Санкт-Петербургский Политехнический Университет
им. Петра Великого

Институт прикладной математики и механики

Кафедра прикладной математики

# Отчёт по лабораторным работам №5-8 по дисциплине “Математическая статистика”

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Баженов Александр Николаевич

Санкт-Петербург

2020 г.

# Оглавление

# Список иллюстраций и таблиц

# 1 Постановка задачи

## 1.1 Выборочные коэффициенты корреляции и эллипсы рассеивания

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 рассеяния.

## 1.2 Простая линейная регрессия

Найти оценки коэффициентов линейной регрессии $y_i = ax_i + b + e_i$, используя 20 точек на отрезке $[-1.8;\ 2]$ с равномерным шагом равным 0.2. Ошибку $e_i$ считать нормально распределённой с параметрами $(0;\ 1)$. В качестве эталонной зависимости взять $y_i = 2x_i + 2 + e_i$. При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей.

Проделать то же самое для выборки, у которой в значения $y_1$ и $y_{20}$ вносятся возмущения 10 и -10.

## 1.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Сгенерировать выборку объёмом 100 элементов для стандартного нормального распределения $N(x, 0{,}1)$. По сгенерированной выборке оценить параметры $\mu$ и $\sigma$ 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 $\chi^2$. В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений $\chi^2$.

## 1.4 Интервальные оценки математического ожидания и стандартного отклонения

Для двух выборок из 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 $\chi^2$ и Стьюдента. В качестве параметра надёжности взять $\gamma = 0.95$.

# 2 Теория

## 2.1 Выборочные коэффициенты корреляции и эллипсы рассеивания

1. Двумерное стандартное нормальное распределение:

$$N(x, y, 0{,}0, 1, 1, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)}$$

2. Коэффициент корреляции Пирсона:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

3. Коэффициент корреляции Спирмена:

$$\rho_n = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n} d_i^2$$

4. Квадрантный коэффициент корреляции:

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} sign(x_i - medx)\, sign(y_i - medy)$$

## 2.2 Простая линейная регрессия

Стандартное нормальное распределение:

$$N(x, 0{,}1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Точная формула:

$$y_i = ax_i + b + e_i \; (i = \overline{1, n}),$$

где $x_i$ – заданные числа, $y_i$ – наблюдаемые значения, $e_i$ – независимые, одинаково распределённые значения ошибок с параметрами $E[e_i] = 0$ и $D[e_i] = \sigma^2$, $a$ и $b$ – неизвестные параметры, подлежащие оценке.

### 2.2.1 Метод наименьших квадратов (МНК)

Критерий – минимизация RSS-функции (***R**esidual **S**um of **S**quares*):

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(\hat{a}x_i + \hat{b} - y_i\right)^2 \to min$$

В случае линейной регрессии оценочные коэффициенты $\hat{a}$ и $\hat{b}$ можно вычислить по формулам:

$$\begin{cases} \hat{a} = \dfrac{\sigma_y}{\sigma_x} r_{xy} \\ \hat{b} = \overline{y} - \hat{a}\overline{x} \end{cases}$$

МНК является несмещённой оценкой. Чувствителен к выбросам (т. к. в вычислениях используется выборочное среднее, крайне неустойчивое к редким, но больших по величине выбросам).

### 2.2.2 Метод наименьших модулей (МНМ)

Критерий – минимизация LAD-функции (***L**east **A**bsolute **D**eviations*):

$$LAD = \sum_{i=1}^{n} \left| \hat{a}x_i + \hat{b} - y_i \right| \to min$$

Коэффициенты так же можно вычислить по формулам:

$$\begin{cases} \hat{a} = \dfrac{\widehat{y_2} - \widehat{y_1}}{\widehat{x_2} - \widehat{x_1}} \\ \hat{b} = -med(\hat{a}x - \hat{y}) \end{cases}$$

МНМ-оценки обладают свойством робастности. Но на практике решение реализуется только численно.

## 2.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Нормальное распределение:

$$N(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

### 2.3.1 Метод максимального правдоподобия (ММП)

МНМ – метод оценивания неизвестного параметра $\theta$ путём максимизации функции правдоподобия $L(X, \theta)$:

$$\hat{\theta}_{\text{ОМП}} = argmax[L(X, \theta)]$$

$$L(X, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Оценкой максимального правдоподобия будем называть такое значение $\hat{\theta}_{\text{ОМП}}$ из множества допустимых значений $\theta$, для которого $L(X, \theta)$ принимает максимальное значение для заданных $x_1, \dots, x_n$.

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

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Отсюда находятся выражения для оценок $\mu$ и $\sigma^2$:

### 2.3.2 Критерий Пирсона

Разобьём генеральную совокупность на $k$ непересекающихся подмножеств $\Delta_1, \dots, \Delta_k$, где $\Delta_i = (x_i, x_{i+1}], p_i = P(X \in \Delta_i), i = \overline{1,k}$ – вероятность того, что точка попала в $i$-ый промежуток.

Так как генеральная совокупность это $\mathbb{R}$, то крайние промежутки будут бесконечными: $\Delta_1 = (-\infty, x_1], \Delta_k = (x_k, \infty], p_i = F(x_i) - F(x_{i-1})$

Пусть $n_i$ – частота попадания выборочных элементов в $\Delta_i$.

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:

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

Тогда

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

Для выполнения гипотезы $H_0$ должны выполняться следующие условия:

$$\chi_B^2 < \chi_{1-\alpha}^2 (k-1),$$

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

## 2.4 Интервальные оценки математического ожидания и стандартного отклонения

Стандартное нормальное распределение:

$$N(x, 0,1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Функция распределения Стьюдента:

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

Функция плотности $\chi^2$:

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x > 0 \\ 0, \text{иначе} \end{cases}$$

Доверительным интервалом или интервальной оценкой числовой характеристики или параметра распределения $\theta$ с доверительной вероятностью $\gamma$ называется интервал со случайными границами $(\theta_1, \theta_2)$, содержащий $\theta$ с вероятностью $\gamma$.

### 2.4.1 Интервальные оценки

Интервальные оценки для математического ожидания нормального распределения:

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

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

Интервальные оценки для стандартного отклонения нормального распределения:

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

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

### 2.4.2 Асимптотические оценки

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

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

Асимптотическая интервальная оценка для произвольного распределения при большой выборке стандартного отклонения:

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

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

## 3 Реализация

Был использован язык *Python 3.8.2*: модуль *numpy* для вычисления описательных статистик, модуль *scipy* для генерации выборок и расчёта

коэффициентов, модуль *matplotlib* для построения и отображения графиков, модуль *pandas* для хранения статистических данных в таблицах и функция *display* из модуля *IPython.display* для их корректного отображения.

