Санкт-Петербургский Политехнический
Университет
им. Петра Великого

Институт прикладной математики и механики

Кафедра "Прикладная математика"

Отчёт по
Лабораторным работам № 5-8
по дисциплине
"Математическая статистика"

Выполнил студент:
Мальцов Дмитрий Дмитриевич
группа: 3630102/70401

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

Санкт-Петербург
2020 год

Содержание

1 Список иллюстраций

2 Список таблиц

3 Постановка задачи

3.1 Вычисление коэффициента корреляции

Необходимо построить выборки объёмом $20, 60, 100, 1000$ для двумерного нормального распределения с коэффициентами корреляции $\rho = 0, 0.5, 0.9$

Вычислить коэффициент корреляции Пирсона, Спирмана и квадрантный коэффициент корреляции для каждой выборки. Эти же вычисления повторить для смеси двумерных нормальных распределений:

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

На графике изобразить точки выборки и эллипс равновероятности.

3.2 Оценки линий регрессии

Необходимо найти оценки линейной регрессии $y_i = a + bx_i + e_i$, используя 20 точек отрезка $[-1.8;\ 2]$ с равномерным шагом 0.2. Ошибку e_i считать нормально распределённой с параметрами $(0,\ 1)$. В качестве эталонной зависимости взять $y_i = 2 + 2x_i + e_i$. При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей. Проделать то же самое для выборки, у которой в значении y_1 и y_{20} вносятся возмущения 10 и -10 соответственно.

3.3 Точечная оценка параметров распределения

Необходимо сгенерировать выборку объемом 100 элементов для нормального распределения $N(x; 0, 1)$. По сгенерированной выборке оценить параметры μ и σ нормального закона методом максимального правдоподобия. В качестве основной гипотезы H_0 будем считать, что сгенерированное распределение имеет вид $N(x, \hat{\mu}, \hat{\sigma})$. Проверить основную гипотезу, используя критерий согласия χ. В качестве ровня значимости взять $\alpha = 0,05$. Проверить гипотизу о нормальности исходного распределения для выборки из распределения Лапласа размером $n = 25$. Привести таблицы вычислений χ^2.

3.4 Интервальные оценки параметров распределения

Для двух выборок 20 и 100 элементов, сгенерированных согласно нормальному закону $N(x, 0, 1)$, для параметров масштаба и положения построить асимптотически нормальные интервальные оценки на основе точечных оценок метода максимального правдоподобия и классические интервальные оценки на основе статистик χ^2 и Стьюдента. В качестве параметра надёжности взять $\gamma = 0.95$.

4 Теория

4.1 Вычисление коэффициента корреляции

1. Двумерное нормально распределение:

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

2. Коэффициент корреляции Пирсона:

$$r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2 \frac{1}{n}\sum(y_i - \bar{y})^2}} \tag{3}$$

3. Квадрантный коэффициент корреляции:

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n} \tag{4}$$

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 $x^{'} = x - medx, y^{'} = y - medy$ и с центром в точке с координатами$(medx, medy)$

4. Коэффициент корреляции Спирмана:

$$r_S = \frac{\frac{1}{n}\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum(u_i - \bar{u})^2 \frac{1}{n}\sum(v_i - \bar{v})^2}} \tag{5}$$

где u и v – ранги, соотвествующие значениям переменной X и Y соответственно.

4.2 Оценки линий регрессии

Простая линейная регрессия :

$$y_i = ax_i + b + e_i,\ i = \overline{1, n}, \tag{6}$$

где x_i – заданные числа, y_i – наблюдаемые значения отклика, e_i – независимые, нормально распределённые с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые), a и b – неизвестные параметры, подлежащие оцениванию.

4.3 Метод наименьших квадратов

Критерий – минимизация функции :

$$Q(a, b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2 \to \min \tag{7}$$

Оценка $\hat{a}$ и $\hat{b}$ параметров a и b, в которых достигается минимум $Q(a, b)$, называются МНК-оценками. Формулы для их вычисления:

$$\begin{cases} \hat{a} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \\ \hat{b} = \overline{y} - \hat{a}\overline{x} \end{cases} \tag{8}$$

Оценка по методу наименьших квадратов является несмещённой оценкой.

МНК чувствителен к выбросам (т.к. в вычислении используется выборочное среднее, значение которого крайне неустойчиво к большим по относительной величине выбросам)

4.4 Метод наименьших модулей

Критерий наименьших модулей – заключается в минимизации следующей функции:

$$M(a,b) = \sum_{i=1}^{n} |y_i - ax_i - b| \to \min \tag{9}$$

Формулы для вычисления робастных параметров:

$$\begin{cases} \hat{a}_R = r_Q \frac{q_y^*}{q_x^*} \\ \hat{b}_R = medy - \hat{a}_R medx \end{cases} \tag{10}$$

, где

$$r_Q = \frac{1}{n} \sum sgn(x_i - medx) sgn(y_i - medy) \tag{11}$$

Статистики выборочной медианы и интерквартильной широты обладают робастными свойствами в силу того, что основаны на центральных порядковых статистиках, малочувствительных к большим по величине выбросам в данных. Статистика выборочного знакового коэффициента корреляции робастна, так как знаковая функция $sgnz$ чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка МНМ обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба.

