Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчет по лабораторным работам № 5-8
по дисциплине: Математическая статика.

Выполнила студентка:
Заболотских Екатерина Дмитриевна
группа: 3630102/70301

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

Санкт-Петербург
2020 г.

Оглавление

Список иллюстраций

Список таблиц

Постановка задачи

1. Лабораторная

Сгенерировать двумерные выборки размера 20, 60, 100 для нормального двумерного распределения $\mathrm{N}(x, y, 0, 0, 1, 1, \rho)$. Коэффициент корреляции ρ взять равным 0, 0.5, 0.9. Каждую выборку сгенерировать 1000 раз и вычислить: среднее значение, среднее значение квадрата, дисперсию коэффициентов корреляции Пирсона, Спирмена и квадратного коэффициента корреляции.

Повторить все вычисления для смеси нормальных распределений:

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

Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

2. Лабораторная

Найти оценки коэффициентов a, b линейной регрессии $y_i = a + bx_i + \varepsilon_i$, используя 20 точек на отрезке $[-1.8, 2]$ с равномерным шагом равным 0.2. Ошибку ε_i считать нормально распределённой с параметрами (0,1). В качестве эталонной зависимости взять $y_i = 2 + 2x_i + \varepsilon_i$. 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_2 вносятся возмущения 10 и -10.

3. Лабораторная

Сгенерировать выборку объемом 100 элементов для нормального распределения $\mathcal{N}(0, 1)$. По ней оценить параметры μ и σ нормального закона методом максимального правдоподобия. В качестве основной гипотезы H_0 будем считать, что сгенерированное распределение имеет вид $\mathcal{N}(\widehat{\mu}, \widehat{\sigma})$. Проверить основную гипотезу, используя критерий согласия χ^2. В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений χ^2.

Дополнительное задание:
Создать выборку распределения Лапласа мощностью 20 событий и проверь гипотезу ее «нормальности».

4. Лабораторная

Для двух выборок размерами 20 и 100 элементов, сгенерированных согласно нормальному закону $N(x, 0, 1)$ для параметров положения масштаба 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 надёжности взять $\gamma = 0.95$.

Теория

Двумерное нормальное распределение

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

$$\mathrm{N}(x, y, m_1, m_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right) \quad (1)$$

В свою очередь компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $m_X = m_1, m_Y = m_2$ и среднеквадратичными отклонениями $\sigma_X = \sigma_1, \sigma_Y = \sigma_2$. В свою очередь, параметр ρ – коэффициент корреляции.

Ковариация и коэффициент корреляции

Ковариацией двух случайных величин X и Y называется величина:

$$K_{XY} = M\,[\,(X - m_X)(Y - m_Y)] \quad (2)$$

В свою очередь коэффициентом корреляции называется:

$$\rho_{XY} = \frac{K_{XY}}{\sigma_X\sigma_Y} \quad (3)$$

Коэффициент корреляции характеризует зависимость между случайными величинами X и Y. Именно его мы задаем в двумерном нормальном распределении как ρ. Если случайные величины X и Y независимы, то $\rho_{XY} = 0$ т.к. в этом случае очевидно $K_{XY} = 0$.

Выборочные коэффициенты корреляции Пирсона

Пусть по выборке значений $\{x_i, y_i\}_{i=1}^n$ двумерной случайной величины (X, Y). Естественной оценкой для ρ_{XY} служит выборочный коэффициент корреляции (Пирсона):

$$r = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \;\; \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

Важное свойство: при данной оценке гипотеза $\rho_{XY} \neq 0$ может быть принята на уровне значимости 0.05 если выполнено:

$$|r|\sqrt{n-1} > 2.5 \quad (5)$$

Квадратный

Выборочным квадрантным коэффициентом корреляции называется величина:

$$r_Q = \frac{(n_1 + n_3)(n_2 - n_4)}{n} \quad (6)$$

где n_1, n_2, n_3, n_4- количества элементов выборки попавших соответственно в I, II, III и IV квадранты декартовой системы координат с центром в $(med\ x, med\ y)$ и осями

$x_1 = x - med\ x, y_1 = y - med\ y$, где med – выборочная медиана.

Формулу (6) можно переписать эквивалентным образом:

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} sign(x_i - med\ x)sign(y_i - med\ y) \quad (7)$$

Спирмена

На практике нередко требуется оценить степень взаимодействия между качественными признаками изучаемого объекта. Качественным называется признак, который нельзя измерить точно, но который позволяет сравнивать изучаемые объекты между собой и располагать их в порядке убывания или возрастания их качества. Для этого объекты выстраиваются в определённом порядке в соответствии с рассматриваемым признаком. Процесс упорядочения называется ранжированием, и каждому члену упорядоченной последовательности объектов присваивается ранг, или порядковый номер.

Например, объекту с наименьшим значением признака присваивается ранг 1, следующему за ним объекту – ранг 2, и т.д. Таким образом, происходит сравнение каждого объекта со всеми объектами изучаемой выборки. Если объект обладает не одним, а двумя качественными признаками – переменными X и Y, то для исследования их взаимосвязи используют выборочный коэффициент корреляции между двумя последовательностями рангов этих признаков.