# 4 Результаты

## 4.1 Выборочные коэффициенты корреляции и эллипсы рассеивания

### 4.1.1 Таблицы

*Таблица 1 Двумерное стандартное нормальное распределение, n=20, r=0*

| n=20 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.012 | 0.011 | 0.003 |
| **E(z^2)** | 0.048 | 0.05 | 0.051 |
| **D(z)** | 0.048 | 0.049 | 0.051 |

*Таблица 2 Двумерное стандартное нормальное распределение, n=60, r=0*

| n=60 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.004 | 0.004 | 0.003 |
| **E(z^2)** | 0.016 | 0.017 | 0.017 |
| **D(z)** | 0.016 | 0.017 | 0.017 |

*Таблица 3 Двумерное стандартное нормальное распределение, n=100, r=0*

| n=100 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.001 | 0.001 | -0.001 |
| **E(z^2)** | 0.01 | 0.01 | 0.011 |
| **D(z)** | 0.01 | 0.01 | 0.011 |

*Таблица 4 Двумерное стандартное нормальное распределение, n=20, r=0.5*

| n=20 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.489 | 0.46 | 0.322 |
| **E(z^2)** | 0.271 | 0.246 | 0.148 |
| **D(z)** | 0.032 | 0.035 | 0.044 |

*Таблица 5 Двумерное стандартное нормальное распределение, n=60, r=0.5*

| n=60 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.496 | 0.476 | 0.331 |
| **E(z^2)** | 0.255 | 0.237 | 0.124 |
| **D(z)** | 0.009 | 0.01 | 0.014 |

*Таблица 6 Двумерное стандартное нормальное распределение, n=100, r=0.5*

| n=100 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.497 | 0.477 | 0.331 |
| **E(z^2)** | 0.253 | 0.233 | 0.118 |
| **D(z)** | 0.005 | 0.006 | 0.009 |

*Таблица 7 Двумерное стандартное нормальное распределение, n=20, r=0.9*

| n=20 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.896 | 0.867 | 0.696 |
| **E(z^2)** | 0.806 | 0.756 | 0.513 |
| **D(z)** | 0.002 | 0.004 | 0.029 |

*Таблица 8 Двумерное стандартное нормальное распределение, n=60, r=0.9*

| n=60 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.898 | 0.883 | 0.707 |
| **E(z^2)** | 0.808 | 0.78 | 0.508 |
| **D(z)** | 0.001 | 0.001 | 0.009 |

*Таблица 9 Двумерное стандартное нормальное распределение, n=100, r=0.9*

| n=100 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | 0.899 | 0.886 | 0.708 |
| **E(z^2)** | 0.809 | 0.786 | 0.507 |
| **D(z)** | 0 | 0.001 | 0.005 |

*Таблица 10 Смесь распределений, n=20*

| n=20 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | -0.08 | -0.078 | -0.05 |
| **E(z^2)** | 0.061 | 0.061 | 0.056 |
| **D(z)** | 0.054 | 0.055 | 0.054 |

*Таблица 11 Смесь распределений, n=60*

| n=60 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | -0.092 | -0.086 | -0.06 |
| **E(z^2)** | 0.025 | 0.024 | 0.021 |
| **D(z)** | 0.016 | 0.016 | 0.017 |

*Таблица 12 Смесь распределений, n=100*

| n=100 | Pearson | Spearman | quadrant |
|---|---|---|---|
| **E(z)** | -0.097 | -0.092 | -0.063 |
| **E(z^2)** | 0.019 | 0.019 | 0.014 |
| **D(z)** | 0.01 | 0.01 | 0.01 |