4.5 Точечная оценка параметров распределения

4.6 Метод максимального правдоподобия

Метод максимального правдоподобия – метод оценивания неизвестного параметра путём максимимзации функции правдоподобия.

$$\hat{\theta}_{\text{МП}} = argmax \mathbf{L}(x_1, x_2, \ldots, x_n, \theta) \tag{12}$$

Где $\mathbf{L}$ это функция правдоподобия, которая представляет собой совместную плотность вероятности независимых случайных величин $X_1, x_2, \ldots, x_n$ и является функцией неизвестного параметра θ

$$\mathbf{L} = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \cdots \cdot f(x_n, \theta) \tag{13}$$

Оценкой максимального правдоподобия будем называть такое значение $\hat{\theta}_{\text{МП}}$ из множества допустимых значений параметра θ, для которого функция правдоподобия принимает максимальное значение при заданных $x_1, x_2, \ldots, x_n$.

Тогда при оценивании математического ожидания m и дисперсии σ^2 нормального распределения $N(m, \sigma)$ получим:

$$\ln(\mathbf{L}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln\left(\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 \tag{14}$$

Отсюда находятся выражения для оценок m и σ^2:

$$\begin{cases} m = \bar{x} \\ \sigma^2 = s^2 \end{cases} \tag{15}$$

4.7 Критерий согласия Пирсона

Разобьём генеральную совокупность на k неперсекающихся подмножеств $\Delta_1, \Delta_2, \ldots, \Delta_k$, $\Delta_i = (a_i, a_{i+1}]$, $p_i = P(X \in \Delta_i)$, $i = 1, 2, \ldots, k$ – вероятность того, что точка попала в iый промежуток.

Так как генеральная совокупность это $\mathbb{R}$, то крайние промежутки будут бесконечными: $\Delta_1 = (-\infty, a_1]$, $\Delta_k = (a_k, \infty)$, $p_i = F(a_i) - F(a_{i-1})$

n_i – частота попадания выборочных элементов в Δ_i, $i = 1, 2, \ldots, k$.

В случае справедливости гипотезы H_0 относительно частоты $\frac{n_i}{n}$ при больших n должны быть близки к p_i, значит в качестве меры имеет смысл взять:

$$Z = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2 \tag{16}$$

Тогда

$$\chi^2_B = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \tag{17}$$

Для выполнения гипотезы H_0 должны выполняться следующие условия:

$$\chi^2_B < \chi^2_{1-\alpha}(k-1) \tag{18}$$

где $\chi^2_{1-\alpha}(k-1)$ – квантиль распределения χ^2 с $k-1$ степенями свободы порядка $1-\alpha$, где α заданный уровень значимости.

4.8 Интервальные оценки параметров распределения

Доверительным интервалом или интервальной оценкой числовой характеристики или параметра распределения θ с доверительной вероятностью γ называется интервал со случайными границами (θ_1, θ_2), содержащий параметр θ с вероятностью γ.

Функция распределения Стьюдента:

$$T = \sqrt{n-1} \frac{\overline{x} - \mu}{\delta} \tag{19}$$

Функция плотности распределения χ^2:

$$f(x) = \begin{cases} 0, & x \le 0 \\ \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \end{cases} \tag{20}$$

Интервальные оценки для нормального распределения
математического ожидания:

$$P = \left(\overline{x} - \frac{\sigma t_{1-\frac{a}{2}}(n-1)}{\sqrt{n-1}} < \mu < \overline{x} + \frac{\sigma t_{1-\frac{a}{2}}(n-1)}{\sqrt{n-1}} \right) = \gamma, \tag{21}$$

где $t_{1-\frac{a}{2}}$ – квантиль распределения Стьюдента порядка $1 - \frac{a}{2}$.

стандартного отклонения:

$$P = \left(\frac{\sigma\sqrt{n}}{\sqrt{\chi^2_{1-\frac{a}{2}}(n-1)}} < \sigma < \frac{\sigma\sqrt{n}}{\sqrt{\chi^2_{\frac{a}{2}}(n-1)}} \right) = \gamma, \tag{22}$$

где $\chi^2_{1-\frac{a}{2}}$, $\chi^2_{\frac{a}{2}}$ – квантили распределения Стьюдента порядков $1 - \frac{a}{2}$ и $\frac{a}{2}$ соответственно.

Асимптотическая интервальная оценка для произвольного распределения при большой выборке
математического ожидания:

$$P = \left(\overline{x} - \frac{\sigma u_{1-\frac{\alpha}{2}}}{\sqrt{n}} < \mu < \overline{x} + \frac{\sigma u_{1-\frac{\alpha}{2}}}{\sqrt{n}} \right) = \gamma, \tag{23}$$

стандартного отклонения:

$$P = \left(s(1+U)^{-1/2} < \sigma < s(1-U)^{-1/2} \right) = \gamma, \tag{24}$$

где $u_{1-\frac{\alpha}{2}}$ – квантиль нормального распределения $N(x,0,1)$ порядка $1-\frac{\alpha}{2}$., $U = u_{1-\alpha/2}\sqrt{(e+2)/n}$, $e = m_4/s^4 - 3$

5 Реализация

Работы была выполнена на языке *Python*3.8.2 Для генерации выборок использовался модуль *numpy*. Для построения графиков использовалась библиотека *matplotlib*. Регрессионные модели использовались из библиотеки *statsmodels*.