Обозначим ранги, соответствующие значениям переменной X, через u, а ранги, соответствующие значениям переменной Y, – через v. Выборочный коэффициент ранговой корреляции Спирмена определяется как выборочный коэффициент корреляции Пирсона между рангами u, v переменных X, Y:

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

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

Эллипсы рассеивания

Рассмотрим поверхность распределения, изображающую функцию (1). Она имеет вид холма, вершина которого находится над точкой $(\bar{x}, \bar{y})$.

В сечении поверхности распределения плоскостями, параллельными оси $N(x, y, x, y, \sigma x, \sigma y, \rho)$, получаются кривые, подобные нормальным кривым распределения. В сечении поверхности распределения плоскостями, параллельными плоскости xOy, получаются эллипсы. Напишем уравнение проекции такого эллипса на плоскость xOy:

$$\frac{(x-\bar{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} = const \quad (9)$$

Уравнение эллипса (9) можно проанализировать обычными методами аналитической геометрии. Применяя их, убеждаемся, что центр эллипса (9) находится в точке с координатами $(\bar{x}, \bar{y})$; что касается направления осей симметрии эллипса, то они составляют с осью Ox углы, определяемые уравнением:

$$tg2\alpha = \frac{2p\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \quad (10)$$

Это уравнение дает два значения углов: α и $\alpha 1$, различающиеся на $\frac{\pi}{2}$.

Таким образом, ориентация эллипса (9) относительно координатных осей находится в прямой зависимости от коэффициента корреляции ρ системы (X, Y); если величины не коррелированы (т.е. в данном случае и независимы), то оси симметрии эллипса параллельны координатным осям; в противном случае они составляют с координатными осями некоторый угол. Пересекая поверхность распределения плоскостями, параллельными плоскости xOy, и проектируя сечения на плоскость xOy мы получим целое семейство подобных и одинаково расположенных эллипсов с общим центром (x, y). Во всех точках каждого из таких эллипсов плотность распределения $N(x, y, x, y, \sigma x, \sigma y, \rho)$ постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче эллипсами рассеивания. Общие оси всех эллипсов рассеивания называются главными осями рассеивания.

В данной работе, для выборки построенной по распределению $\mathrm{N}(x, y, m_1, m_2, \sigma_1, \sigma_2, \rho)$ эллипсы равновероятности строились таким образом, чтобы покрыть все элементы выборки т.е. в качестве константы, стоящей в правой части уравнения (9) бралась:

$$R = \max_{\{(x_i, y_i)\}_{i=1}^n} \left(\frac{(x_i - m_1)^2}{\sigma_1^2} - 2\rho\frac{(x_i - m_1)(y_i - m_2)}{\sigma_1\sigma_2} + \frac{(y_i - m_2)^2}{\sigma_2^2} \right) \quad (11)$$

Простая линейная регрессия

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (12)$$

, где $\{x_i\}_{i=1}^n$ - значения фактора, $\{y_i\}_{i=1}^n$ - наблюдаемые значения отклика, а $\{\varepsilon_i\}_{i=1}^n$ - независимые, нормально распределенные по закону $\mathcal{N}(0, \sigma)$ случайные величины, а β_0, β_1 - оцениваемые параметры. Для оценки применяются различные методы, в данной работе рассмотрен следующий подход: вводится критерий рассогласования отклика и регрессионной функции, после чего оценки параметров регрессии выводятся из задачи минимизации критерия.

Критерий наименьших квадратов

Достаточно простые расчетные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции:

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

Оценки $\widehat{\beta_0}, \widehat{\beta_1}$ параметров β_0, β_1, реализующие минимум критерия (13), называют МНК-оценками:

$$\widehat{\beta_1} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \quad \widehat{\beta_0} = \bar{y} - \bar{x} \cdot \widehat{\beta_1} \quad (14)$$

Критерий наименьших модулей

Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию данных редких, но больших по величине выбросов) может быть обеспечена различными способами. Одним из них является метод нами меньших модулей вместо метода МНК:

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

Использование метода наименьших модулей в задаче оценивания параметра сдвига распределений приводит к оценке в виде выборочной медианы, обладающей робастными свойствами. В отличие от этого случая и задач метода МНК, на практике задача (15) решается численно.

В данной работе был использован метод Нелдера-Мида, применимый к негладким функциям (в том числе к $M(\beta_0, \beta_1)$).