### 4.1.2 Иллюстрации

*Рис. 1 Двумерное стандартное нормальное распределение для n=20, r=0*

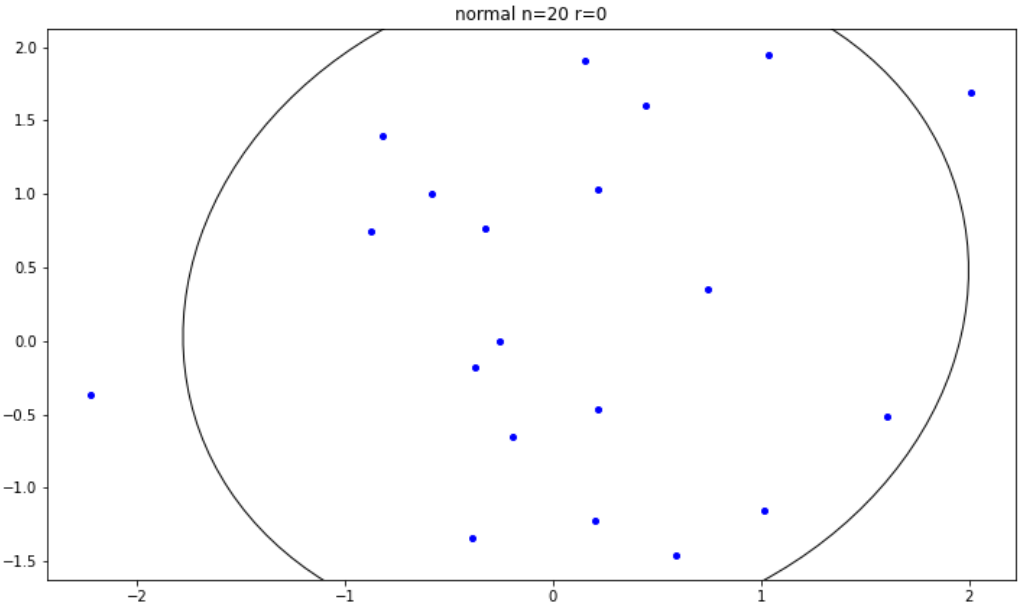



*Рис. 2 Двумерное стандартное нормальное распределение для n=60, r=0*

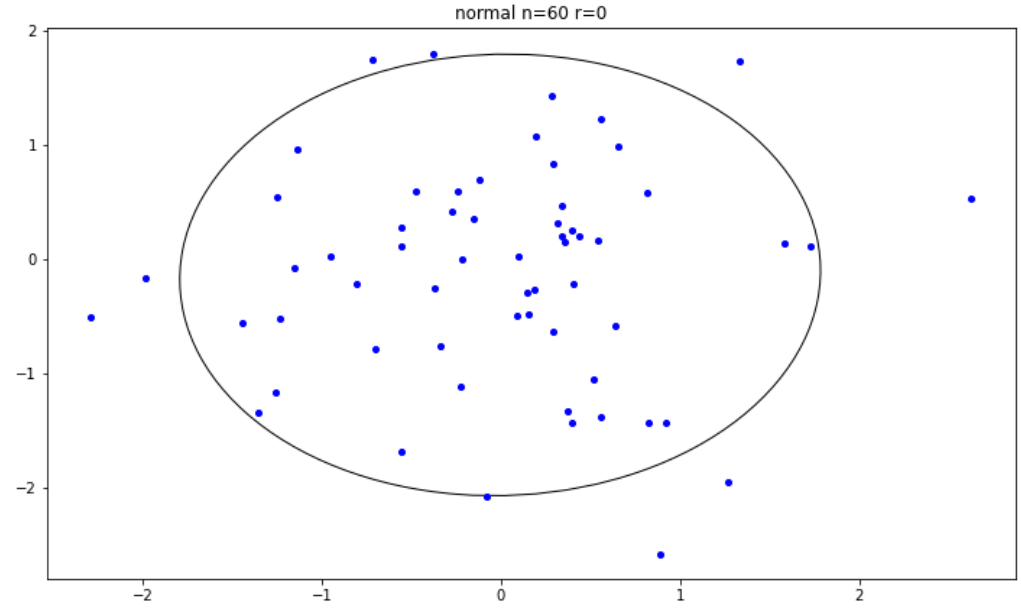


*Рис. 3 Двумерное стандартное нормальное распределение для n=100, r=0*

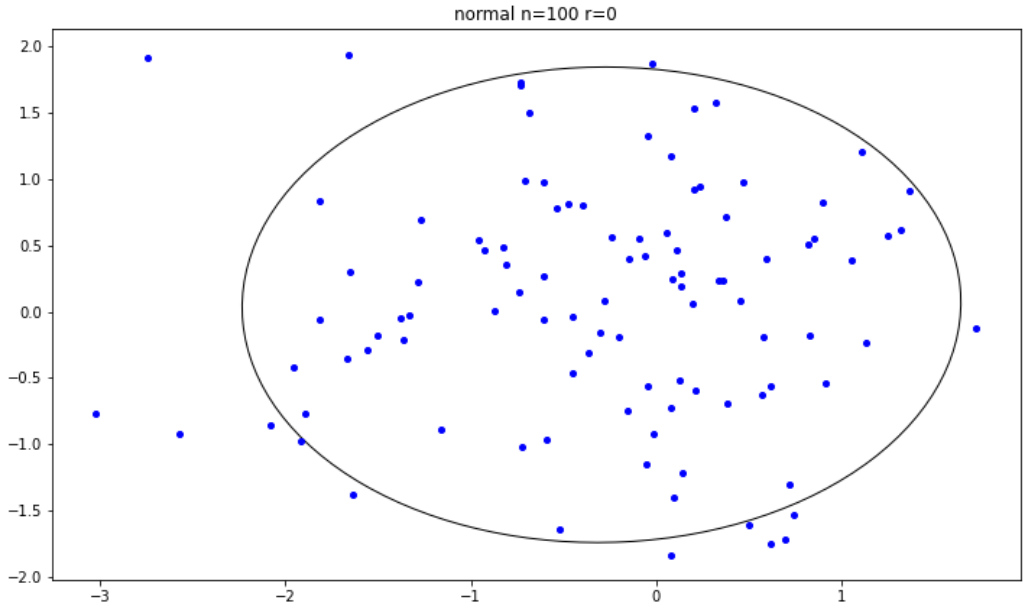



