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

Содержание

Список иллюстраций

Страница

Список таблиц

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 ρ 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

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

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

2.2 Корреляционный момент и коэффициент корреляции

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

$$K_{XY} = \operatorname{cov}(X, Y) = \mathbf{M}\left[(X - \overline{x})(Y - \overline{y})\right]. \tag{2}$$

Коэффициент корреляции ρ_{XY} случайных величин X, Y:

$$\rho_{XY} = \frac{K_{XY}}{\sigma_x \sigma_y}. \tag{3}$$

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

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

Кореляционной матрицей случайного вектора (X, Y) называется нормированная ковариационная матрица вида

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

2.3 Выборочные коэффициенты корреляции

2.3.1 Выборочный коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона:

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

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

2.3.2 Выборочный квадрантный коэффициент корреляции

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

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 $x' = x - \operatorname{med} x$, $y' = y - \operatorname{med} y$ и с центром в точке с координатами $(\operatorname{med} x, \operatorname{med} y)$.

2.3.3 Выборочный коэффициент ранговой корреляции Спирмена

Обозначим ранги, соотвествующие значениям переменной X, через u, а ранги, соответствующие значениям переменной Y, – через v.

Выборочный коэффициент ранговой корреляции Спирмена:

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

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

2.4 Эллипсы рассеивания

Уравнение проекции эллипса рассеивания на плоскость xOy:

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

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

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

2.5 Простая линейная регрессия

2.5.1 Модель простой линейной регрессии

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, ..., n, \tag{11}$$

где $x_1, ..., x_n$ – заданные числа (значения фактора); $y_1, ..., y_n$ – наблюдаемые значения отклика; $\varepsilon_1, ..., \varepsilon_n$ – независимые, нормально распределенные $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1 – неизвестные параметры, подлежащие оцениванию.

2.5.2 Метод наименьших квадратов

Метод наименьших квадратов (МНК):

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

2.5.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров β_0 и β_1:

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

$$\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1. \tag{14}$$

2.6 Робастные оценки коэффициентов линейной регрессии

Метод наименьших модулей:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1}. \tag{15}$$

$$\widehat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \tag{16}$$

$$\widehat{\beta}_{0R} = \operatorname{med} y - \widehat{\beta}_{1R} \operatorname{med} x, \tag{17}$$

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

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

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

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

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

Уравнение регрессии здесь имеет вид

$$y = \widehat{\beta}_{0R} + \widehat{\beta}_{1R} \cdot x. \tag{20}$$

$$k_q(20) = 1.491.$$

2.7 Метод максимального правдоподобия

$L(x_1, ..., x_n, \theta)$ – функция правдоподобия(ФП), рассматриваемая как функция неизвестного параметра θ:

$$L(x_1, ..., x_n, \theta) = f(x_1, \theta) f(x_2, \theta) ... f(x_n, \theta). \tag{21}$$

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

Таблица 1: Двумерное нормальное распределение, n = 20

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

При построении всех эллипсов использовался параметр $C = 3$.