Метод максимального правдоподобия

Пусть $x_1, \dots, x_n$ – случайная выборка из генеральной совокупности с плотностью вероятности $f(x, \theta)$; L ($x1$, ..., xn, θ) – функция правдоподобия (ФП), представляющая собой совместную плотность вероятности независимых с. в. $x_1, \dots, x_n$ и рассматриваемая как функция неизвестного параметра θ:

$$L(x_1, \dots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta) \quad (16)$$

Определение. Оценкой максимального правдоподобия (о.м.п) будем называть такое значение $\hat{\theta}$мп из множества допустимых значений параметра θ, для которого ФП принимает наибольшее значение при заданных $x_1, \dots, x_n$:

$$\hat{\theta}_{\text{мп}} = \arg\max_{\theta} L(x_1, \dots, x_n, \theta) \quad (17)$$

Если ФП дважды дифференцируема, то её стационарные значения даются корнями уравнения

$$\frac{\partial L(x_1, \dots, x_n, \theta)}{\partial \theta} = 0 \quad (18)$$

Достаточным условием того, чтобы некоторое стационарное значение $\tilde{\theta}$ было локальным максимумом, является неравенство

$$\frac{\partial^2 L}{\partial \theta^2}(x_1, \dots, x_n, \tilde{\theta}) < 0 \quad (19)$$

Определив точки локальных максимумов ФП (если их несколько), находят наибольший, который и даёт решение задачи (16).

Часто проще искать максимум логарифма ФП, так как он имеет максимум в одной точке с ФП:

$\frac{\partial lnL}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta}$, если $L > 0$,

и соответственно решать уравнение

$$\frac{\partial lnL}{\partial \theta} = 0, \quad (20)$$

Которое называют уравнением правдоподобия.

В задаче оценивания векторного параметра $\theta = (\theta 1, ..., \theta m)$ аналогично (17) находится максимум ФП нескольких аргументов:

$$\hat{\theta}_{\text{мп}} = arg \max_{\theta_1, \dots, \theta_m} L(x_1, \dots, x_n, \theta_1, \dots, \theta_m) \quad (21)$$

и в случае дифференцируемости ФП выписывается система уравнений правдоподобия

$$\frac{\partial L}{\partial \theta_k} = 0 \text{ или } \frac{\partial LnL}{\partial \theta_k} = 0, k = 1, \dots, m \quad (22)$$

Критерий согласия χ^2

Для проверки гипотезы о законе распределения применяются критерии согласия. Таких критериев существует много. Мы рассмотрим наиболее обоснованный и наиболее часто используемый в практике – критерий χ^2 (хи-квадрат), введённый К. Пирсоном (1900 г.) для случая, когда параметры распределения известны. Этот критерий был существенно уточнён Р. Фишером (1924 г.), когда параметры распределения оцениваются по выборке, используемой для проверки.

Мы ограничимся рассмотрением случая одномерного распределения.

Итак, выдвинута гипотеза H_0 о генеральном законе распределения с функцией распределения F*(x)*. Рассматриваем случай, когда гипотетическая функция распределения *F(x)* не содержит неизвестных параметров.

Разобьём генеральную совокупность, т.е. множество значений изучаемой случайной величины χ на *k* непересекающихся подмножеств $\Delta_1, \dots, \Delta_k$. Пусть $pi = P(X \in \Delta_i), i = 1, \dots, k.$

Если генеральная совокупность – вся вещественная ось, то подмножества – полуоткрытые промежутки. Крайние промежутки будут полу бесконечными:

$$\Delta_i = (a_{i-1}, a_i], i = 2, \dots, k-1,\ \Delta_1 = (-\infty, a_1],\ \Delta_k = (a_{k-1}, +\infty). \quad (23)$$

В этом случае $pi = F(a_i) - F(a_{i-1}); a_0 = -\infty, a_k = +\infty\ (i = 1, \dots, k)$.

Пусть $n_1, n_2, \dots, n_k$ – частоты попадания выборочных элементов в подмножества $\Delta_1, \dots, \Delta_k$ соответственно.В случае справедливости гипотезы H_0 относительные частоты $\frac{n_i}{n}$ при большом n должны быть близки к вероятностям $p_i\ (i = 1, \dots, k)$, поэтому за меру отклонения выборочного распределения от гипотетического с функцией $F(x)$ естественно выбрать величину

$$Z = \sum_{i=1}^{n} c_i \left(\frac{n_i}{n} - p_i\right)^2, \quad (24)$$

где c_i – какие-нибудь положительные числа (веса). К. Пирсоном в качестве весов выбраны числа $c_i = \frac{n}{p_i}\ (i = 1, \dots, k)$. Тогда получается статистика критерия хи-квадрат К. Пирсона

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \quad (25)$$

К. Пирсоном доказана теорема об асимптотическом поведении статистики χ^2, указывающая путь её применения.

Теорема К. Пирсона. Статистика критерия χ^2 асимптотически распределена по закону χ^2 с $k-1$ степенями свободы.

Это означает, что независимо от вида проверяемого распределения, т.е. функции $F(x)$, выборочная функция распределения статистики χ^2 при $n \to \infty$ стремится к функции распределения случайной величины с плотностью вероятности

$$f_{k-1}(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{1}{2^{\frac{k-1}{2}}\,\Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}, & x > 0 \end{cases} \quad (26)$$

Для пояснения сущности метода χ^2 сделаем ряд замечаний.