*Рис. 4 Двумерное стандартное нормальное распределение для n=20, r=0.5*

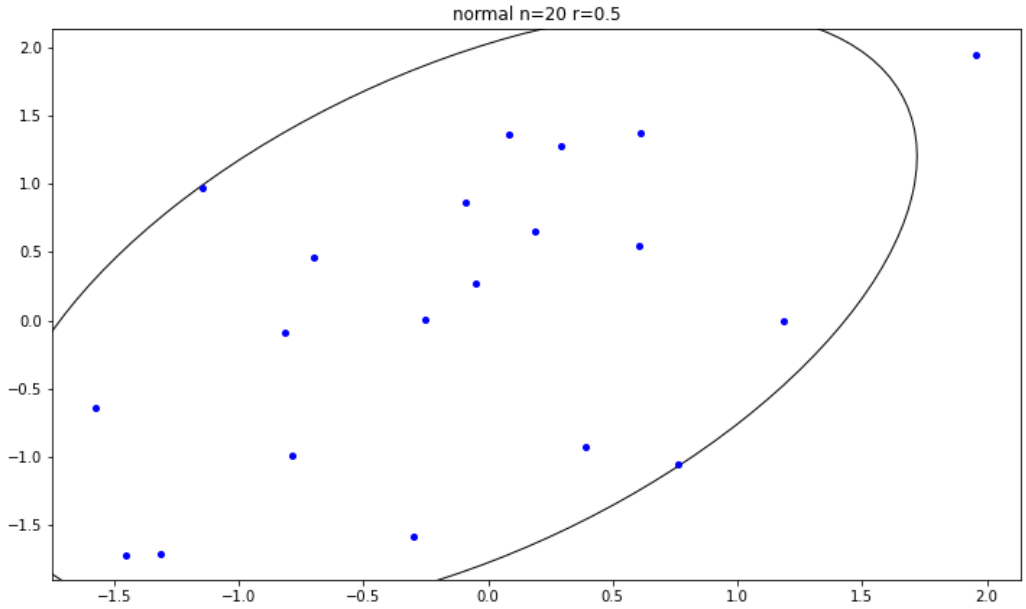


*Рис. 5 Двумерное стандартное нормальное распределение для n=60, r=0.5*

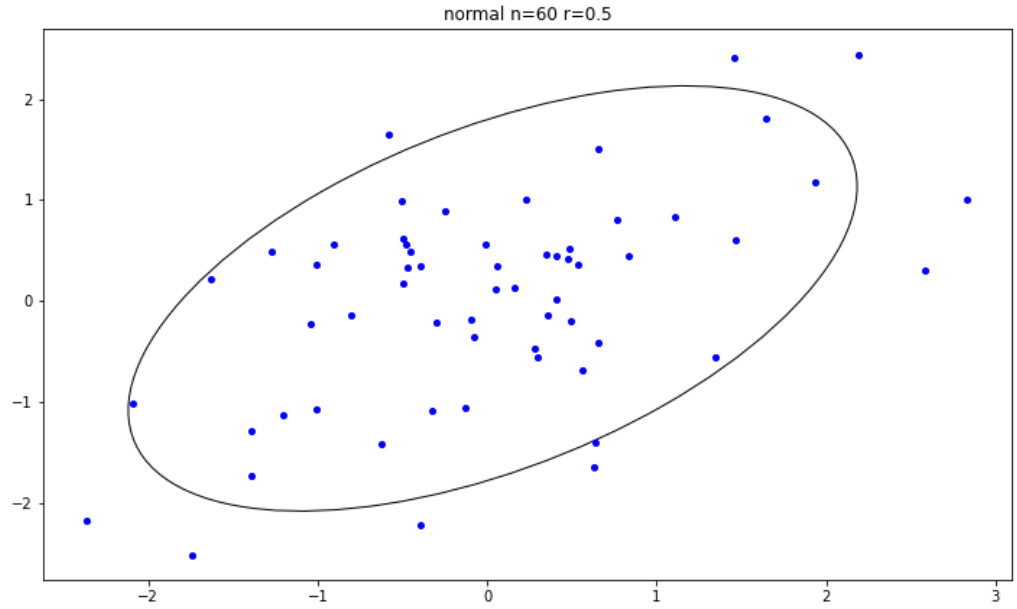



*Рис. 6 Двумерное стандартное нормальное распределение для n=100, r=0.5*

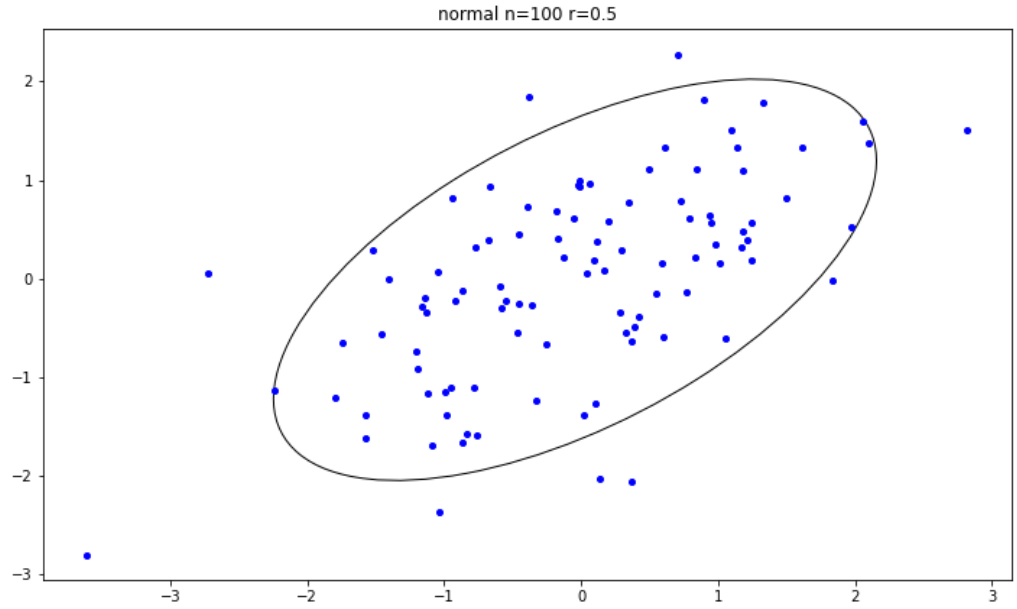