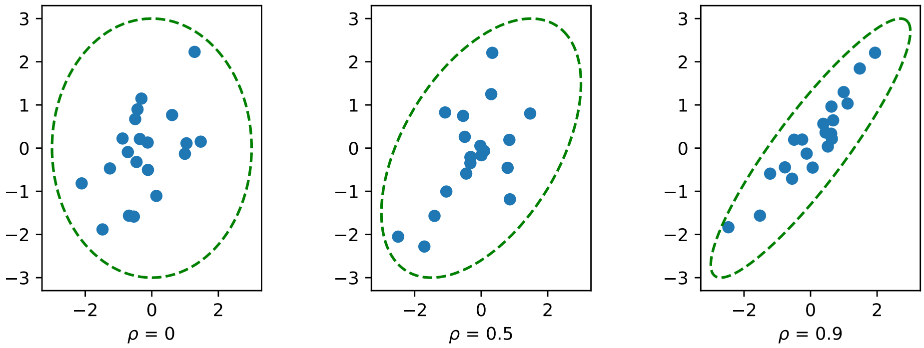

Рис. 1: Двумерное нормальное распределение, n = 20

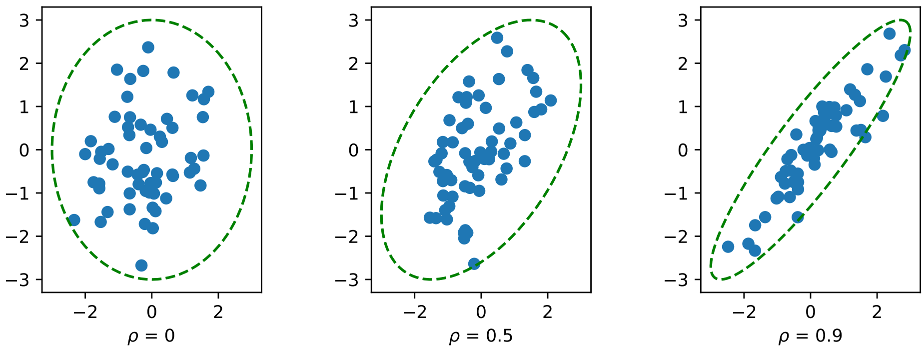

Рис. 2: Двумерное нормальное распределение, n = 60

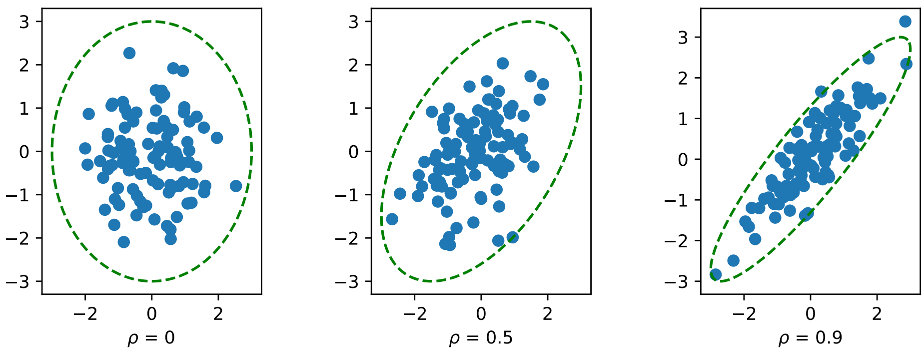

Рис. 3: Двумерное нормальное распределение, n = 100

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: $\hat{\beta}_1 = 2.0006$, $\hat{\beta}_0 = 2.4235$.

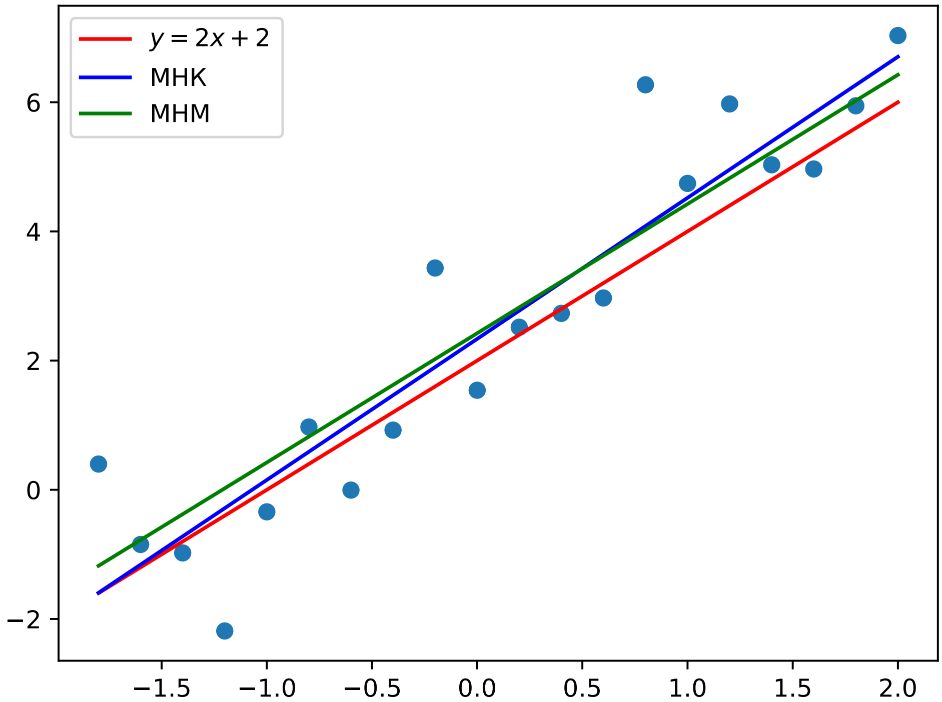

Рис. 4: Выборка без возмущений

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: $\hat{\beta}_1 = 0.8314$, $\hat{\beta}_0 = 1.8622$.

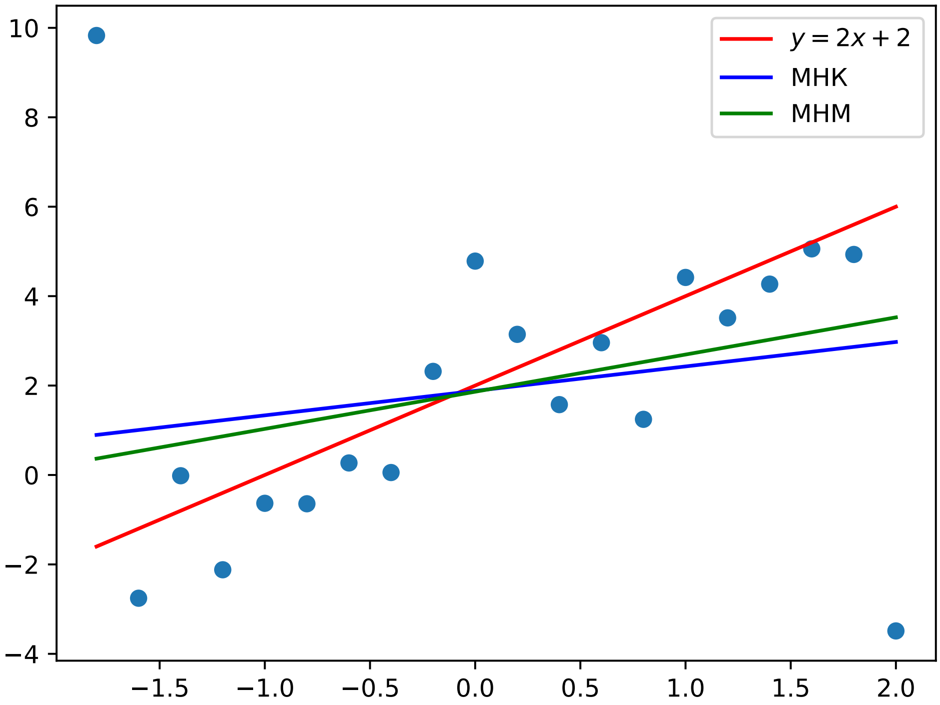

Рис. 5: Выборка с возмущениями

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

С исходным кодом работы и данного отчета можно ознакомиться в репозитории `https://github.com/Stasychbr/MatStat`

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).