Замечание 2.: Выбор подмножеств $\Delta_1, \dots, \Delta_k$ и их числа k в принципе ничем не регламентируется, так как $n \to \infty$. Но так как число n хотя и очень большое, но конечное, то k, должно быть с ним согласовано. Обычно его берут таким же, как и для построения гистограммы, т.е. можно руководствоваться формулой Райса

$$k \approx 1.72 \sqrt[3]{n} \quad (27)$$

Или формулой Старджесса

$$k \approx 1 + 3.3\, lg\, n \quad (28)$$

При этом, если $\Delta_1, \Delta_2, \dots, \Delta_k$ – промежутки, то их длины удобно сделать равными за исключением крайних – полу бесконечных.

В данной работе применялось правило Скотта для ширины (считаем все интервалы кроме крайних одинаковой ширины):

$$a_i = med\ \mathcal{N}(\widehat{\mu}, \widehat{\sigma}) + \left(i - \frac{k-1}{2}\right)h, \text{где } h = \frac{3.49\widehat{\sigma}}{\sqrt[3]{n}} \quad (29)$$

Замечание 2. (о числе степеней свободы).

Числом степеней свободы функции (по старой терминологии) называется число её независимых аргументов. Аргументами статистики χ^2 являются частоты $n_1, n_2, \dots, n_k$. Эти частоты связаны одним равенством $n_1 + n_2 + \ \dots + n_k = n$, а в остальном независимы в силу независимости элементов выборки. Таким образом, функция χ^2 имеет $k-1$ независимых аргументов: число частот минус одна связь. В силу теоремы Пирсона число степеней свободы статистики χ^2 отражается на виде асимптотической плотности $f_{k-1}(x)$.

На основе общей схемы проверки статистических гипотез сформулируем следующее правило.

Правило проверки гипотезы о законе распределения по методу χ^2.

Выбираем значимости α.

По таблице находим квантиль $\chi^2_{1-\alpha}(k-1)$ распределения хи-квадрат с k-1 степенями свободы порядка $1-\alpha$.

С помощью гипотетической функции распределения $F(x)$ вычисляем вероятности $pi = P(X \in \Delta_i), i = 1, \dots, k$.

Находим частоты n_i попадания элементов выборки в подмножества $\Delta_i, i = 1, \dots, k$.

Вычисляем выборочное значение статистики критерия χ^2 :

$$\chi^2_B = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

Сравниваем χ^2_B и квантиль $\chi^2_{1-\alpha}(k-1)$.

Если $\chi^2_B < \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 на данном этапе проверки принимается.

Если $\chi^2_B \geq \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

Замечание 3. Из формулы (25) видим, что веса $c_i = n/p_i$ пропорциональны n, т.е. с ростом n увеличиваются. Отсюда следует, что если выдвинутая гипотеза неверна, то относительные частоты n_i/n не будут близки к вероятностям p_i, и с ростом n величина χ^2_B будет увеличиваться. При фиксированном уровне значимости α будет фиксированное пошаговое число – квантиль $\chi^2_{1-\alpha}(k-1)$, поэтому, увеличивая n, мы придём к неравенству $\chi^2_B > \chi^2_{1-\alpha}(k-1)$, т.е. с увеличением объема выборки неверная гипотеза будет отвергнута.

Отсюда следует, что при сомнительной ситуации, когда $\chi^2_B \approx \chi^2_{1-\alpha}(k-1)$, можно попытаться увеличить объем выборки (например, в 2 раза), чтобы требуемое неравенство было более чётким.

Замечание 4. Теория и практика применения критерия χ^2 указывают, что если для каких-либо подмножеств Δ_i $(i = 1, \dots, k)$ условие $np_i \geq 5$ не выполняется, то следует объединить соседние подмножества (промежутки).

Это условие выдвигается требованием близости величин

$$(n_i - np_i)/\sqrt{np_i},$$

Квадраты которых являются слагаемыми χ^2 к нормальным $N(0,1)$. Тогда случайная величина в формуле (25) будет распределена по закону, близкому к хи-квадрат. Такая близость обеспечивается достаточной численностью элементов в подмножествах Δ_i.

Интервальное оценивание

Интервальной оценкой (доверительным интервалом) числовой характеристики или параметра распределения θ генеральной совокупности с доверительной вероятностью γ называется интервал (θ_1, θ_2), границы которого являются случайными функциями $\theta_1 = \theta_1(x_1, \dots x_n)$, который покрывает θ с вероятностью γ:

$$P(\theta_1 < \theta < \theta_2) = \gamma \tag{30}$$

Часто вместо доверительной вероятности γ рассматривается уровень значимости $\alpha = 1 - \gamma$. Важной характеристикой данной интервальной оценки является половина длины доверительного интервала, она называется точностью интервального оценивания.

$$\Delta = \frac{\theta_1 - \theta_2}{2} \tag{31}$$

Общий вид интервальных оценок:

Пусть известна статистика $Y(\hat{\theta}, \theta)$, содержащая оцениваемый параметр θ и его точечную оценку $\hat{\theta}$. Функция $Y(\hat{\theta}, \theta)$ непрерывна и строго монотонна (для определенности строго возрастает) по θ. Известна функция распределения $F_Y(x)$, и она зависит от θ.

Зададим уровень значимости α и будем строить доверительный интервал так, чтобы $(-\infty, \alpha_1), (\alpha_2, \infty)$ накрывали θ с вероятностью $\frac{\alpha}{2}$.

