

Санкт-Петербургский политехнический университет  
Высшая школа прикладной математики и  
вычислительной физики, ИПММ

Направление подготовки  
01.03.02 «Прикладная математика и информатика»

Отчет по лабораторным работам №5-6  
по дисциплине «Математическая статистика»

Выполнил студент гр. 3630102/80201

Кирпиченко С. Р.

Руководитель

Баженов А. Н.

Санкт-Петербург

2021

	Страница
<b>1 Постановка задачи</b>	<b>6</b>
<b>2 Теория</b>	<b>6</b>
2.1 Двумерное нормальное распределение . . . . .	6
2.2 Корреляционный момент и коэффициент корреляции . . .	7
2.3 Выборочные коэффициенты корреляции . . . . .	7
2.3.1 Выборочный коэффициент корреляции Пирсона . .	7
2.3.2 Выборочный квадрантный коэффициент корреляции	7
2.3.3 Выборочный коэффициент ранговой корреляции Спир- мена . . . . .	8
2.4 Эллипсы рассеивания . . . . .	8
2.5 Простая линейная регрессия . . . . .	8
2.5.1 Модель простой линейной регрессии . . . . .	8
2.5.2 Метод наименьших квадратов . . . . .	9
2.5.3 Расчётные формулы для МНК-оценок . . . . .	9
2.6 Робастные оценки коэффициентов линейной регрессии . .	9
2.7 Метод максимального правдоподобия . . . . .	10
<b>3 Реализация</b>	<b>10</b>
<b>4 Результаты</b>	<b>11</b>
4.1 Выборочные коэффициенты корреляции . . . . .	11
4.2 Эллипсы рассеивания . . . . .	13
4.3 Оценки коэффициентов линейной регрессии . . . . .	13
4.3.1 Выборка без возмущений . . . . .	13
4.3.2 Выборка с возмущениями . . . . .	14
<b>5 Обсуждение</b>	<b>15</b>
5.1 Выборочные коэффициенты корреляции и эллипсы рассеи- вания . . . . .	15

5.2	Оценки коэффициентов линейной регрессии . . . . .	16
-----	---	----

## Список иллюстраций

		Страница
1	Двумерное нормальное распределение, $n = 20$ . . . . .	13
2	Двумерное нормальное распределение, $n = 60$ . . . . .	13
3	Двумерное нормальное распределение, $n = 100$ . . . . .	13
4	Выборка без возмущений . . . . .	14
5	Выборка с возмущениями . . . . .	15

## Список таблиц

		Страница
1	Двумерное нормальное распределение, $n = 20$ . . . . .	11
2	Двумерное нормальное распределение, $n = 60$ . . . . .	11
3	Двумерное нормальное распределение, $n = 100$ . . . . .	12
4	Смесь нормальных распределений . . . . .	12

## 1 Постановка задачи

1. Сгенерировать двумерные выборки размерами 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:

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

Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

## 2 Теория

### 2.1 Двумерное нормальное распределение

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

$$N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho_{XY}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{XY}^2}} \times \\ \times \exp \left\{ -\frac{1}{2(1-\rho_{XY}^2)} \left[ \frac{(x-\bar{x})^2}{\sigma_x^2} - 2\rho_{XY}\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} \right] \right\}, \quad (1)$$

где  $\bar{x}, \bar{y}, \sigma_x, \sigma_y$  - математические ожидания и средние квадратические отклонения компонент  $X, Y$  соответственно, а  $\rho_{XY}$  - коэффициент корреляции.

## 2.2 Корреляционный момент и коэффициент корреляции

*Корреляционный момент (ковариация) двух случайных величин  $X, Y$ :*

$$K_{XY} = \text{cov}(X, Y) = \mathbf{M}[(X - \bar{x})(Y - \bar{y})]. \quad (2)$$

*Коэффициент корреляции  $\rho_{XY}$  случайных величин  $X, Y$ :*

$$\rho_{XY} = \frac{K_{XY}}{\sigma_x \sigma_y}. \quad (3)$$

*Ковариационной матрицей* случайного вектора  $(X, Y)$  называется симметричная матрица вида

$$K = \begin{pmatrix} D_X & K_{XY} \\ K_{YX} & D_Y \end{pmatrix}. \quad (4)$$

*Корреляционной матрицей* случайного вектора  $(X, Y)$  называется нормированная ковариационная матрица вида

$$R = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{YX} & 1 \end{pmatrix}. \quad (5)$$

## 2.3 Выборочные коэффициенты корреляции

### 2.3.1 Выборочный коэффициент корреляции Пирсона

*Выборочный коэффициент корреляции Пирсона:*

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{K_{XY}}{s_X s_Y}, \quad (6)$$

где  $K, s_X^2, s_Y^2$  — выборочные ковариация и дисперсии случайных величин  $X, Y$ .

