САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Институт прикладной математики и механики

Высшая школа прикладной математики и вычислительной физики

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

Выполнил студент:
Колосков Александр
группа: 3630102/80301

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

Санкт-Петербург
2021 г.

Содержание

Список иллюстраций

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:

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

   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

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

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:

$$K_{XY} = \operatorname{cov}(X, Y) = \mathbf{M}\left[(X - \overline{x})(Y - \overline{y})\right]. \tag{3}$$

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

$$\rho_{XY} = \frac{K_{XY}}{\sigma_x\sigma_y}. \tag{4}$$

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

$$K = \begin{pmatrix} D_X & K_{XY} \\ K_{YX} & D_Y \end{pmatrix}. \tag{5}$$

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

$$R = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{YX} & 1 \end{pmatrix}. \tag{6}$$

2.3 Выборочные коэффициенты корреляции

2.3.1 Выборочный коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{K_{XY}}{s_X s_Y}, \tag{7}$$

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

2.3.2 Выборочный квадрантный коэффициент корреляции

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \tag{8}$$

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:

$$r_S = \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})^2 \frac{1}{n}\sum_{i=1}^{n}(v_i - \overline{v})^2}}, \tag{9}$$

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:

$$\frac{(x-\overline{x})^2}{\sigma_x^2} - 2\rho_{XY}\frac{(x-\overline{x})(y-\overline{y})}{\sigma_x\sigma_y} + \frac{(y-\overline{y})^2}{\sigma_y^2} = C, \;\; C - \text{const}. \tag{10}$$

Центр эллипса (10) находится в точке с координатами $(\overline{x}, \overline{y})$, оси симметрии эллипса составляют с осью Ox углы, определяемые уравнением

$$\tan 2\alpha = \frac{2\rho_{XY}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \tag{11}$$

2.5 Простая линейная регрессия

2.5.1 Модель простой линейной регрессии

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; i = 1, ..., n, \tag{12}$$

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):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0, \beta_1}. \tag{13}$$

2.5.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров β_0 и β_1:

$$\widehat{\beta}_1 = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - (\overline{x})^2}, \tag{14}$$

$$\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1. \tag{15}$$

2.6 Робастные оценки коэффициентов линейной регрессии

Метод наименьших модулей:

$$\sum_{i=1}^{n}|y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0,\beta_1}. \tag{16}$$

$$\widehat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \tag{17}$$

$$\widehat{\beta}_{0R} = \operatorname{med} y - \widehat{\beta}_{1R} \operatorname{med} x, \tag{18}$$

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}(x_i - \operatorname{med} x) \operatorname{sign}(y_i - \operatorname{med} y), \tag{19}$$

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)}, \quad q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)} \tag{20}$$

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

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

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

Уравнение регрессии здесь имеет вид

$$y = \widehat{\beta}_{0R} + \widehat{\beta}_{1R} \cdot x. \tag{21}$$

$$k_q(20) = 1.491.$$

3 Реализация

Лабораторная работа выполнена на языке Python в среде PyCharm с использованием библиотек numpy, scipy.stats, matplotlib.pyplot, statsmodels

4 Результаты

4.1 Выборочные коэффициенты корреляции

n=20	$r(7)$	$r_S(9)$	$r_Q(8)$
$E(z)$	-0.0077	-0.0037	0.0067
$E(z^2)$	0.0498	0.0498	0.0497
$D(z)$	0.0497	0.0497	0.0496
n=60	r	r_S	r_Q
$E(z)$	-0.0057	-0.0066	-0.0168
$E(z^2)$	0.0535	0.0529	0.0507
$D(z)$	0.0535	0.0528	0.0504
n=100	r	r_S	r_Q
$E(z)$	-0.0048	-0.0069	-0.0088
$E(z^2)$	0.0517	0.051	0.056
$D(z)$	0.0517	0.051	0.056

Таблица 1: Выборочные коэффициенты корреляции двумерного нормального распределения. $\rho = 0(4)$

n=20	$r(7)$	$r_S(9)$	$r_Q(8)$
$E(z)$	0.4922	0.4701	0.3296
$E(z^2)$	0.252	0.2318	0.1233
$D(z)$	0.0098	0.0108	0.0147
n=60	r	r_S	r_Q
$E(z)$	0.4986	0.4768	0.334
$E(z^2)$	0.2574	0.2379	0.1269
$D(z)$	0.0088	0.0106	0.0153
n=100	r	r_S	r_Q
$E(z)$	0.4988	0.4744	0.329
$E(z^2)$	0.2584	0.2359	0.1237
$D(z)$	0.0097	0.0109	0.0155