Пусть $y_{\alpha/2}, y_{1-\alpha/2}$ – квантили распределения Y соответствующих порядков, тогда:

$$\begin{aligned} P(y_{\alpha/2} < Y(\hat{\theta}, \theta) < y_{1-\alpha/2}) &= F_Y(y_{1-\alpha/2}) - F_Y(y_{\alpha/2}) \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha = \gamma \end{aligned} \tag{32}$$

Так как $Y(\hat{\theta}, \theta)$ строго возрастает по θ, то у нее есть обратная функция $Y^{-1}(y)$. Она в свою очередь строго возрастает и тоже зависит от θ, следовательно:

$$\begin{aligned} &y_{\alpha/2} < Y(\hat{\theta}, \theta) < y_{1-\alpha/2} \\ &Y^{-1}(y_{\alpha/2}) < \theta < Y^{-1}(y_{1-\alpha/2}) \end{aligned} \tag{33}$$

Получаем границы интервала: $\theta_1 = Y^{-1}(y_{\alpha/2}), \theta_2 = Y^{-1}(y_{1-\alpha/2})$.

Классическое оценивание

Для математического ожидания m

Доказано, что случайная величина

$$T = \sqrt{n-1} * \frac{\overline{x} - m}{s} \tag{34}$$

называется статистикой Стьюдента, распределена по закону Стьюдента с $n-1$ степенями свободы. После некоторых выкладок имеем оценки границ интервала:

$$\begin{aligned} m_1 &= \bar{x} - \frac{x t_{1-\alpha/2}(n-1)}{\sqrt{n-1}} \\ m_2 &= \bar{x} + \frac{x t_{1-\alpha/2}(n-1)}{\sqrt{n-1}} \end{aligned} \tag{35}$$

, где $t_{1-\alpha/2}(n-1)$ – квантиль порядка $1-\alpha/2$ распределения Стьюдента с n – 1 степенями свободы.

Для среднего квадратичного отклонения σ

Доказано, что случайная величина ${ns^2}/{\sigma^2}$ распределена по закону χ^2с n – 1 степенями свободы. Применяя общий метод построения интервальных оценок, получаем оценки границ интервала:

$$\begin{aligned} \sigma_1 &= \frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} \\ \sigma_2 &= \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}} \end{aligned} \tag{36}$$

, где $\chi^2_{1-\alpha/2}(n-1), \chi^2_{\alpha/2}(n-1)$ – квантили соответствующих порядков χ^2-распределения с n – 1 степенями свободы.

Асимптотически нормальные оценки

Для математического ожидания m

В силу центральной предельной теоремы, центрированная и нормированная случайная величина ${\sqrt{n}(\bar{x} - m)}/{\sigma}$ распределена приблизительно нормально с параметрами 0 и 1. Исходя из этого получаем:

$$\begin{aligned} m_1 &= \bar{x} - \frac{s u_{1-\alpha/2}}{\sqrt{n}} \\ m_2 &= \bar{x} + \frac{s u_{1-\alpha/2}}{\sqrt{n}} \end{aligned} \tag{37}$$

, где $u_{1-\alpha/2}$ – квантиль нормального распределения $\mathcal{N}(0,1)$ порядка $1-\alpha/2$

Для среднего квадратичного отклонения σ

Аналогично, в силу ЦПТ, центрированная и нормированная случайная величина $(s^2 - M_{s^2})/\sqrt{D_{s^2}}$ при большом объеме выборки распределена приблизительно нормально с параметрами 0 и 1. Исходя из этого получаем оценку:

$$\sigma_1 = s\left(1 + u_{1-\alpha/2}\sqrt{(e+2)/n}\right)^{-1/2}$$
$$\sigma_2 = s\left(1 - u_{1-\alpha/2}\sqrt{(e+2)/n}\right)^{-1/2} \quad (38)$$

, где e – выборочный эксцесс, определяемый по формуле:

$$e = \frac{m_4}{s^4} - 3 \quad (39)$$

, где m_4 – четвертый выборочный центральный момент, определяемый по формуле:

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{s})^4 \quad (40)$$

Реализация

Код программ, реализующий данные задачи, был написан на языке Python в интегрированной среде разработке PyCharm.

Были использованы библиотеки:

- **Numpy** – библиотека для работы с данными.
- **Matplotlib** – вывод графиков.
- **SciPy** – модуль “stats” для генерации данных, и встроенных вычислений.