### 2.3.2 Выборочный квадрантный коэффициент корреляции

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \quad (7)$$

где  $n_1, n_2, n_3, n_4$  — количества точек с координатами  $(x_i, y_i)$ , попавшими соответственно в I, II, III и IV квадранты декартовой системы с осями  $x' = x - \text{med } x$ ,  $y' = y - \text{med } y$  и с центром в точке с координатами  $(\text{med } x, \text{med } y)$ .

### 2.3.3 Выборочный коэффициент ранговой корреляции Спирмена

Обозначим ранги, соответствующие значениям переменной  $X$ , через  $u$ , а ранги, соответствующие значениям переменной  $Y$ , — через  $v$ .

*Выборочный коэффициент ранговой корреляции Спирмена:*

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

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

## 2.4 Эллипсы рассеивания

Уравнение проекции эллипса рассеивания на плоскость  $xOy$ :

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - 2\rho_{XY} \frac{(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = C, \quad C - \text{const.} \quad (9)$$

Центр эллипса (9) находится в точке с координатами  $(\bar{x}, \bar{y})$ , оси симметрии эллипса составляют с осью  $Ox$  углы, определяемые уравнением

$$\tan 2\alpha = \frac{2\rho_{XY}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \quad (10)$$

## 2.5 Простая линейная регрессия

### 2.5.1 Модель простой линейной регрессии

Регрессионную модель описания данных называют *простой линейной регрессией*, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (11)$$



где  $x_1, \dots, x_n$  — заданные числа (значения фактора);  $y_1, \dots, y_n$  — наблюдаемые значения отклика;  $\varepsilon_1, \dots, \varepsilon_n$  — независимые, нормально распределенные  $N(0, \sigma)$  с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые);  $\beta_0, \beta_1$  — неизвестные параметры, подлежащие оцениванию.

### 2.5.2 Метод наименьших квадратов

*Метод наименьших квадратов* (МНК):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \rightarrow \min_{\beta_0, \beta_1}. \quad (12)$$

### 2.5.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров  $\beta_0$  и  $\beta_1$ :

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad (13)$$

$$\hat{\beta}_0 = \bar{y} - \bar{x} \hat{\beta}_1. \quad (14)$$

## 2.6 Робастные оценки коэффициентов линейной регрессии

*Метод наименьших модулей*:

$$\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \rightarrow \min_{\beta_0, \beta_1}. \quad (15)$$

$$\hat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \quad (16)$$

$$\hat{\beta}_{0R} = \text{med } y - \hat{\beta}_{1R} \text{med } x, \quad (17)$$

$$r_Q = \frac{1}{n} \sum_{i=1}^n \text{sign}(x_i - \text{med } x) \text{sign}(y_i - \text{med } y), \quad (18)$$

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)}, \quad q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)} \quad (19)$$

$$l = \begin{cases} [n/4] + 1 & \text{при } n/4 \text{ дробном,} \\ n/4 & \text{при } n/4 \text{ целом.} \end{cases}$$

$$j = n - l + 1.$$

$$\text{sign } z = \begin{cases} 1 & \text{при } z > 0, \\ 0 & \text{при } z = 0, \\ -1 & \text{при } z < 0. \end{cases}$$

Уравнение регрессии здесь имеет вид

$$y = \hat{\beta}_{0R} + \hat{\beta}_{1R} \cdot x. \quad (20)$$

$$k_q(20) = 1.491.$$

## 2.7 Метод максимального правдоподобия

$L(x_1, \dots, x_n, \theta)$  — функция правдоподобия(ФП), рассматриваемая как функция неизвестного параметра  $\theta$ :

$$L(x_1, \dots, x_n, \theta) = f(x_1, \theta)f(x_2, \theta) \dots f(x_n, \theta). \quad (21)$$

## 3 Реализация

Лабораторная работа выполнена на языке Python 3.9 с использованием библиотек numpy, scipy, matplotlib, seaborn.

