x - m_1}{\sigma_1} = \frac{y - m_2}{\sigma_2} \quad (10)$$

Аналогично рассматривается предельный случай  $\rho = -1$ .

### 3 Реализация

Работа выполнена с использованием языка **Python** в интегрированной среде разработки **PyCharm**, были задействованы библиотеки:

- **NumPy** - векторизация вычислений, работа с массивами данных, включая вычисление среднего и дисперсии
- **SciPy** - модуль **stats** для генерации данных по распределениям, вычисления коэффициентов корреляции
- **Matplotlib** - построение эллипсов рассеяния

Исходный код работы приведен в приложении.

## 4 Результаты

### 4.1 Коэффициенты корреляции

$n = 20$	$r$	$r_S$	$r_Q$
$E(z)$	0.0	0.0	0.0
$E(z^2)$	0.1	0.1	0.1
$D(z)$	0.053556	0.053729	0.054264
$n = 60$	$r$	$r_S$	$r_Q$
$E(z)$	0.0	0.0	0.0
$E(z^2)$	0.02	0.02	0.02
$D(z)$	0.016997	0.017097	0.018564
$n = 100$	$r$	$r_S$	$r_Q$
$E(z)$	0.0	0.0	0.0
$E(z^2)$	0.01	0.01	0.01
$D(z)$	0.010095	0.010177	0.010416

Таблица 1:  $\rho = 0$

$n = 20$	$r$	$r_S$	$r_Q$
$E(z)$	0.5	0.5	0.3
$E(z^2)$	0.3	0.2	0.2
$D(z)$	0.03309	0.036427	0.046232
$n = 60$	$r$	$r_S$	$r_Q$
$E(z)$	0.49	0.47	0.33
$E(z^2)$	0.25	0.23	0.12
$D(z)$	0.009512	0.010584	0.014097
$n = 100$	$r$	$r_S$	$r_Q$
$E(z)$	0.5	0.48	0.33
$E(z^2)$	0.25	0.23	0.12
$D(z)$	0.006043	0.00652	0.008687

Таблица 2:  $\rho = 0.5$



$n = 20$	$r$	$r_S$	$r_Q$
$E(z)$	0.891	0.86	0.7
$E(z^2)$	0.797	0.75	0.52
$D(z)$	0.002823	0.005097	0.02752
$n = 60$	$r$	$r_S$	$r_Q$
$E(z)$	0.899	0.882	0.71
$E(z^2)$	0.809	0.78	0.51
$D(z)$	0.000641	0.001065	0.008925
$n = 100$	$r$	$r_S$	$r_Q$
$E(z)$	0.898	0.885	0.71
$E(z^2)$	0.807	0.784	0.5
$D(z)$	0.000417	0.000637	0.004951

Таблица 3:  $\rho = 0.9$

$n = 20$	$r$	$r_S$	$r_Q$
$E(z)$	-0.0	0.5	0.5
$E(z^2)$	1.0	0.3	0.3
$D(z)$	0.448101	0.078135	0.039344
$n = 60$	$r$	$r_S$	$r_Q$
$E(z)$	-0.6	0.48	0.56
$E(z^2)$	0.5	0.26	0.33
$D(z)$	0.079885	0.027009	0.01148
$n = 100$	$r$	$r_S$	$r_Q$
$E(z)$	-0.7	0.47	0.56
$E(z^2)$	0.51	0.24	0.33
$D(z)$	0.029483	0.015814	0.006452