Результаты

1. Лабораторная

1.1. Коэффициенты корреляции

n = 20	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.0	0.0	0.0
$E(z^2)$	0.05	0.05	0.05
$D(z)$	0.050394	0.050121	0.051036
n = 60	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.01	-0.01	-0.01
$E(z^2)$	0.016	0.016	0.018
$D(z)$	0.016349	0.015956	0.017728
n = 100	r (4)	r_s(8)	r_Q(7)
$E(z)$	-0.01	-0.0	0.0
$E(z^2)$	0.0101	0.0108	0.0102
$D(z)$	0.01010	0.010811	0.010161

Таблица 1: ρ = 0

n = 20	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	0.5	0.5	0.3
$E(z^2)$	0.27	0.25	0.15
$D(z)$	0.035037	0.035054	0.044572
n = 60	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	0.5	0.47	0.33
$E(z^2)$	0.26	0.23	0.12
$D(z)$	0.00981	0.01102	0.014527
n = 100	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	0.5	0.48	0.33
$E(z^2)$	0.26	0.24	0.12
$D(z)$	0.005412	0.005892	0.008509

Таблица 2: p = 0.5

n = 20	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	0.897	0.86	0.69
$E(z^2)$	0.81	0.75	0.5
$D(z)$	0.002323	0.004809	0.029723
n = 60	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	0.899	0.881	0.71
$E(z^2)$	0.808	0.78	0.51
$D(z)$	0.00063	0.001155	0.008627
n = 100	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	0.899	0.886	0.71
$E(z^2)$	0.809	0.786	0.51
$D(z)$	0.000407	0.000588	0.004774

Таблица 3: p = 0.9

n = 20	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	-0.0	0.5	0.5
$E(z^2)$	0.6	0.3	0.3
$D(z)$	0.457086	0.080878	0.038778
n = 60	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	-0.6	0.47	0.56
$E(z^2)$	0.5	0.25	0.32
$D(z)$	0.083955	0.029063	0.010997
n = 100	r (4)	$r_s(8)$	$r_Q(7)$
$E(z)$	-0.7	0.48	0.56
$E(z^2)$	0.5	0.25	0.32
$D(z)$	0.031831	0.016868	0.006436