*Рис. 7 Двумерное стандартное нормальное распределение для n=20, r=0.9*

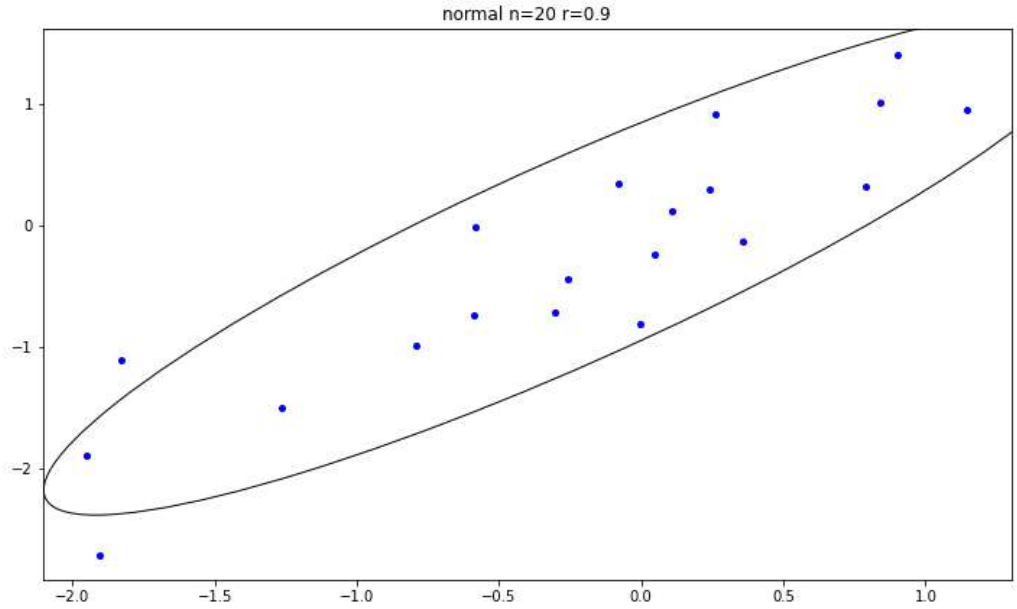



*Рис. 8 Двумерное стандартное нормальное распределение для n=60, r=0.9*

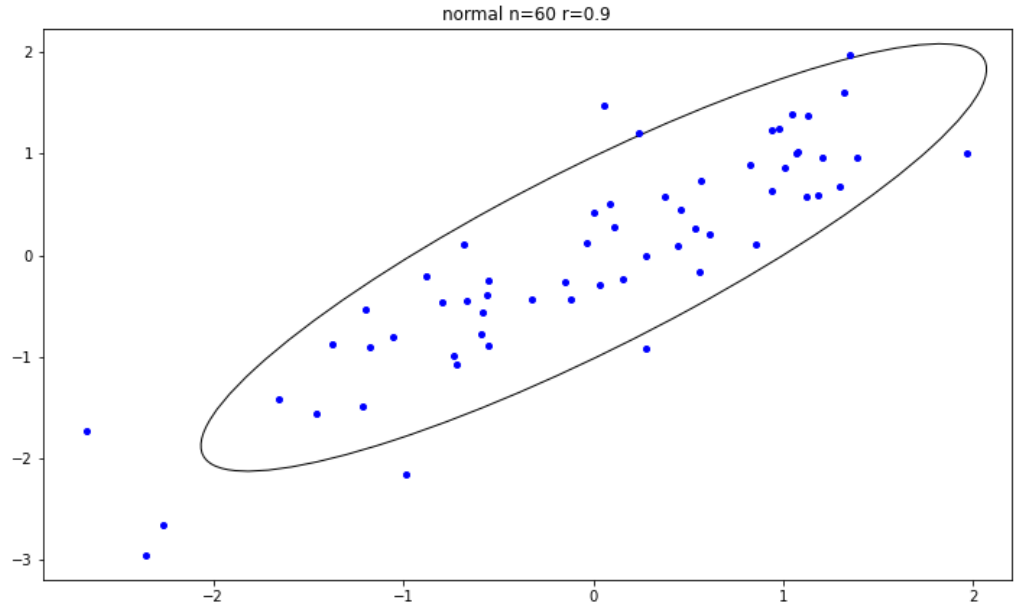


*Рис. 9 Двумерное стандартное нормальное распределение для n=100, r=0.9*

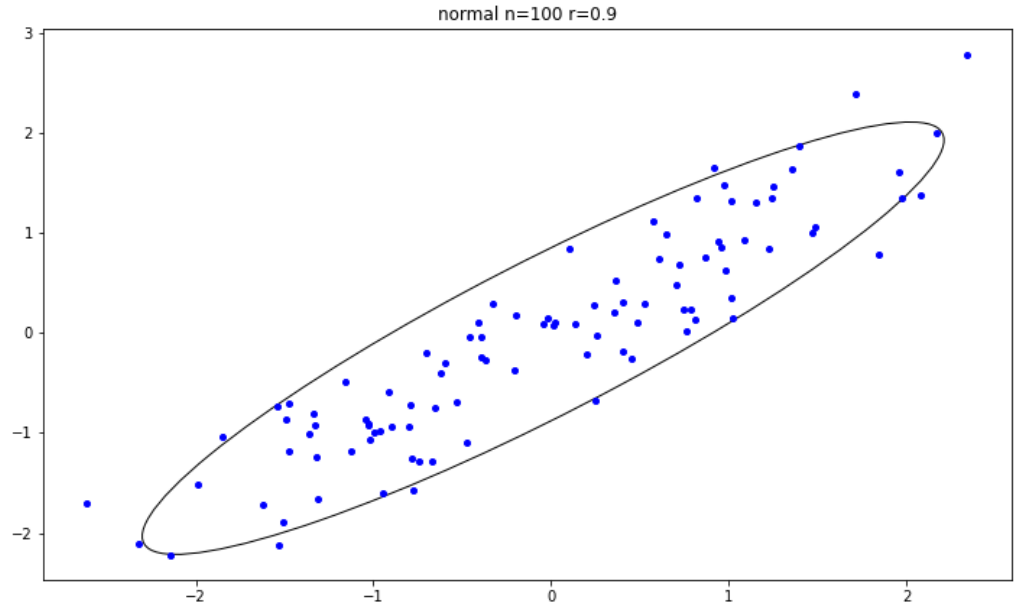



*Рис. 10 Смесь распределений для n=20*