6 Результаты

6.1 Вычисление коэффициента корреляции

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

$\rho = 0$	Pearson	Spearman	Quad
$E(z)$	0.009	0.001	0.004
$E(z^2)$	0.05	0.05	0.05
$D(z)$	0.05	0.05	0.05
$\rho = 0.5$	Pearson	Spearman	Quad
$E(z)$	0.49	0.46	0.32
$E(z^2)$	0.27	0.25	0.15
$D(z)$	0.03	0.03	0.05
$\rho = 0.9$	Pearson	Spearman	Quad
$E(z)$	0.893	0.865	0.69
$E(z^2)$	0.801	0.754	0.5
$D(z)$	0.003	0.05	0.03

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

$\rho = 0$	Pearson	Spearman	Quad
$E(z)$	−0.003	−0.004	−0.0004
$E(z^2)$	0.02	0.2	0.02
$D(z)$	0.02	0.02	0.02
$\rho = 0.5$	Pearson	Spearman	Quad
$E(z)$	0.497	0.47	0.33
$E(z^2)$	0.257	0.24	0.13
$D(z)$	0.009	0.01	0.02
$\rho = 0.9$	Pearson	Spearman	Quad
$E(z)$	0.8984	0.883	0.706
$E(z^2)$	0.8078	0.782	0.508
$D(z)$	0.0007	0.001	0.009

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

$\rho = 0$	Pearson	Spearman	Quad
$E(z)$	−0.004	−0.004	−0.001
$E(z^2)$	0.01	0.01	0.01
$D(z)$	0.01	0.01	0.01
$\rho = 0.5$	Pearson	Spearman	Quad
$E(z)$	0.499	0.481	0.329
$E(z^2)$	0.254	0.237	0.118
$D(z)$	0.005	0.006	0.009
$\rho = 0.9$	Pearson	Spearman	Quad
$E(z)$	0.8981	0.8841	0.708
$E(z^2)$	0.8071	0.7823	0.506
$D(z)$	0.0004	0.0007	0.004

Таблица 4: Смесь нормальных распределений

$n = 20$	Pearson	Spearman	Quad
$E(z)$	−0.09	−0.09	−0.06
$E(z^2)$	0.05	0.06	0.06
$D(z)$	0.05	0.05	0.05
$n = 60$	Pearson	Spearman	Quad
$E(z)$	−0.09	−0.004	−0.06
$E(z^2)$	0.03	0.02	0.02
$D(z)$	0.02	0.02	0.02
$n = 100$	Pearson	Spearman	Quad
$E(z)$	−0.094	−0.079	−0.06
$E(z^2)$	0.018	0.016	0.01
$D(z)$	0.009	0.009	0.01

Рис. 1: Графики двумерного нормального распределения и смеси для размера выборки $n = 20$