Таблица 2: Выборочные коэффициенты корреляции двумерного нормального распределения. $\rho = 0.5(4)$

n=20	$r(7)$	$r_S(9)$	$r_Q(8)$
$E(z)$	0.8997	0.8869	0.7142
$E(z^2)$	0.81	0.7872	0.5151
$D(z)$	0.0004	0.0007	0.005
n=60	r	r_S	r_Q
$E(z)$	0.8977	0.8843	0.7133
$E(z^2)$	0.8062	0.7826	0.5138
$D(z)$	0.0004	0.0007	0.0051
n=100	r	r_S	r_Q
$E(z)$	0.8985	0.8854	0.7118
$E(z^2)$	0.8076	0.7846	0.512
$D(z)$	0.0004	0.0006	0.0053

Таблица 3: Выборочные коэффициенты корреляции двумерного нормального распределения. $\rho = 0.9(4)$

n=20	r	r_S	r_Q
$E(z)$	-0.3039	0.4771	0.1509
$E(z^2)$	0.5442	0.3042	0.2747
$D(z)$	0.4518	0.0765	0.2519
n=60	r	r_S	r_Q
$E(z)$	-0.6378	0.4742	0.3501
$E(z^2)$	0.487	0.2513	0.2017
$D(z)$	0.0802	0.0264	0.0791
n=100	r	r_S	r_Q
$E(z)$	-0.6888	0.4756	0.3935
$E(z^2)$	0.5051	0.2422	0.2105
$D(z)$	0.0307	0.016	0.0556

Таблица 4: Выборочные коэффициенты корреляции смешанного распределения. (1)