Таблица 4: Смесь нормальных распределений

## 4.2 Эллипсы равновероятности

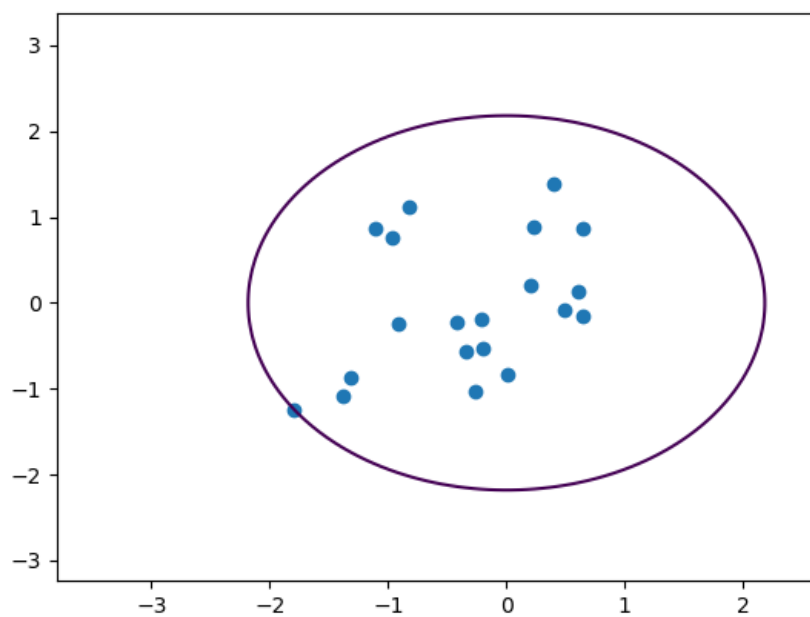


Рис. 1:  $\rho = 0.0, n = 20$

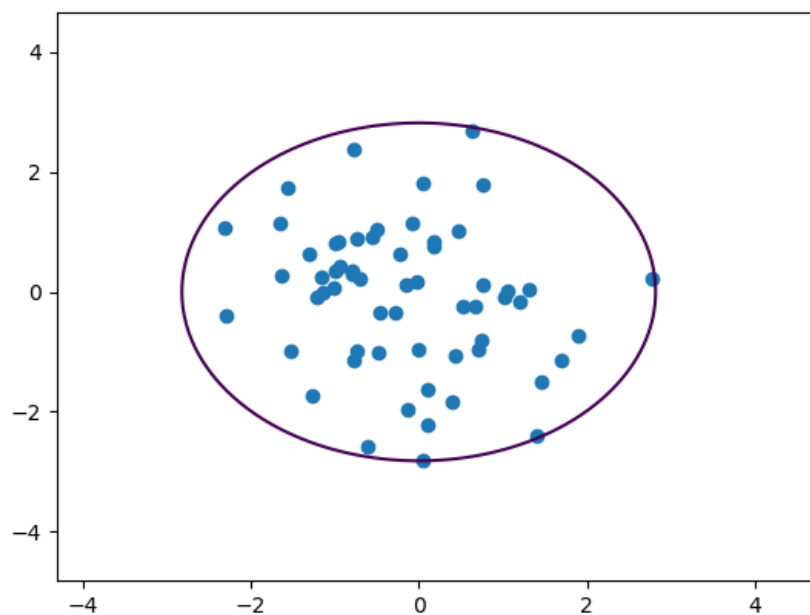


Рис. 2:  $\rho = 0.0, n = 60$

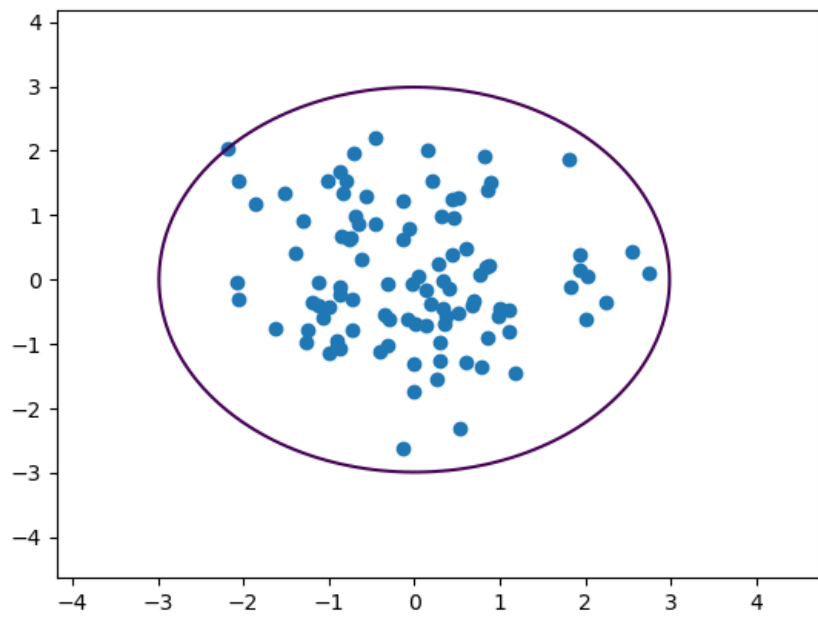


Рис. 3:  $\rho = 0.0, n = 100$

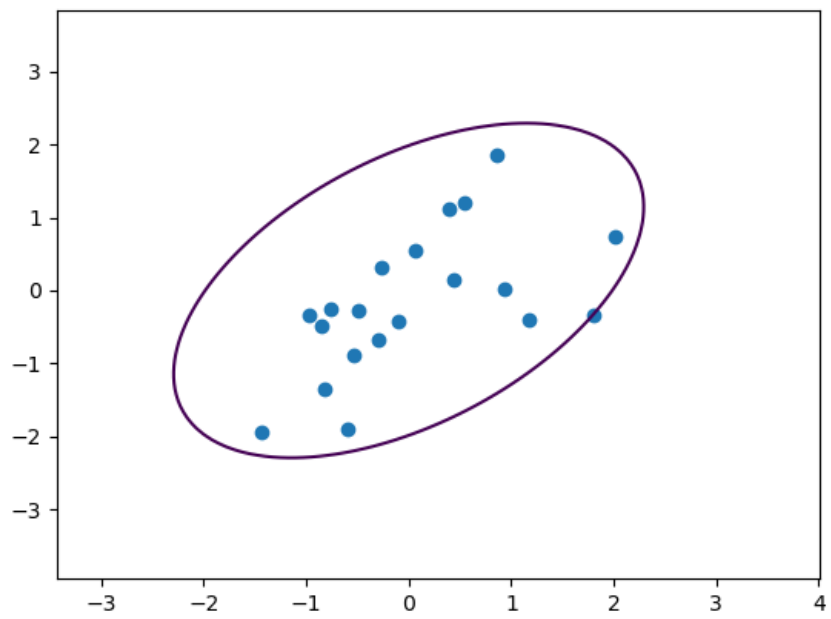


Рис. 4:  $\rho = 0.5, n = 20$

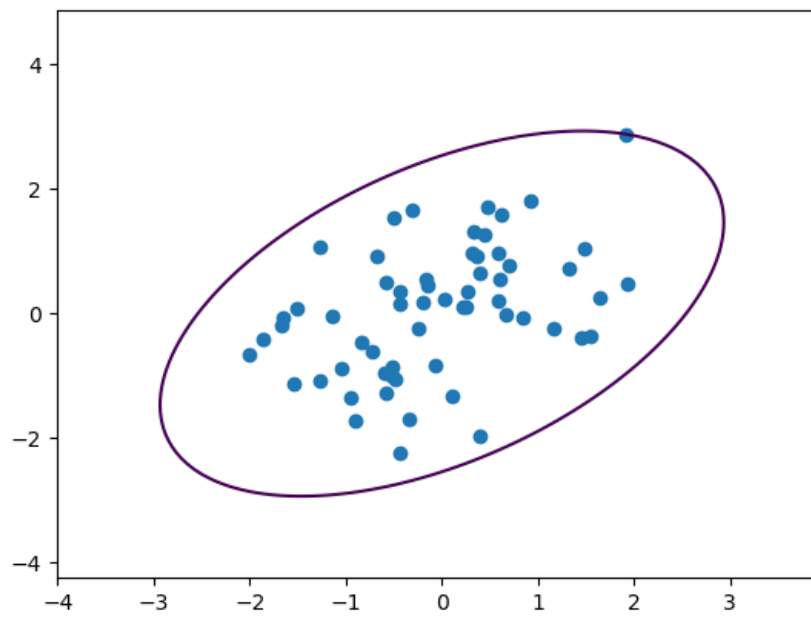


Рис. 5:  $\rho = 0.5, n = 60$

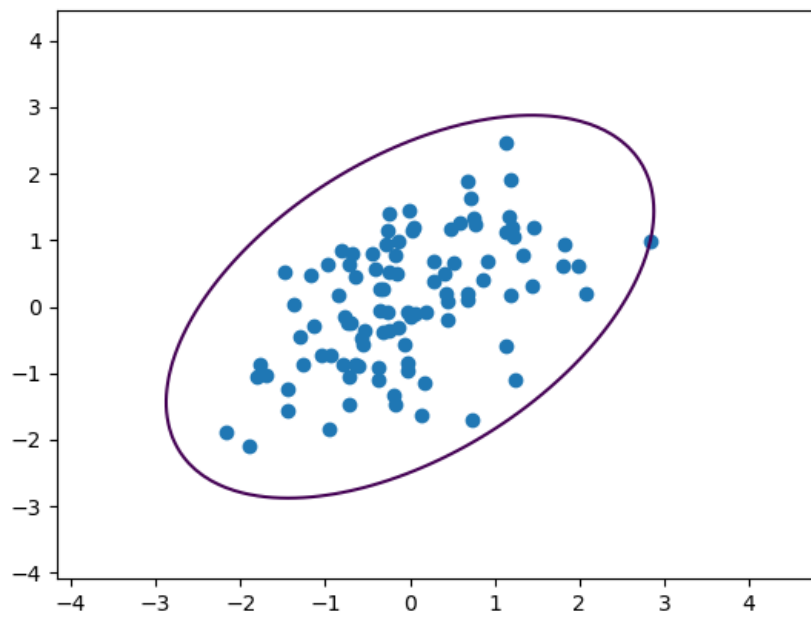


Рис. 6:  $\rho = 0.5, n = 100$