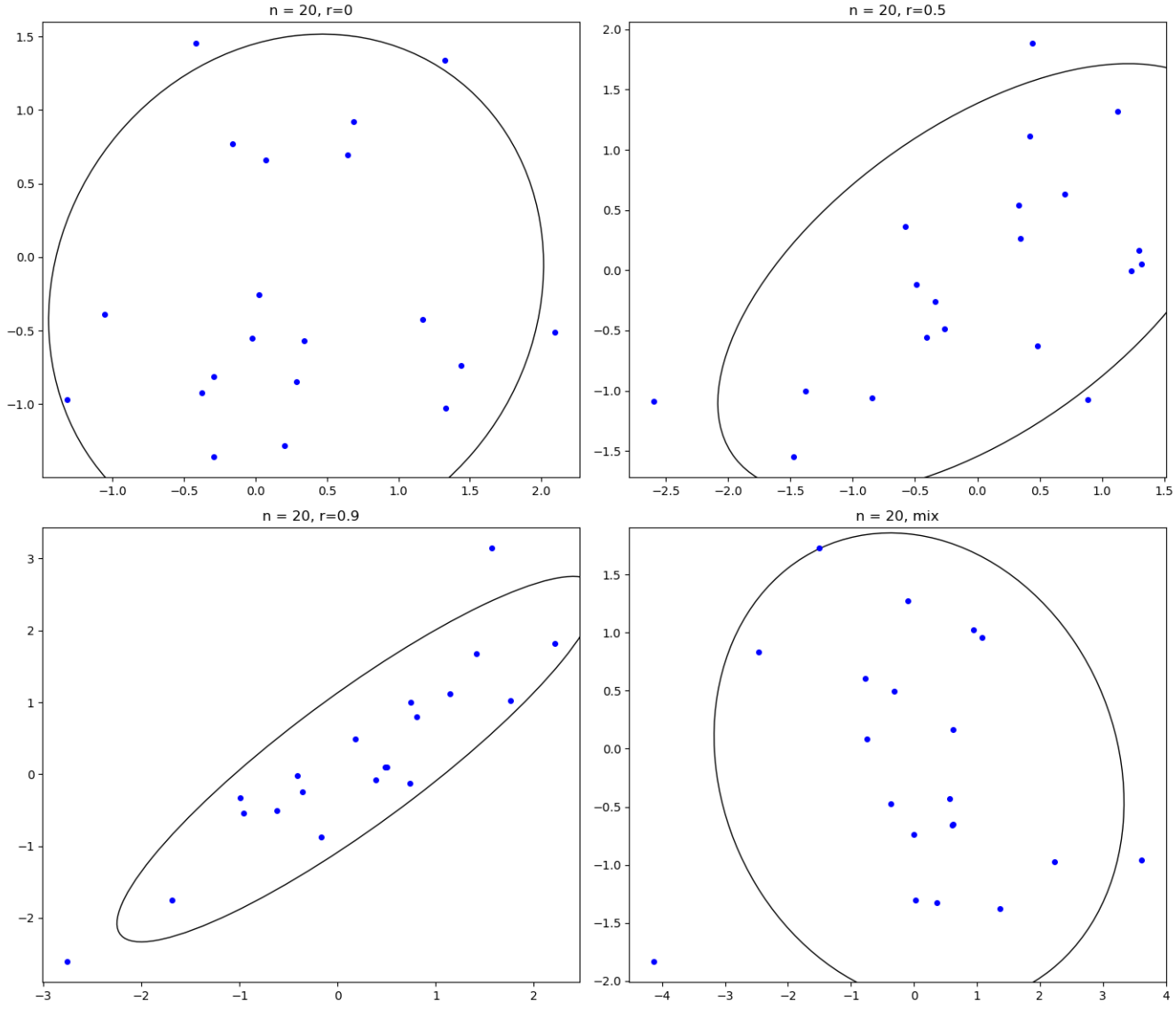


Рис. 2: Графики двумерного нормального распределения и смеси для размера выборки $n = 60$