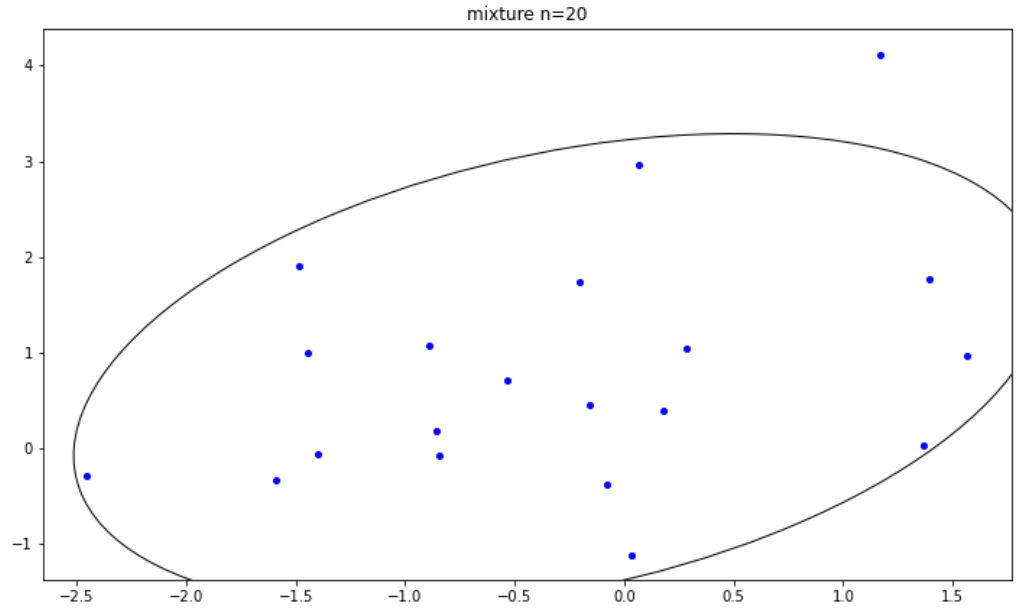


*Рис. 11 Смесь распределений для n=60*

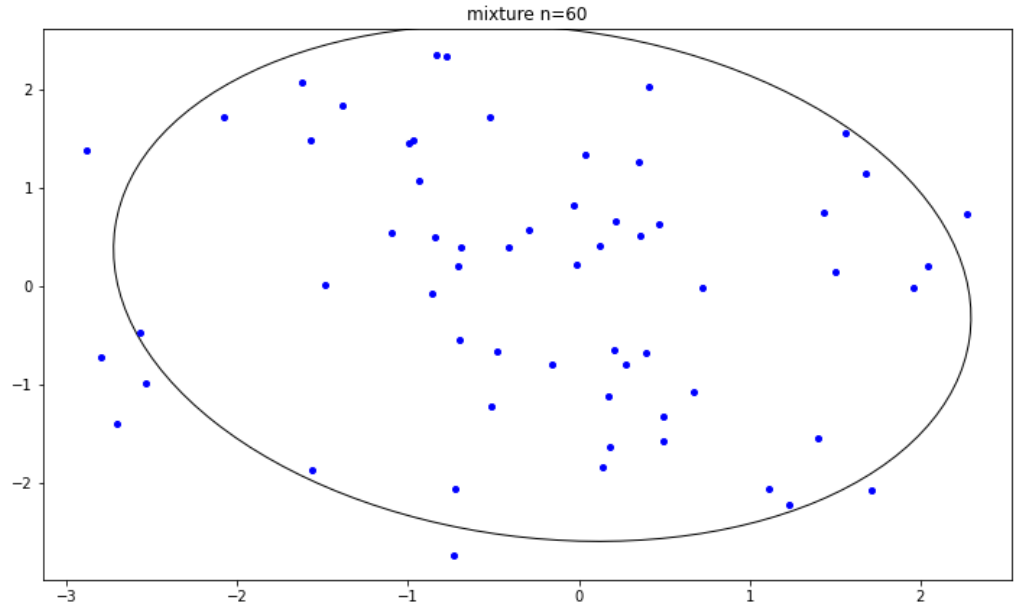



*Рис. 12 Смесь распределений для n=100*

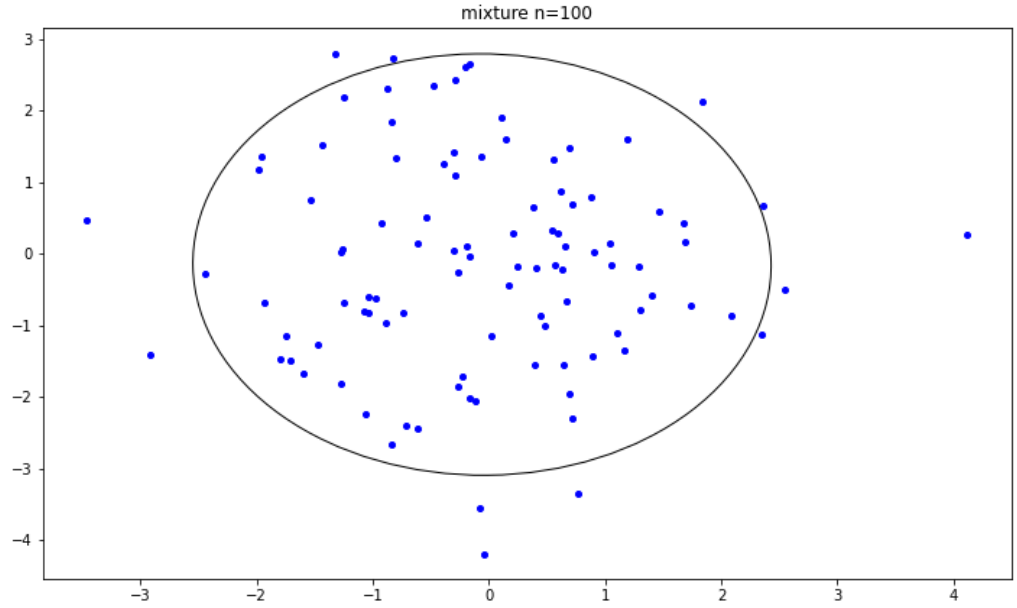


*Рис. 13 Эллипс рассеивания для 2-х точек при r=0*

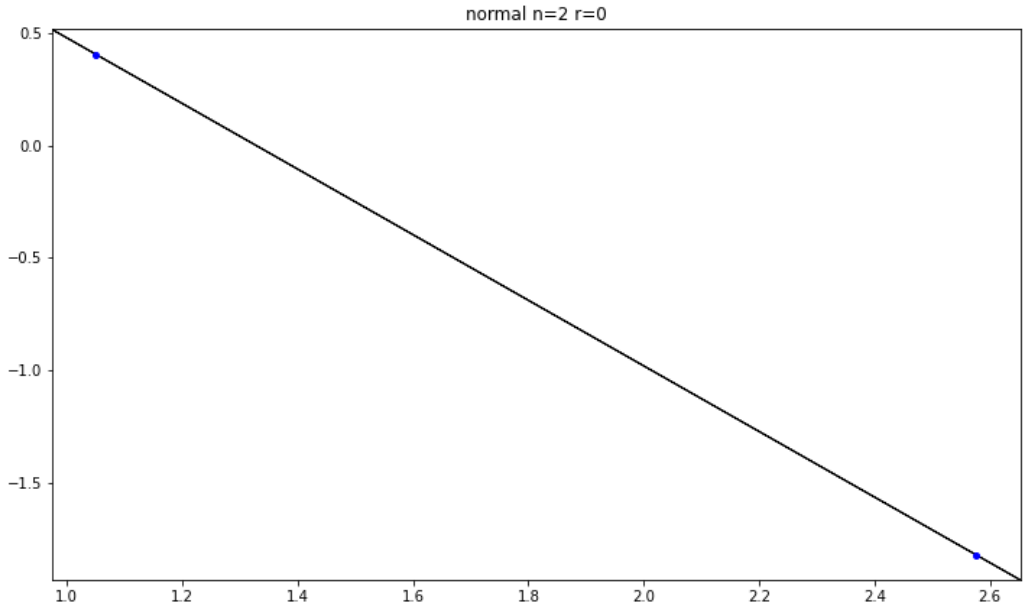



*Рис. 14 Эллипс рассеивания для 2-х точек при r=0.5*