## 4 Результаты

### 4.1 Выборочные коэффициенты корреляции

$\rho = 0$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.0026	0.003	0.0029
$E(z^2)$	0.0537	0.055	0.0555
$D(z)$	0.0537	0.055	0.0554
$\rho = 0.5$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.4865	0.46	0.338
$E(z^2)$	0.2679	0.246	0.1576
$D(z)$	0.0313	0.034	0.0433
$\rho = 0.9$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.8948	0.866	0.7068
$E(z^2)$	0.803	0.755	0.5259
$D(z)$	0.0024	0.004	0.0263

Таблица 1: Двумерное нормальное распределение,  $n = 20$

$\rho = 0$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.0029	0.002	0.0018
$E(z^2)$	0.0167	0.017	0.0172
$D(z)$	0.0167	0.017	0.0172
$\rho = 0.5$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.5004	0.479	0.3371
$E(z^2)$	0.2601	0.24	0.1277
$D(z)$	0.0097	0.01	0.0141
$\rho = 0.9$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.8974	0.882	0.7121
$E(z^2)$	0.8059	0.779	0.5151
$D(z)$	0.0006	0.001	0.008

Таблица 2: Двумерное нормальное распределение,  $n = 60$

$\rho = 0$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.002	0.002	-0.0015
$E(z^2)$	0.0098	0.01	0.0106
$D(z)$	0.0098	0.01	0.0106
$\rho = 0.5$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.4989	0.478	0.3317
$E(z^2)$	0.2547	0.235	0.1188
$D(z)$	0.0059	0.006	0.0088
$\rho = 0.9$			
	$r$	$r_S$	$r_Q$
$E(z)$	0.8989	0.886	0.7107
$E(z^2)$	0.8084	0.785	0.5098
$D(z)$	0.0004	0.001	0.0048

Таблица 3: Двумерное нормальное распределение,  $n = 100$

$n = 20$			
	$r$	$r_S$	$r_Q$
$E(z)$	-0.1029	-0.095	-0.0678
$E(z^2)$	0.0588	0.057	0.0526
$D(z)$	0.0482	0.048	0.048
$n = 60$			
	$r$	$r_S$	$r_Q$
$E(z)$	-0.087	-0.083	-0.0567
$E(z^2)$	0.0234	0.023	0.0204
$D(z)$	0.0158	0.016	0.0172
$n = 100$			
	$r$	$r_S$	$r_Q$
$E(z)$	-0.0935	-0.089	-0.0589
$E(z^2)$	0.0192	0.018	0.0134
$D(z)$	0.0104	0.01	0.0099