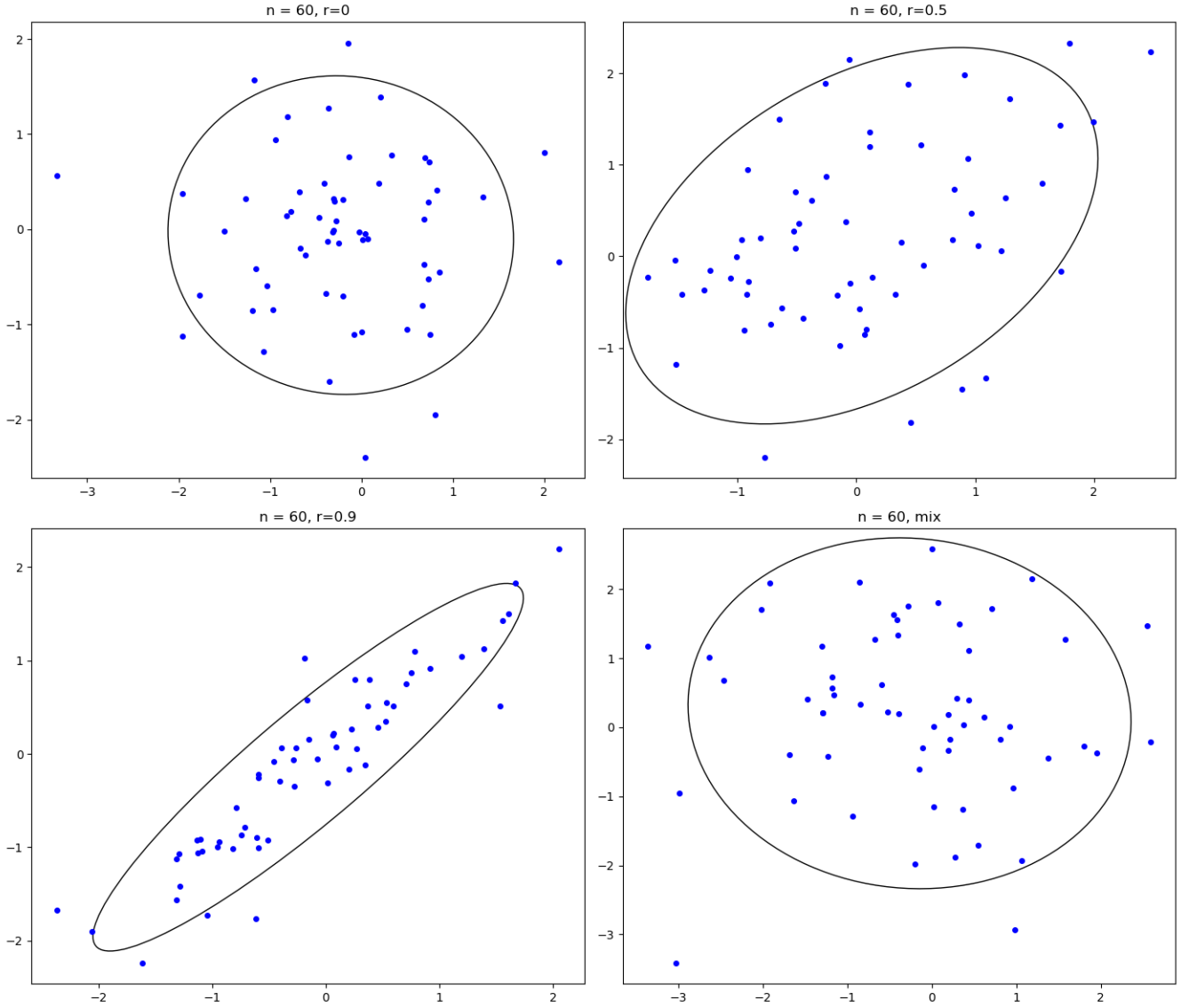


Рис. 3: Графики двумерного нормального распределения и смеси для размера выборки $n = 100$

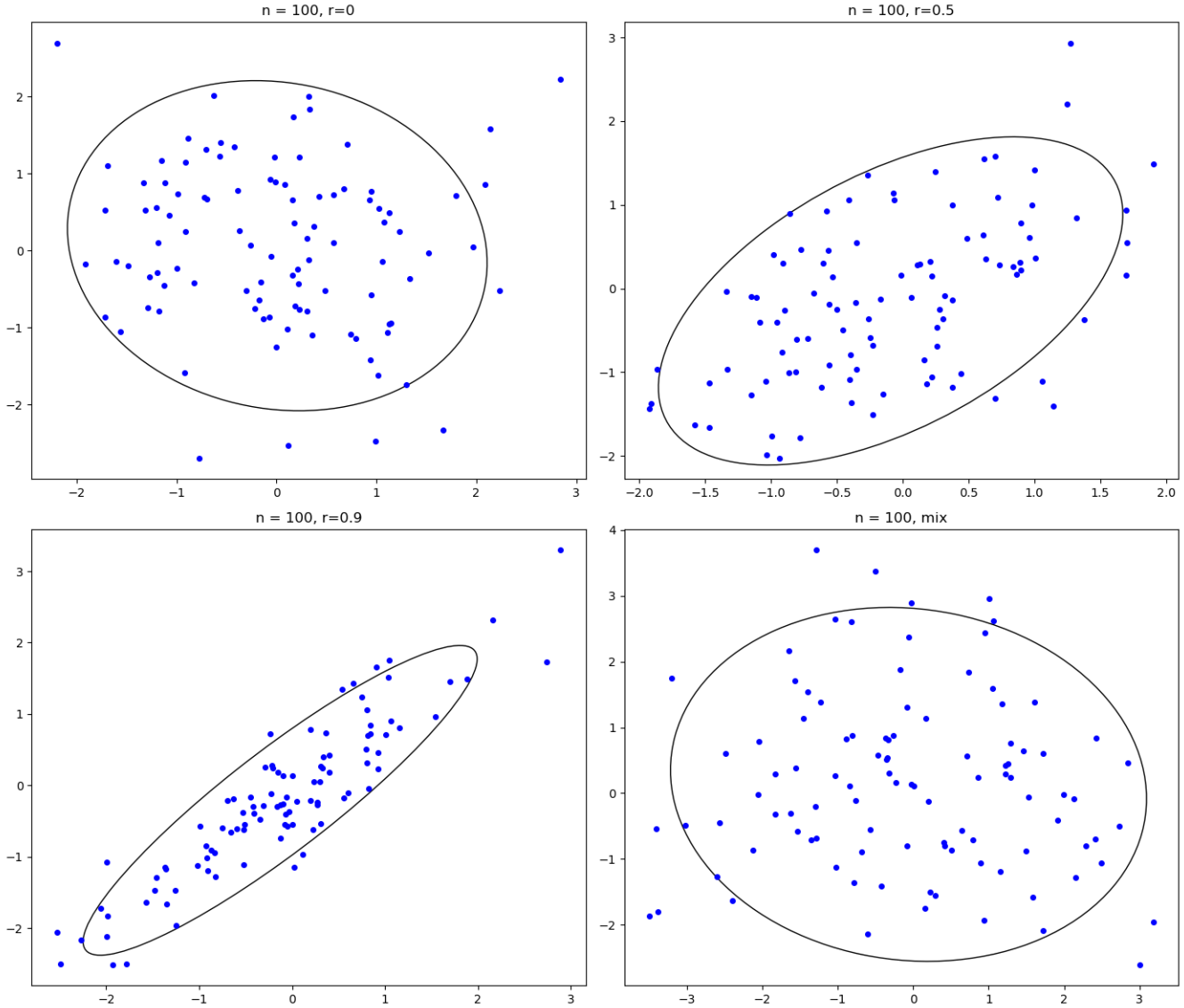


Рис. 4: Графики эллипса рассеивания для двумерного нормального распределения для 2 точек