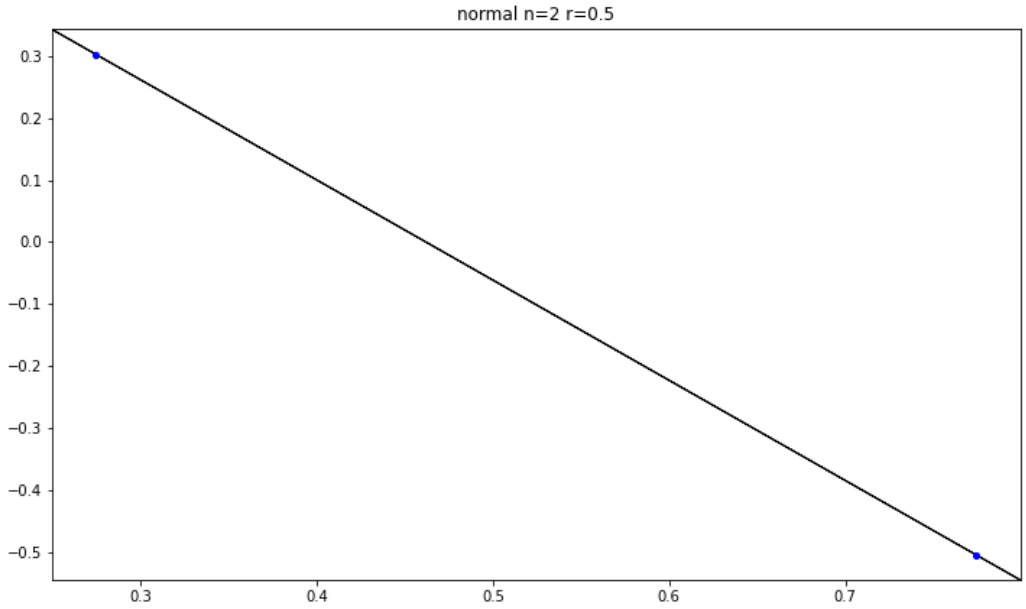


*Рис. 15 Эллипс рассеивания для 2-х точек при r=0.9*

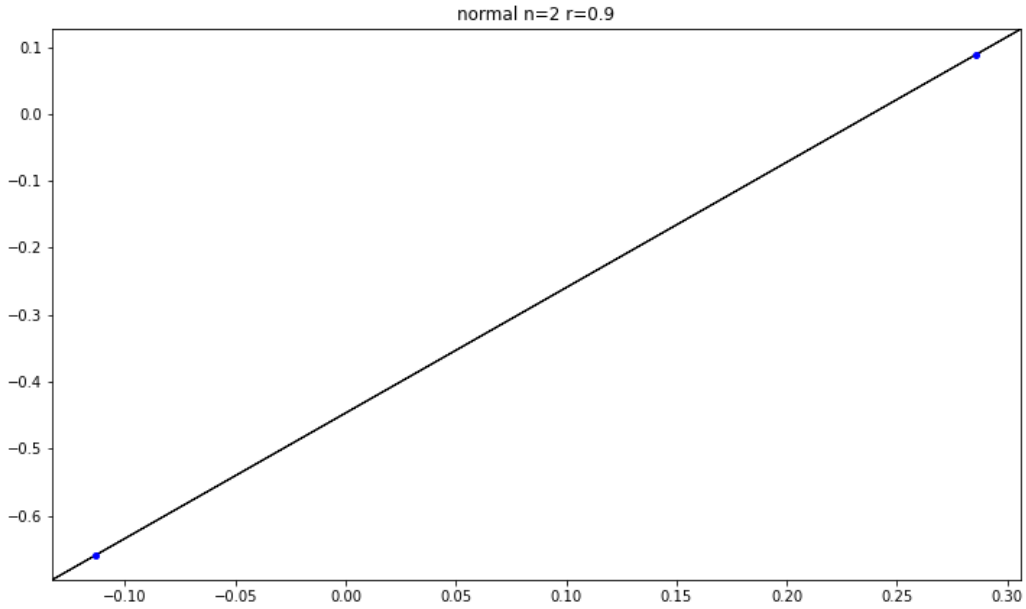



*Рис. 16 Эллипс рассеивания для 3-х точек при r=0*

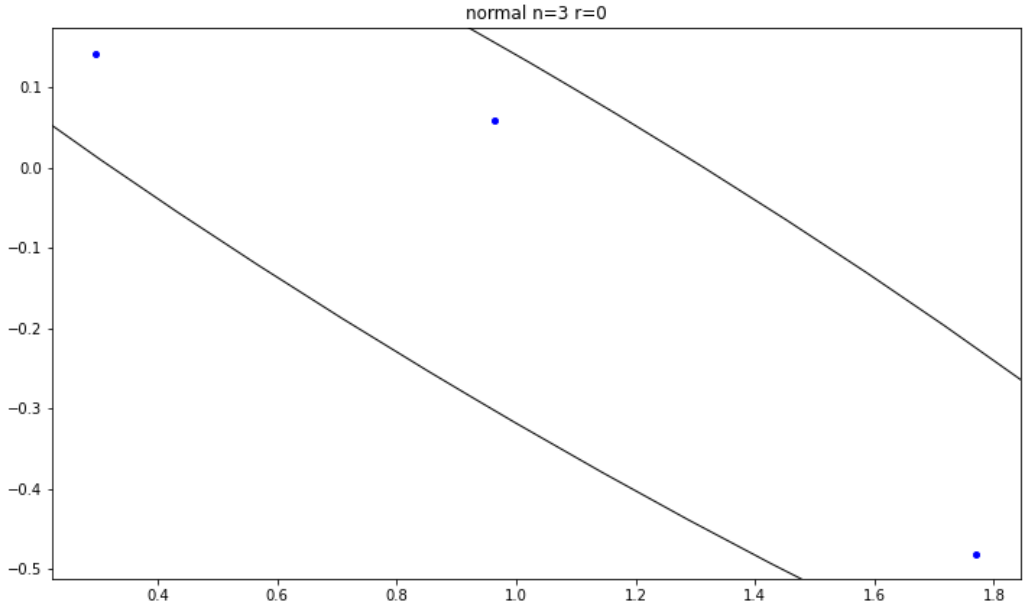


*Рис. 17 Эллипс рассеивания для 3-х точек при r=0.5*

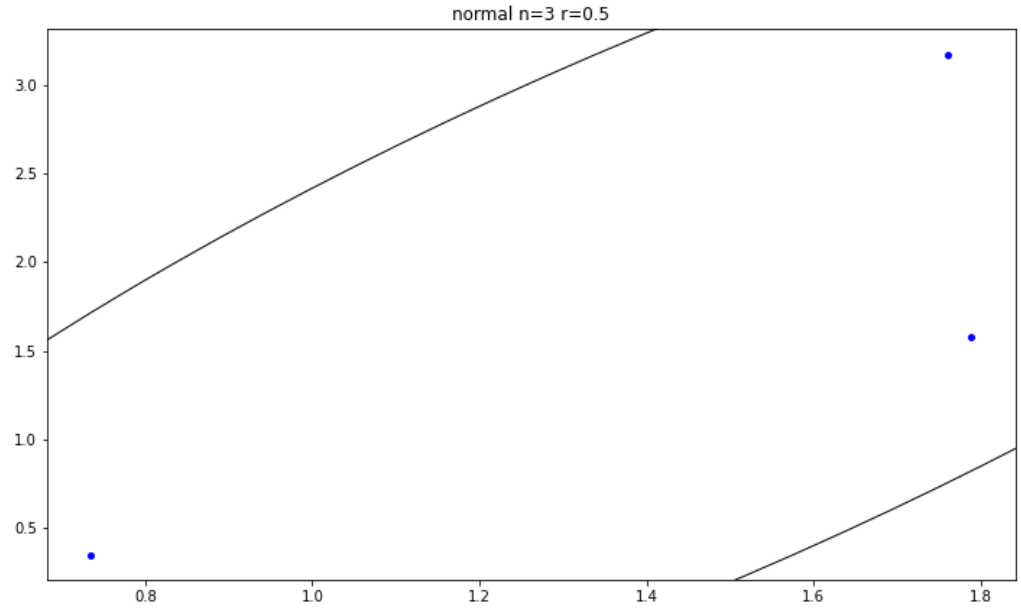
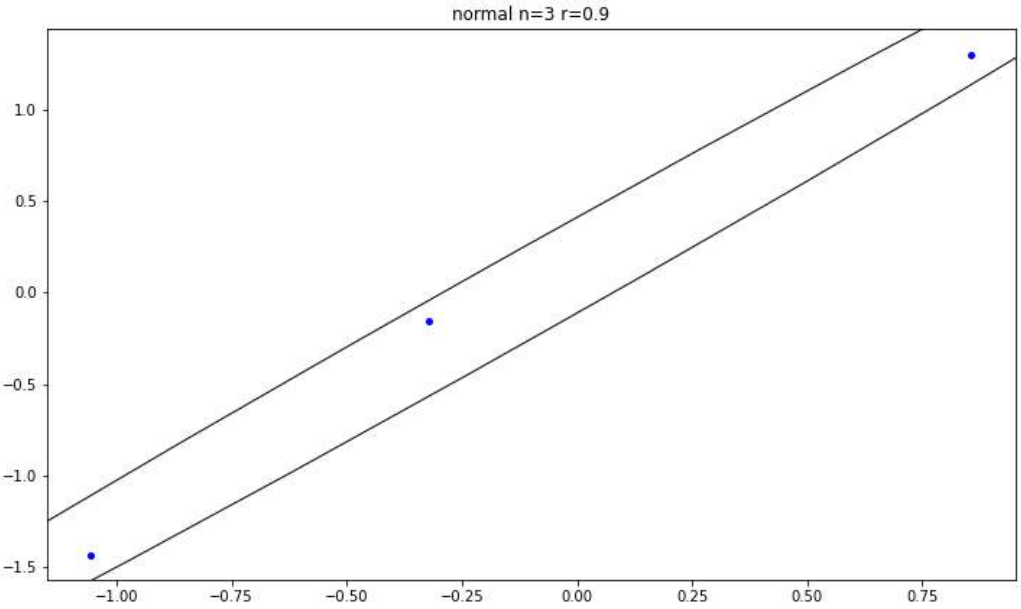