4.2 Эллипсы рассеивания

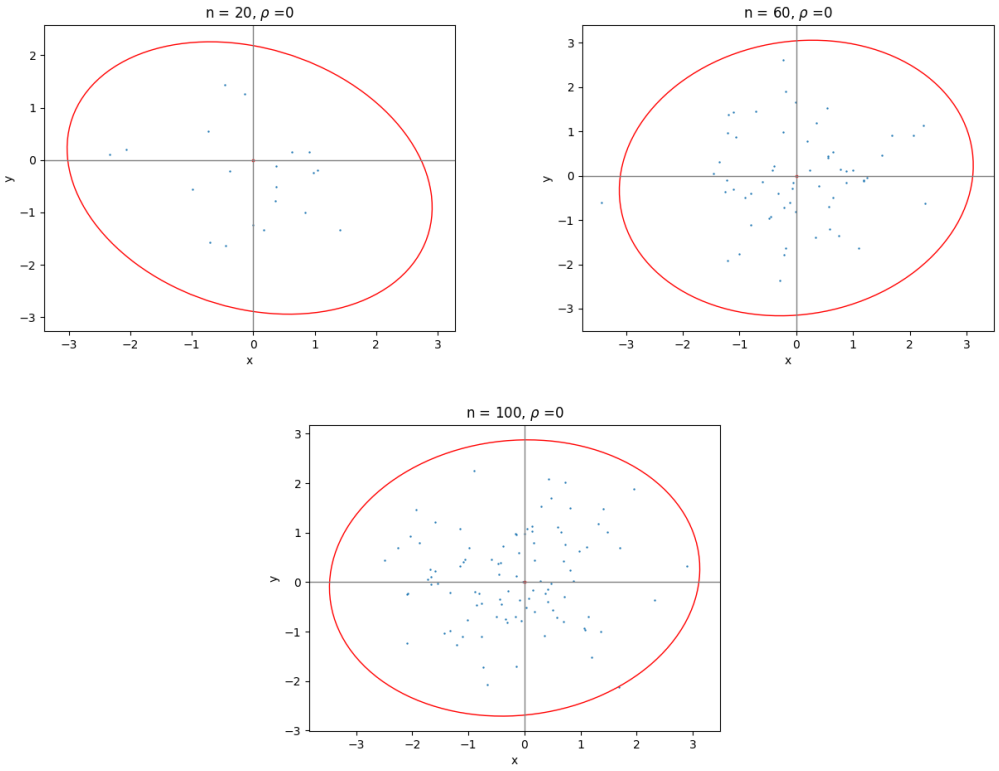

Рис. 1: Эллипсы рассеивания. $\rho = 0(4)$

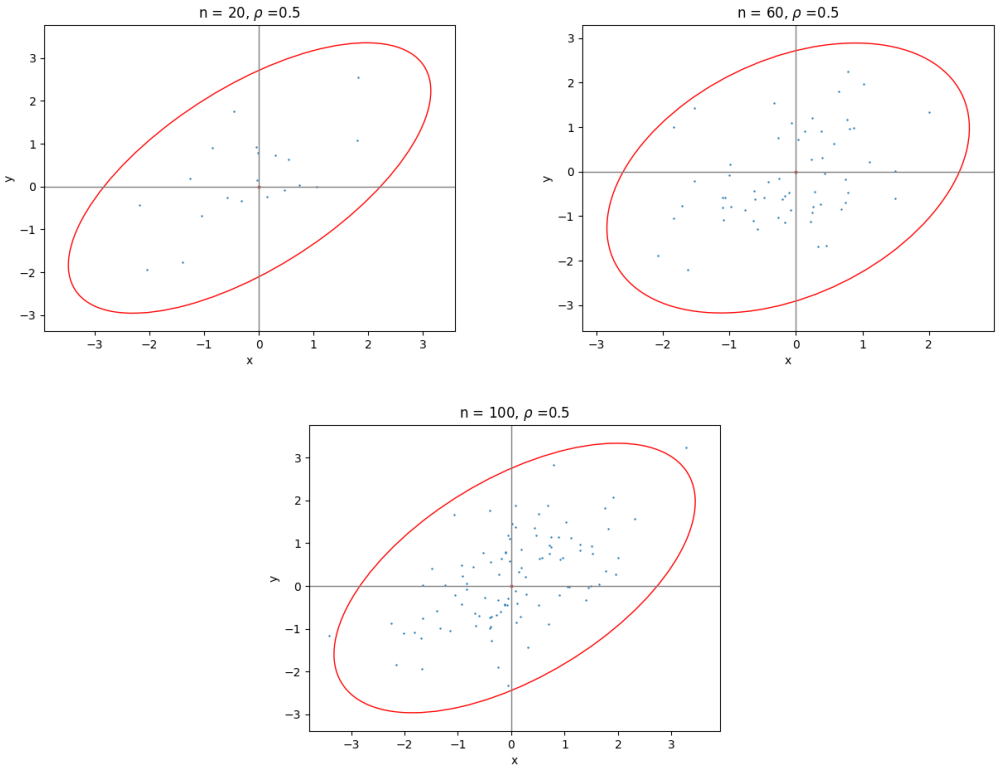

Рис. 2: Эллипсы рассеивания. $\rho = 0.5(4)$

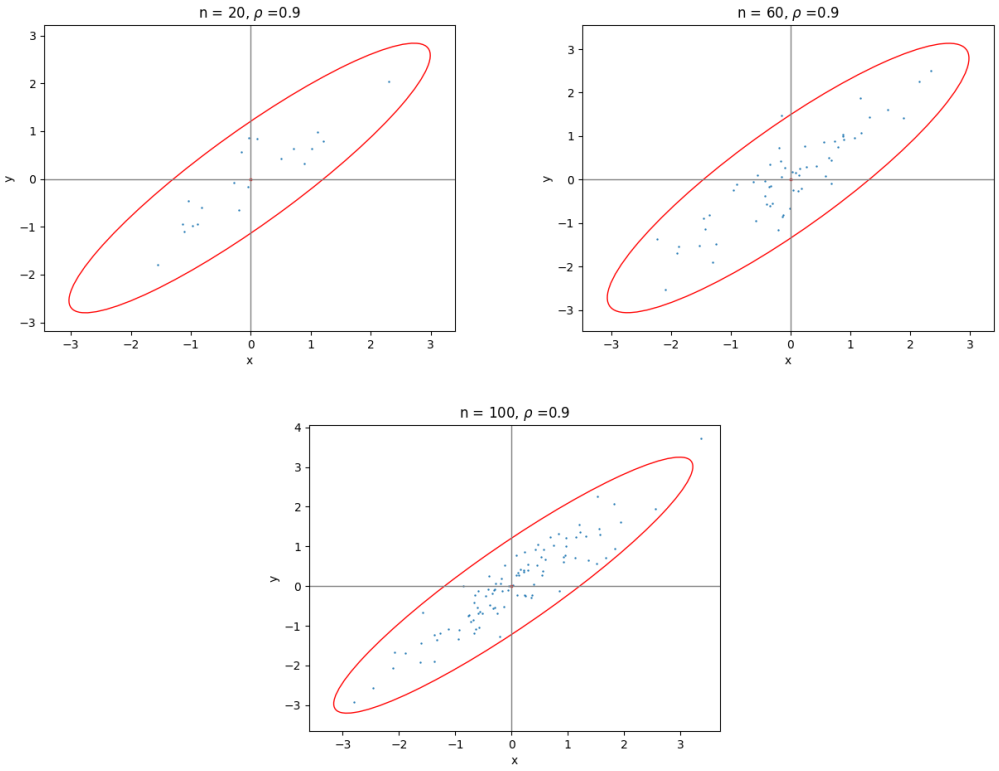

Рис. 3: Эллипсы рассеивания. $\rho = 0.9(4)$

4.3 Оценки коэффициентов линейной регрессии

4.3.1 Выборка без возмущений

- Критерий наименьших квадратов:

$$\hat{a} = 1.8986 \quad \hat{b} = 1.9533$$

- Критерий наименьших модулей:

$$\hat{a} = 1.9332 \quad \hat{b} = 1.6795$$

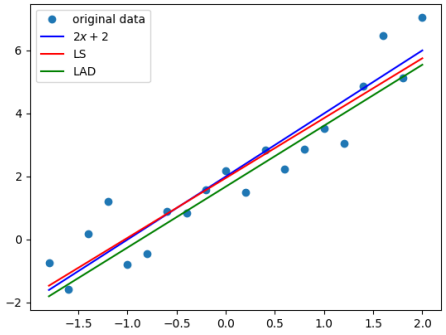

Рис. 4: Выборка без возмущений

4.3.2 Выборка с возмущениями

- Критерий наименьших квадратов:

$$\hat{a} = 0.3261 \quad \hat{b} = 2.2993$$

- Критерий наименьших модулей:

$$\hat{a} = 0.4383 \quad \hat{b} = 2.0391$$

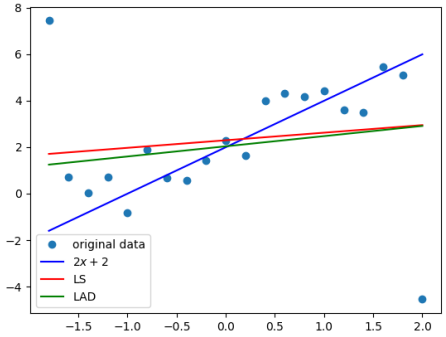

Рис. 5: Выборка с возмущениями

5 Обсуждение

5.1 Выборочные коэффициенты корреляции и эллипсы рассеивания

Для двумерного нормального распределения дисперсии выборочных коэффициентов корреляции упорядочены следующим образом: $D(r) < D(r_S) < D(r_Q)$. Аналогичные неравенства наблюдаются для величины $|\rho - r_{ind}|$, из чего можно сделать вывод, что оценка Пирсона коэффициента корреляции (7) является оптимальной в анализе двумерного нормального распределения.
Для смеси нормальных распределений наименьшая выборочная дисперсия наблюдается у коэффициента корреляции Спирмена. Кроме того коэффициент Спирмена наиболее устойчив к изменению размеров анализируемой выборки, из чего можно сделать вывод, что оценка Спирмена коэффициента корреляции (9) является оптимальной в анализе смеси нормальных распределений.
Процент попавших элементов выборки в эллипс рассеивания (95%-ная доверительная область) примерно равен его теоретическому значению.

5.2 Оценки коэффициентов линейной регрессии

Для сравнительно небольшой выборки (n=20) без возмущений критерий наименьших квадратов и критерий наименьших модулей дают сравнимые результаты (с небольшим выигрышем МНК в оценке коэффициента сдвига и небольшим выигрышем МНМ в оценке коэффициента наклона)
Для выборки с возмущениями МНК и МНМ также дают приблизительно одинаковые результаты. Стоит отметить, что в соответствии с построением эксперимента (резкие и равнозначные отклонения выборки на краях), хотя обоим методам сильно вредят отклонения в оценке коэффициента наклона, оценке коэффициента сдвига они вредят в меньшей степени.

Репозиторий

`https://github.com/KoloskovAleksandr/MathStatLabs2021`