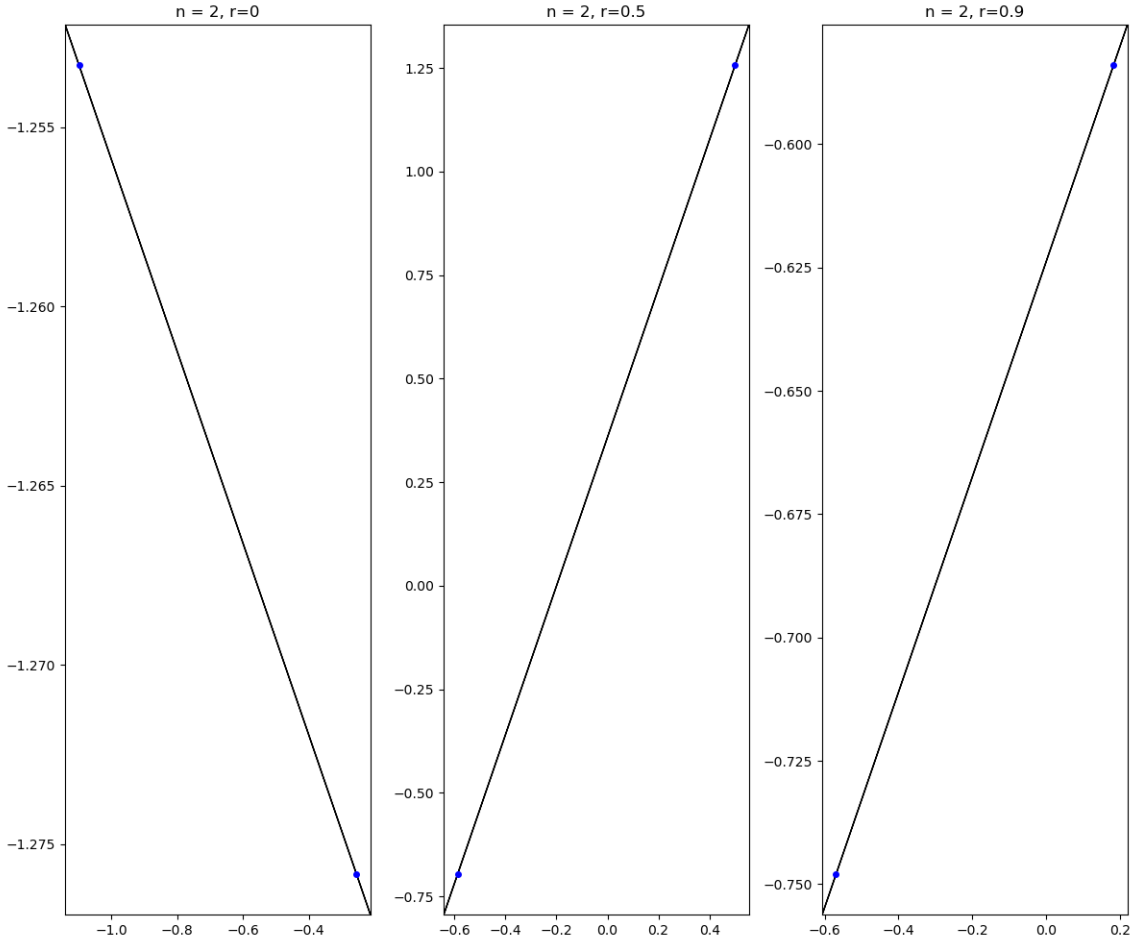
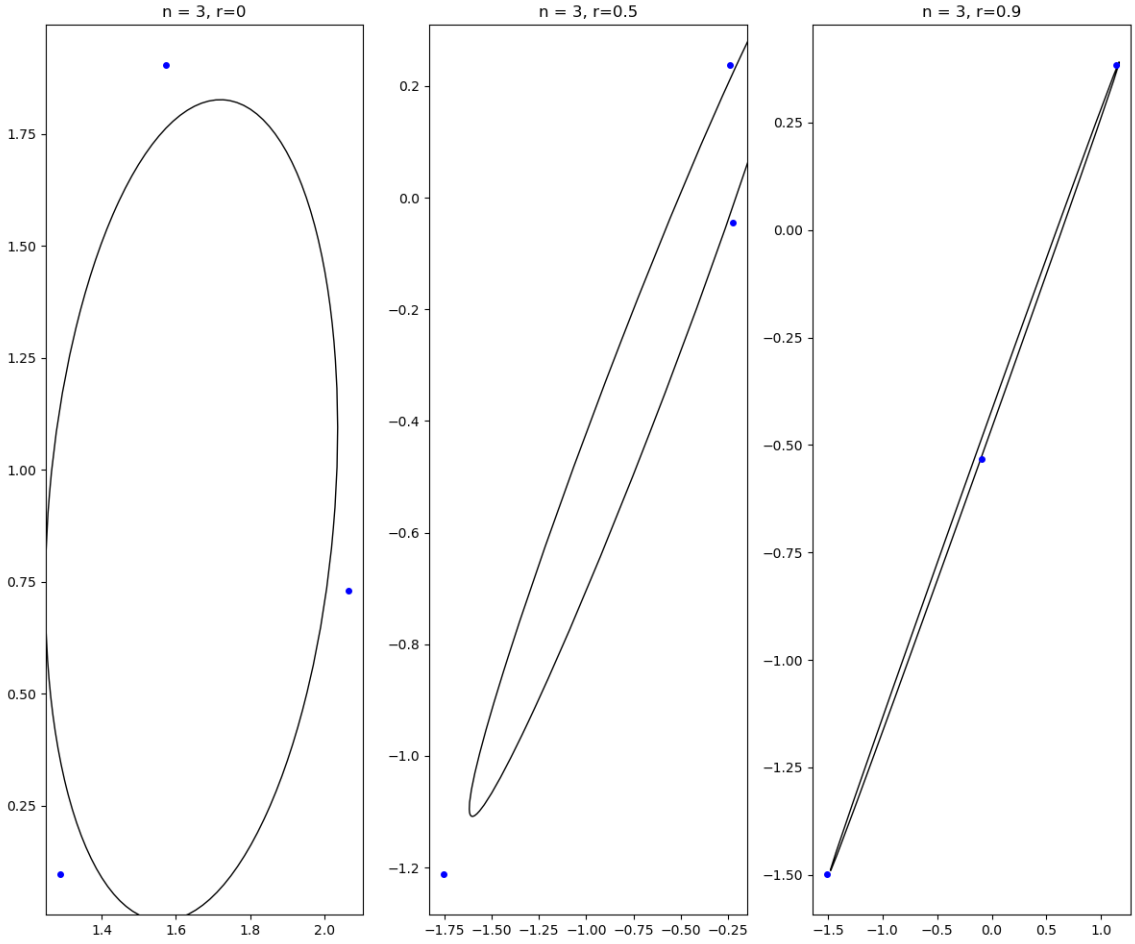