Таблица 4: Смесь нормальных распределений

1.2. Эллипсы равновероятности

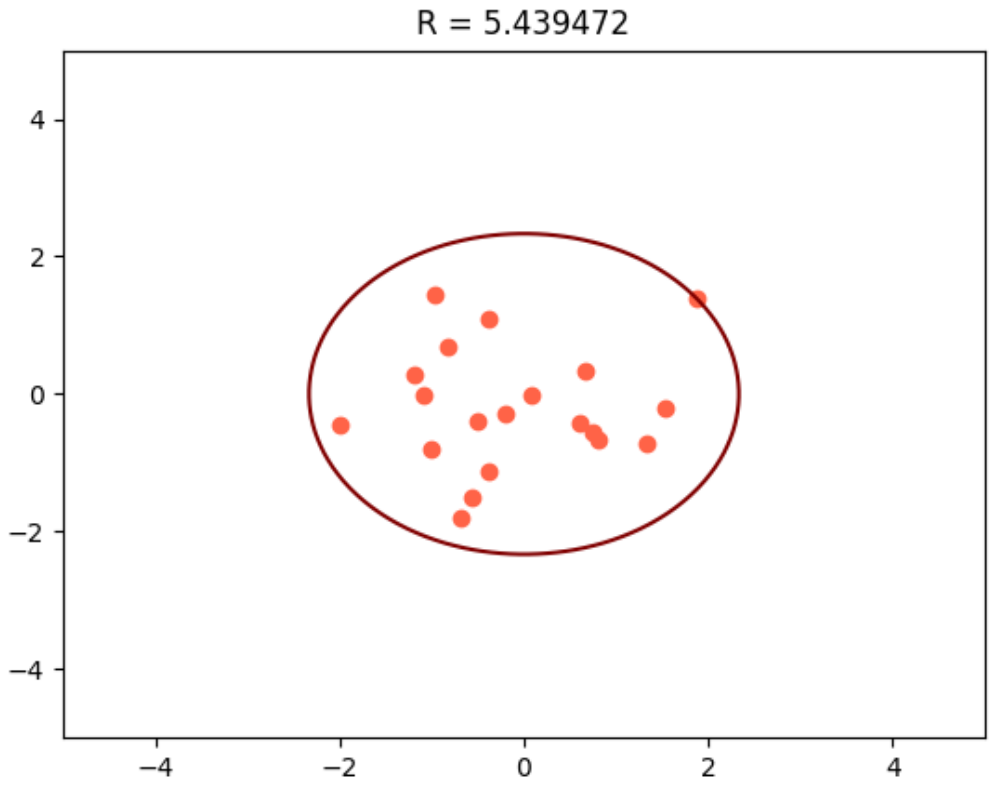

Рисунок 1: p = 0; n = 20

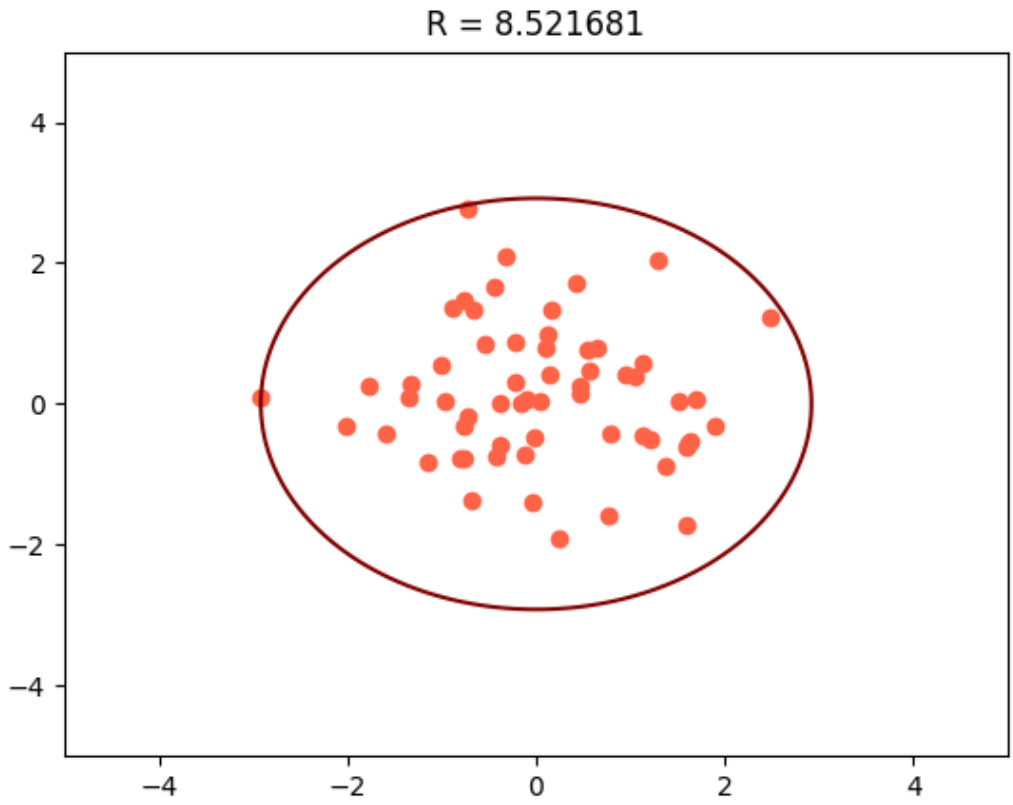

Рисунок 2: p = 0; n = 60