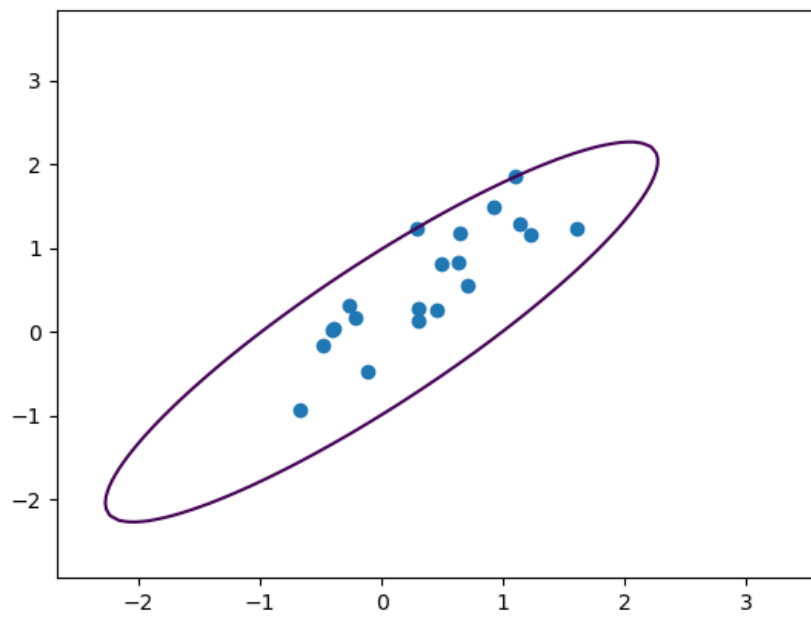


Рис. 7:  $\rho = 0.9, n = 20$

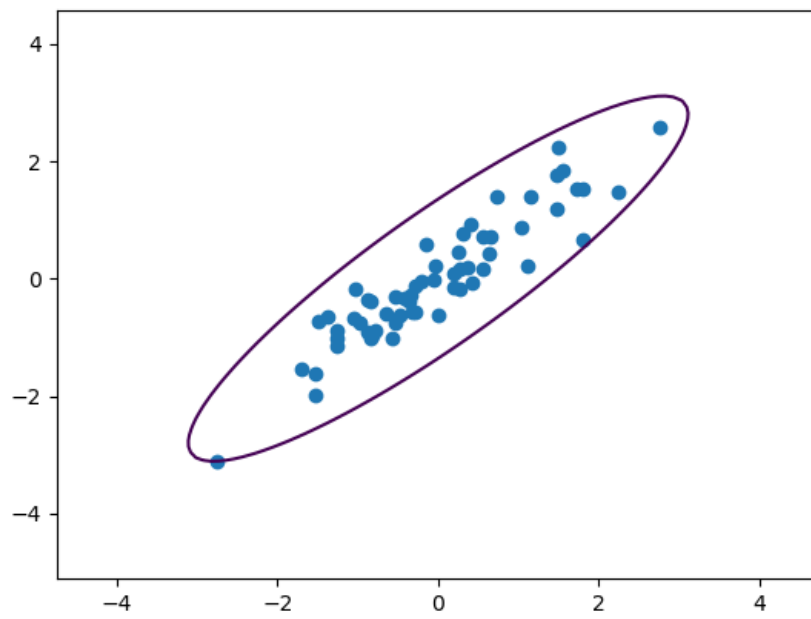


Рис. 8:  $\rho = 0.9, n = 60$

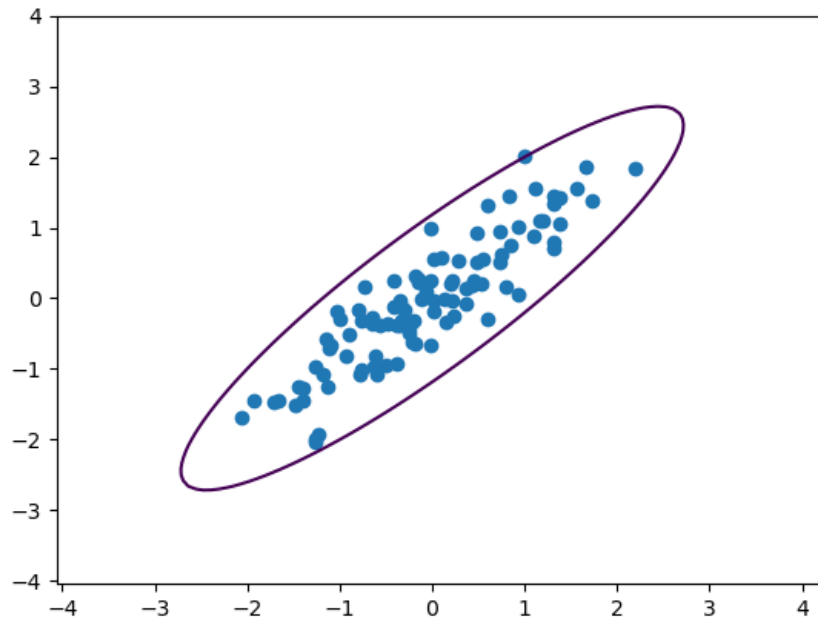


Рис. 9:  $\rho = 0.9, n = 100$

## 5 Обсуждение

### 5.1 Коэффициенты корреляции

Для начала воспользуемся (5) для анализа экспериментов по которым были получены таблицы 1, 2. Выясним можно ли принять гипотезу о зависимости между случайными величинами на уровне значимости  $\alpha = 0.05$  для  $n = 100$  по коэффициенту Пирсона.