Рис. 5: Графики эллипса рассеивания для двумерного нормального распределения для 3 точек



6.2 Оценки линий регрессии

Рис. 6: Графики линейной регрессии

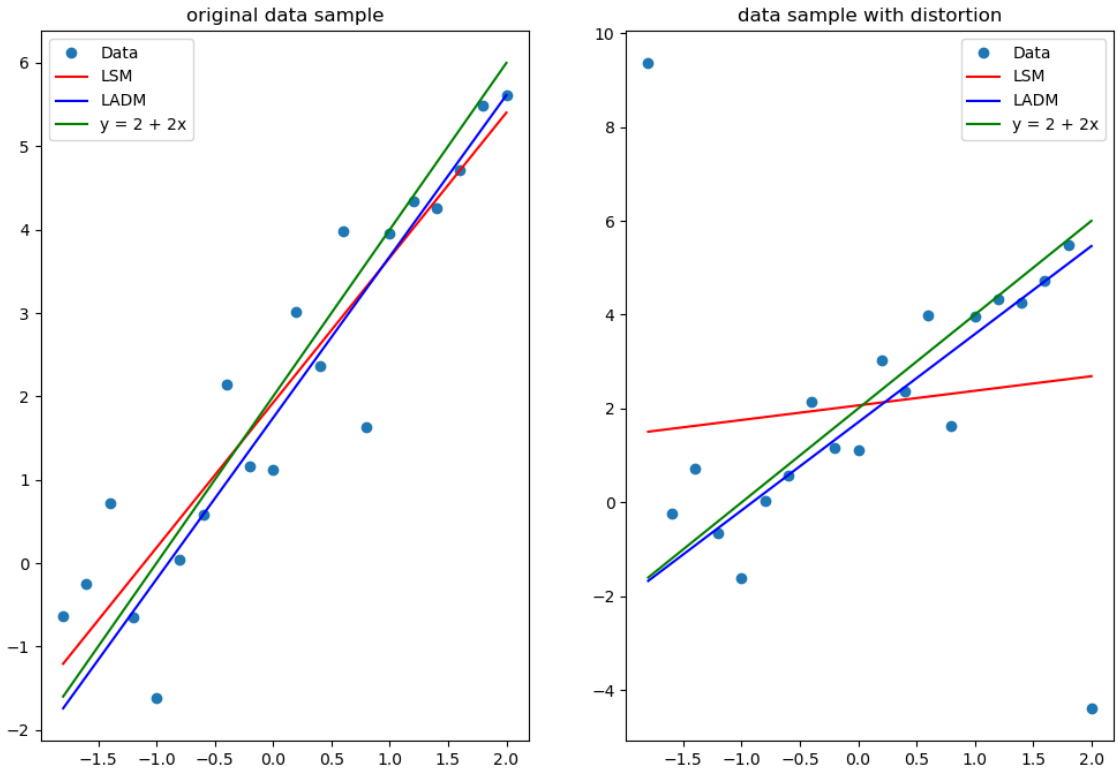


Таблица 5: Таблица оценок коэффициентов линейной регрессии без возмущёний

	$\hat{a}$	$\hat{b}$
МНК	1.739737	1.924716
МНМ	1.935446	1.742526

Таблица 6: Таблица оценок коэффициентов линейной регрессии с возмущёниями

	$\hat{a}$	$\hat{b}$
МНК	0.311165	2.067573
МНМ	1.877536	1.707785

6.3 Точечная оценка параметров распределения

6.4 Метод максимального правдоподобия

При подсчете оценок параметров закона нормального распределения методом максимального правдоподобия были получены следующие значения:

$$\begin{aligned} \hat{m}_{\text{МП}} &= 0.090527 \\ \hat{\sigma}^2_{\text{МП}} &= 0.963167 \end{aligned} \tag{25}$$