Таблица 4: Смесь нормальных распределений

## 4.2 Эллипсы рассеивания

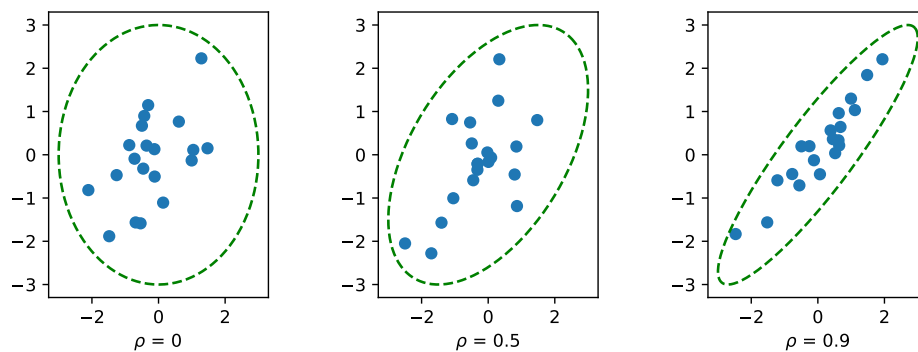


Рис. 1: Двумерное нормальное распределение,  $n = 20$

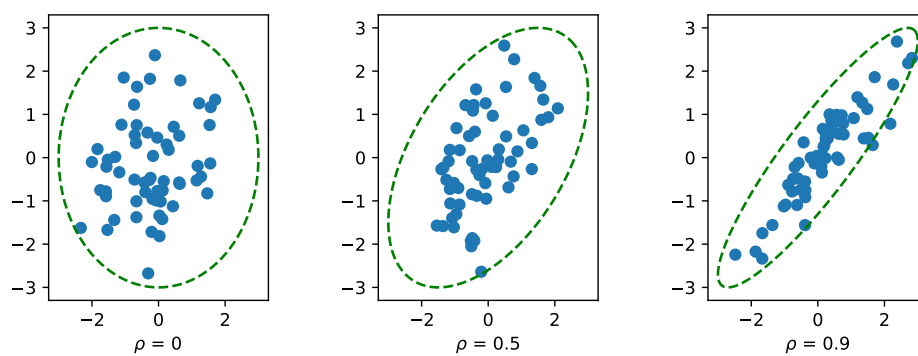


Рис. 2: Двумерное нормальное распределение,  $n = 60$

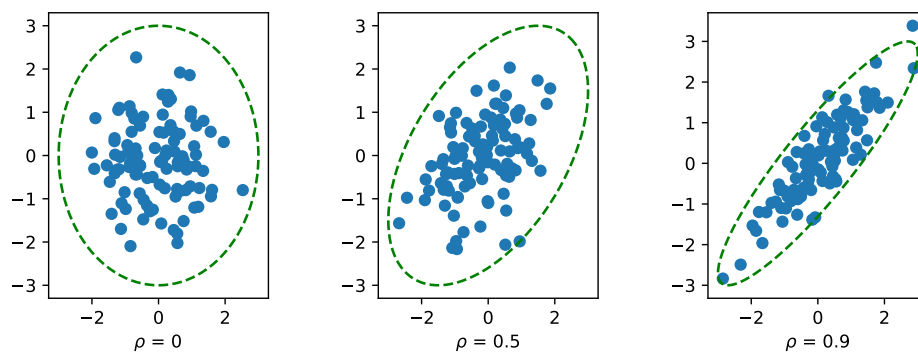


Рис. 3: Двумерное нормальное распределение,  $n = 100$

## 4.3 Оценки коэффициентов линейной регрессии

### 4.3.1 Выборка без возмущений

Коэффициенты прямых:

1. Метод наименьших квадратов:  $\hat{\beta}_1 = 2.1838$ ,  $\hat{\beta}_0 = 2.3362$ ;
2. Метод наименьших модулей:  $\hat{\beta}_1 = 2.0006$ ,  $\hat{\beta}_0 = 2.4235$ .

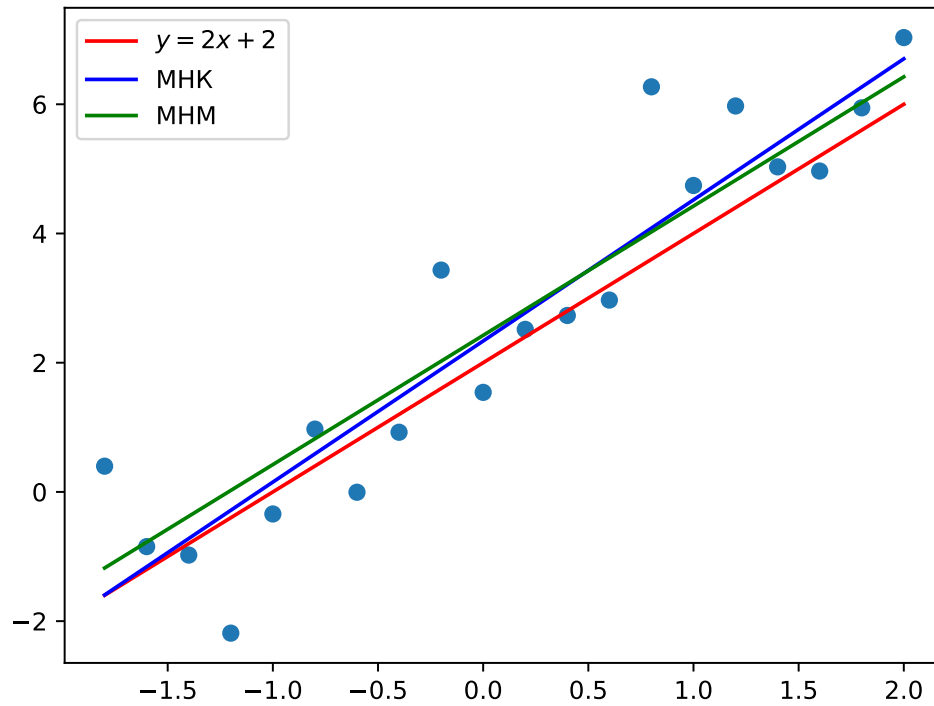


Рис. 4: Выборка без возмущений

#### 4.3.2 Выборка с возмущениями

Коэффициенты прямых:

1. Метод наименьших квадратов:  $\hat{\beta}_1 = 0.5469$ ,  $\hat{\beta}_0 = 1.8807$ ;
2. Метод наименьших модулей:  $\hat{\beta}_1 = 0.8314$ ,  $\hat{\beta}_0 = 1.8622$ .

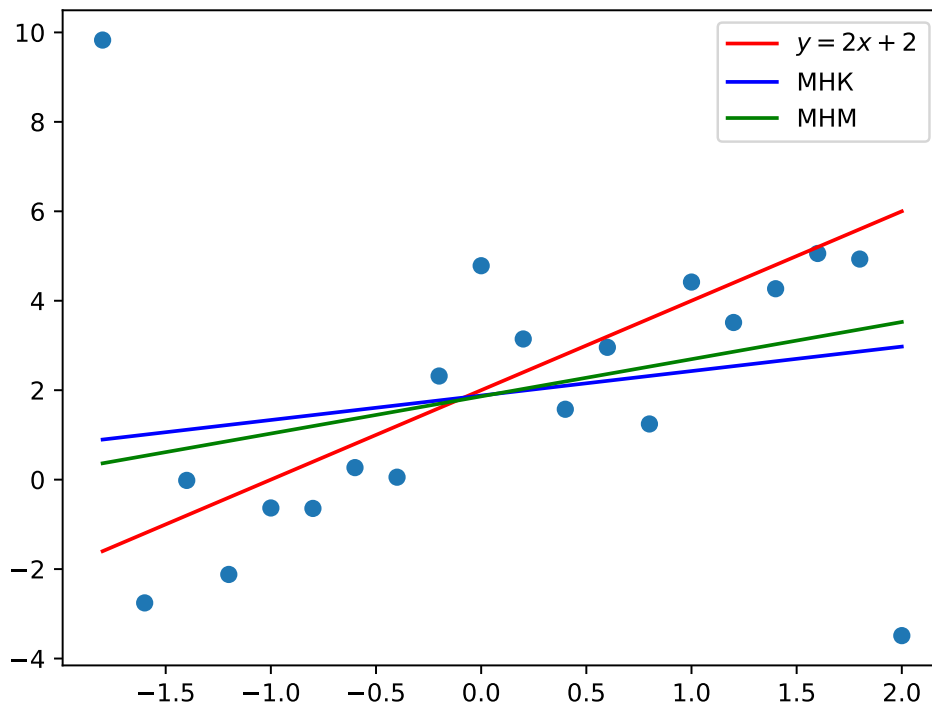


Рис. 5: Выборка с возмущениями

## 5 Обсуждение

### 5.1 Выборочные коэффициенты корреляции и эллипсы рассеивания

Для дисперсий выборочных коэффициентов корреляции можно сделать следующие выводы:

1. Для двумерного нормального распределения справедлив следующий порядок:  $D(r) \leq D(r_S) \leq D(r_Q)$ . Коэффициент Пирсона является оптимальным для анализа подобных выборок.
2. Для смеси нормальных распределений дисперсии всех трех коэффициентов примерно равны.

Процент попадания элементов выборки в эллипс рассеивания примерно равен теоретическому значению (95%).

## 5.2 Оценки коэффициентов линейной регрессии

Для выборки без возмущений методы наименьших квадратов и модулей дают схожие хорошие результаты, однако МНМ дает более параллельную к исходной прямую.

Для выборки с возмущениями МНК и МНМ также дают схожие прямые, ввиду рода возмущений коэффициент наклона сильно отличается от эталона, однако метод наименьших модулей показал большую устойчивость.



## Примечание

## Список литературы

- [1] Максимов Ю.Д. Математика. Теория и практика по математической статистике. Конспект-справочник по теории вероятностей : учеб. пособие / Ю.Д. Максимов; под ред. В.И. Антонова. — СПб. : Изд-во Политехн. ун-та, 2009. — 395 с. (Математика в политехническом университете).