$$0\sqrt{100-1} \leq 2.5, 4.98 \approx 0.5\sqrt{100-1} > 2.5$$

В эксперименте 1 эту гипотезу принять нельзя, а в эксперименте 2 можно. При этом в эксперименте 1 с.в. заведомо независимы, а в эксперименте 2 зависимы, так что все согласуется с теорией.

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

Из таблицы 4 видим что  $r_Q$  устойчивая к выбросам (робастная) оценка. Квадрантный коэффициент корреляции показывает лучшие результаты в устойчивости.

### 5.2 Эллипсы равновероятности

Видно что чем ближе  $\rho$  к 1, тем эллипс равновероятности становится все больше похож на прямую, заданную как (10). Т.е. наглядно показано как с.в.  $X, Y$  становятся линейно-зависимыми.

## 6 Приложения

1. Исходный код лабораторной <https://github.com/zhenyatos/statlabs/tree/master/Lab5>

## Список литературы

- [1] **Вероятностные разделы математики.** Учебник для бакалавров технических направлений. // Под ред. Максимова Ю.Д. - СПб «Иван Федоров», 2001. - 592 с., илл
- [2] Вентцель Е.С. *Теория вероятностей: Учеб. для вузов.* — 6-е изд. стер. — М.: Высш. шк., 1999.— 576 с.