6.5 Критерий Пирсона

Таблица 7: Таблица вычислений χ^2

i	Δ_i	n_i	p_i	$\frac{(n_i - np_i)^2}{np_i}$
1	$(-\infty, -1.0]$	15	0.1288	0.3501
2	$(-1.0, -0.5)$	10	0.1411	1.1988
3	$(-0.5, 0.0)$	24	0.1927	1.1634
4	$(0.0, 0.5)$	19	0.2021	0.0721
5	$(0.5, 1.0)$	13	0.1629	0.6626
6	$(1.0, \infty)$	19	0.1725	0.1771
$\sum$		100	1	3.6241

$$\chi^2_B = 3.6241$$

6.6 Проверка гипотезы о нормальности для распределения Лапласа

Размер выборки $n = 25$ для распределения Лапласа

$$L\left(x, 0, \frac{1}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \tag{26}$$

$$\begin{aligned} \hat{m}_{\text{МП}} &= 0.198045 \\ \hat{\sigma}^2_{\text{МП}} &= 0.656187 \end{aligned} \tag{27}$$

Таблица 8: Таблица вычислений χ^2

i	Δ_i	n_i	p_i	$\frac{(n_i - np_i)^2}{np_i}$
1	$(-\infty, -1.0]$	1	0.0339	0.027
2	$(-1.0, 0.0)$	12	0.3475	1.2641
3	$(0.0, 1.0)$	8	0.5078	1.7359
4	$(1.0, \infty)$	4	0.1108	0.5454

$$\chi^2_B = 3.5725$$

6.7 Интервальные оценки параметров распределения

Таблица 9: Доверительные интервалы для параметров нормального распределения

	m	σ
$n = 20$	$[-0.7367, -0.0098]$	$[0.5906, 1.1343]$
$n = 100$	$[-0.0307, 0.356]$	$[0.8555, 1.1319]$

Таблица 10: Доверительные интервалы для параметров произвольного распределения. Асимптотический подход

	m	σ
$n = 20$	$[-0.705, -0.0415]$	$[0.6199, 1.0608]$
$n = 100$	$[-0.0273, 0.3527]$	$[0.8668, 1.1201]$

7 Выводы

7.1 Вычисление коэффициента корреляции

По таблицам 1,2,3,4 , видно, что, при увеличении объёма выборки, подсчитанные коэффициенты корреляции стремятся к теоретическим.

Ближе всего к теоретическому коэффициенту корреляции находится коэффициент Пирсона.

По графикам также видно, что при уменьшении корреляции эллипс равновероятности стремится к окружности, а при увеличении растягивается, стремясь к прямой.

Из графиков наглядно видно, что для построения эллипса рассеивания необходимое минимальное число событий в выборке – 3 события, так как 2 точки (2 события) вырождаются в прямую линию (для 2 точек мы всегда можем перейти в систему координат, где у одной из компонент вектора (x,y) будет 0 мат. ожидание и 0 дисперсия, то есть переходим в одномерный случай).

7.2 Оценки линий регрессии

По графику 6 видно, что оба метода дают хорошую оценку коэффициентов линейной регрессии, если нет выбросов. Однако выбросы сильно влияют на оценки по МНК.

Выбросы мало влияют на оценку по МНМ. Ценой за это является бóльшая по сравнению с МНК сложность вычисления.На практике зачастую легче просто отсеять выбросы из выборки.

7.3 Точечная оценка параметров распределения

Табличное значение квантиля $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(5) = 11.0705$. Полученное значение критерия согласия Пирсона для нормального распределения $\chi^2_B = 3.6241 < \chi^2_{0.95}(5)$, следовательно основная гипотеза H_0 на исходной выборке не может быть отвергнута на уровне значимости $\alpha = 0.05$.. Для распределения Лапласа полученное значение критерия Пирсона $\chi^2_B = 3.5725 < \chi^2_{0.95}(3) = 7.8147$ означает что из полученной выборки мы не можем отвергнуть гипотезу H_0 о нормальности исходного распределения. Такой результат легко объясним низким размером выборки, так как интервалы в которых мы оцениваем распределение получаются слишком большими, на которых распределение Лапласа очень схоже с нормальным.

7.4 Интервальные оценки параметров распределения

Качество оценок растёт с увеличением объёма выборки, оба метода показывают схожие точности оценки, но у ассимптотического подхода очевидно преимущество в применимости к выборке из произвольного распределения.

8 Литература

Модуль numpy
matplotlib boxplot
Боксплот Тьюки
Модуль scipy

Модуль matplotlib

9 Приложения

Код 5-й лабораторной
Код 6-й лабораторной
Код 7-й лабораторной
Код 8-й лабораторной