*Рис. 18 Эллипс рассеивания для 3-х точек при r=0.9*




## 4.2 Простая линейная регрессия

### 4.2.1 Таблицы

*Таблица 13 Коэффициенты при выборке без возмущений*

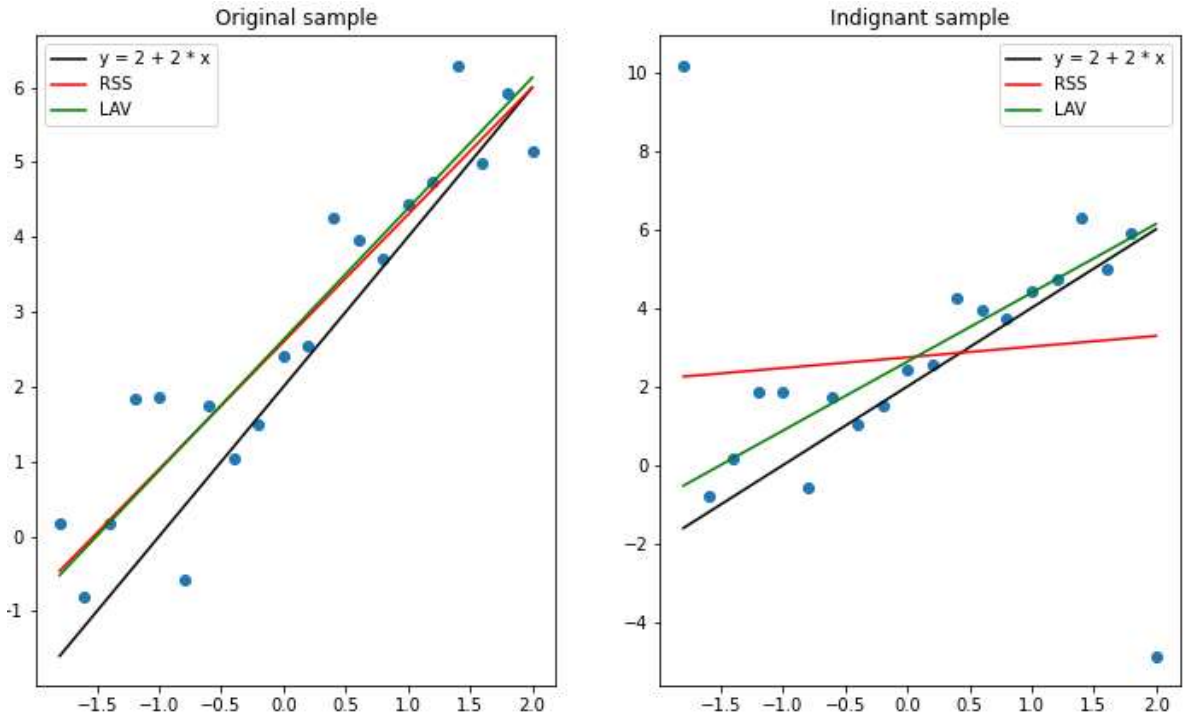
| Исходная выборка | $\hat{a}$ | $\hat{b}$ |
|---|---|---|
| RSS | 1.70073 | 2.599262 |
| LAD | 1.75065 | 2.631271 |

*Таблица 14 Коэффициенты при выборке с возмущениями*

| Выборка с возмущениями | $\hat{a}$ | $\hat{b}$ |
|---|---|---|
| RSS | 0.272159 | 2.742119 |
| LAD | 1.751878 | 2.631394 |

### 4.2.2 Иллюстрации

*Рис. 19 График получившейся линейной регрессии*




## 4.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

### 4.3.1 Нормальное распределение

При подсчёте оценок параметров закона нормального распределения с помощью МНМ были получены следующие результаты:

$$\hat{\mu}_{\text{ОМП}} = 0.0598$$
$$\widehat{\sigma^2}_{\text{ОМП}} = 1.0079$$

*Таблица 15 Вычисления $\chi^2$*

| $i$ | $\Delta_i$ | $n$ | $p_i$ | $\chi_B^2$ |
|---|---|---|---|---|
| **1** | -1 | 15.0 | 0.1465 | 0.0083 |
| **2** | -0.5 | 14.0 | 0.1428 | 0.0055 |
| **3** | 0 | 16.0 | 0.1870 | 0.3909 |
| **4** | 0.5 | 27.0 | 0.1925 | 3.1189 |
| **5** | 1 | 10.0 | 0.1557 | 1.9923 |
| **6** | inf | 18.0 | 0.1755 | 0.0118 |

$$\chi_B^2 = 5.5277$$

### 4.3.2 Равномерное распределение

Размер выборки $n = 20$, заданный отрезок $[-2, 2]$.

$$U(x, -2,2) = \begin{cases} \frac{1}{4}, x \in [-2, 2] \\ 0, \text{иначе} \end{cases}$$

$$\hat{\mu}_{\text{ОМП}} = -0.0309$$

$$\widehat{\sigma^2}_{\text{ОМП}} = 0.8724$$

*Таблица 16 Вычисления $\chi^2$*

| $i$ | $\Delta_i$ | $n$ | $p_i$ | $\chi_B^2$ |
|---|---|---|---|---|
| **1** | -2 | 0.0 | 0.0120 | 0.2399 |
| **2** | 1 | 18.0 | 0.8693 | 0.0216 |
| **3** | 4 | 2.0 | 0.1187 | 0.0587 |
| **4** | inf | 0.0 | 0.0000 | 0.0000 |

$$\chi_B^2 = 0.3203$$

## 4.4 Интервальные оценки математического ожидания и стандартного отклонения

*Таблица 17 Результаты для выборок мощности n=20*

| **n=20** | $\mu$ | $\sigma$ |
|---|---|---|
| **normal_dist** | $[0.1611, 0.9776]$ | $[0.6634, 1.274]$ |
| **random_dist** | $[0.1967, 0.9419]$ | $[0.6909, 1.2196]$ |

*Таблица 18 Результаты для выборок мощности n=100*

| **n=100** | $\mu$ | $\sigma$ |
|---|---|---|
| **normal_dist** | $[-0.1658, 0.2564]$ | $[0.9341, 1.2359]$ |
| **random_dist** | $[-0.1622, 0.2527]$ | $[0.9558, 1.2035]$ |

# 5 Выводы

## 5.1 Выборочные коэффициенты корреляции и эллипсы рассеивания

Ближе всего к теоретическому коэффициенту корреляции находится коэффициент Пирсона.

По графикам видно, что

- при увеличении объёма выборки коэффициенты корреляции стремятся к теоретическим
- при уменьшении корреляции эллипс рассеивания стремится к окружности, а при увеличении – вырождается в прямую с углом наклона в $45^o$ против часовой стрелки

- для построения эллипса рассеивания нужно минимум 3 точки, а при 2-х точках эллипс вырождается в прямую под определённым углом

### 5.2 Простая линейная регрессия

По графикам видно, что оба метода дают хорошую оценку, если нет выбросов. Однако выбросы сильно влияют на оценки по МНК.

Выбросы слабо влияют на оценку по МНМ, но ценой за это является бóльшая вычислительная сложность.

### 5.3 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Табличное значение квантиля $\chi^2_{0.95}(5) = 11.0705$. Полученное значение критерия согласия Пирсона для нормального распределения $\chi^2_B = 5.5277 < 11.0705$, следовательно основная гипотеза $H_0$ не может быть опровергнута на уровне значимости $\alpha = 0.05$.

Для равномерного распределения полученное значение критерия Пирсона $\chi^2_B = 0.3203 < \chi^2_{0.95}(3) = 7.81473$ означает, что из полученной выборки мы не можем опровергнуть гипотезу $H_0$ о нормальности данного распределения.

### 5.4 Интервальные оценки математического ожидания и стандартного отклонения

Точность оценок растёт с увеличением объёма выборки, оба метода показывают примерно одинаковое качество оценок, но у асимптотического подхода (**random_dist**) очевидное преимущество.

## 6 Литература


Основы работы с *numpy* (отдельная глава курса)

Pandas обзор

Документация по *scipy*

Таблица значений $\chi^2$


## 7 Приложения

Код лабораторной №5

Код лабораторной №6

Код лабораторной №7

Код лабораторной №8