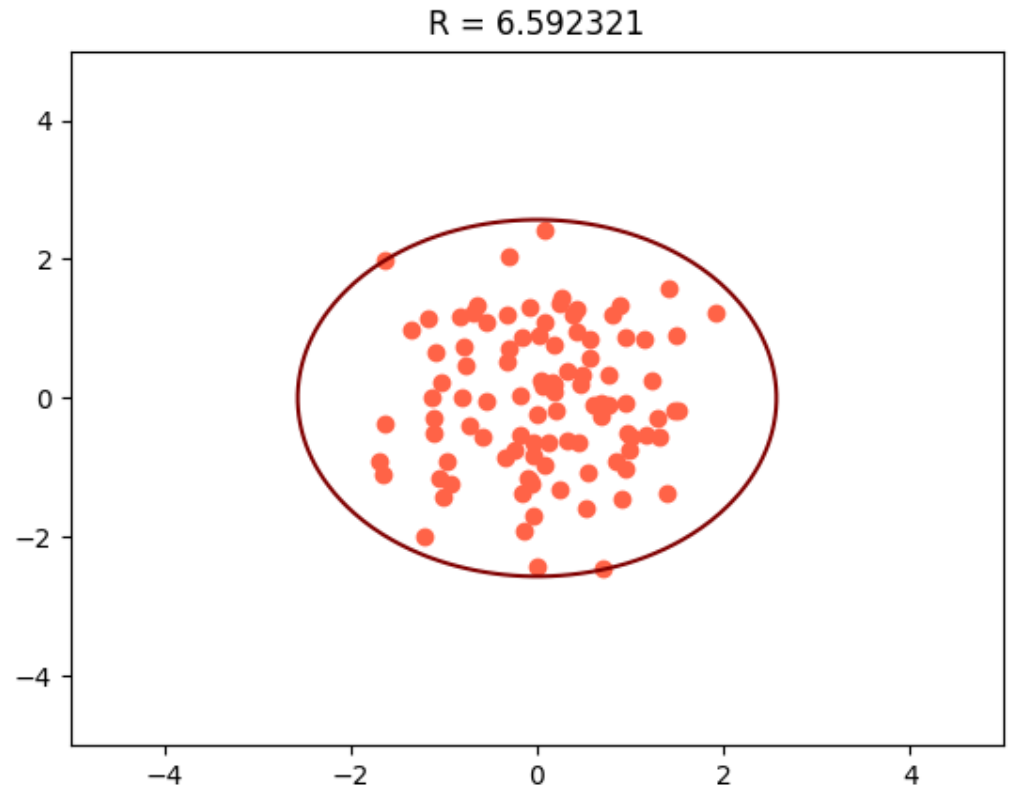

Рисунок 3: p = 0; n = 100

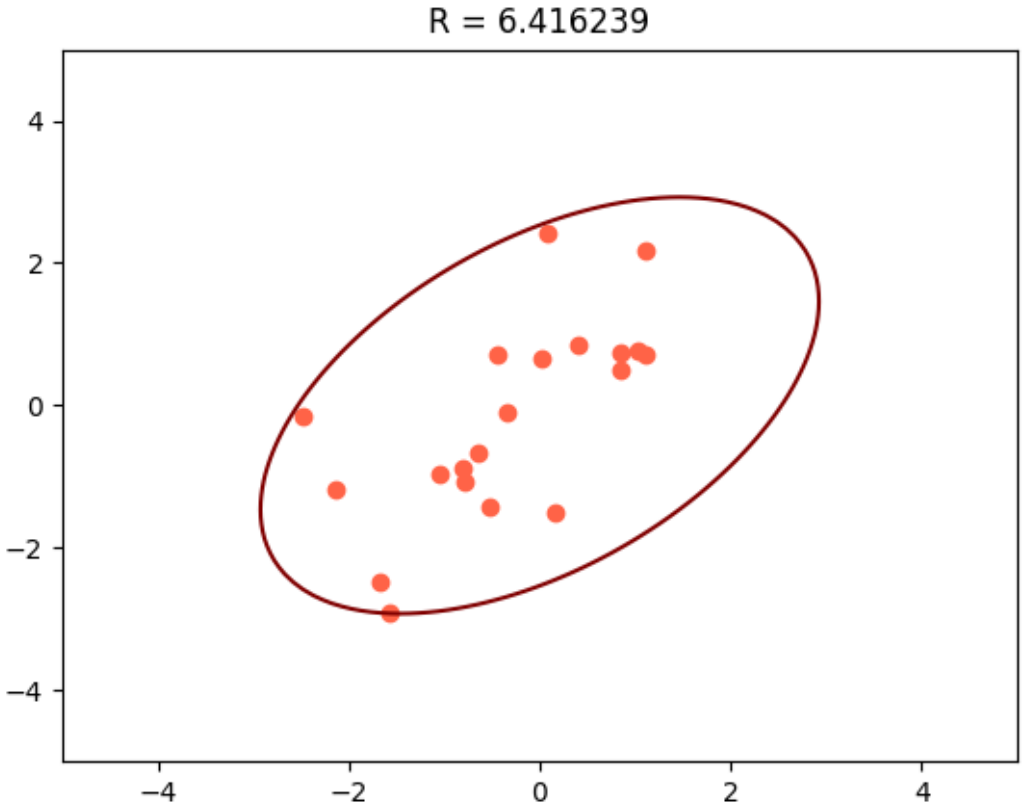

Рисунок 4: p = 0.5; n = 20

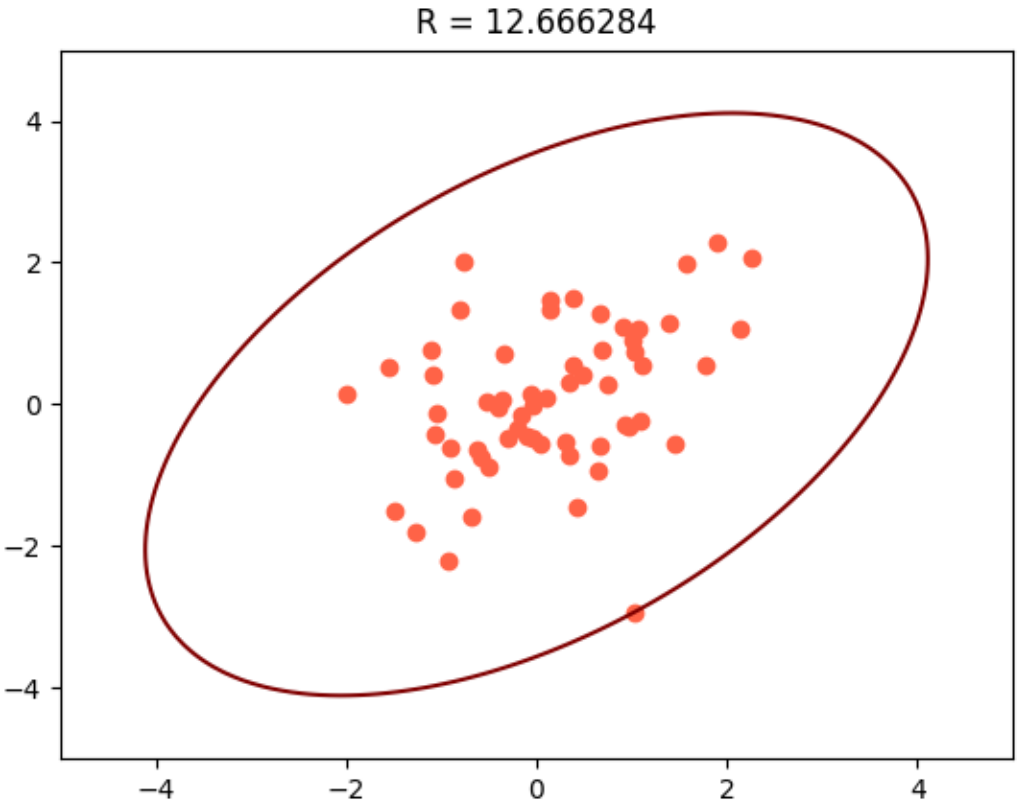

Рисунок 5: p = 0.5; n = 60

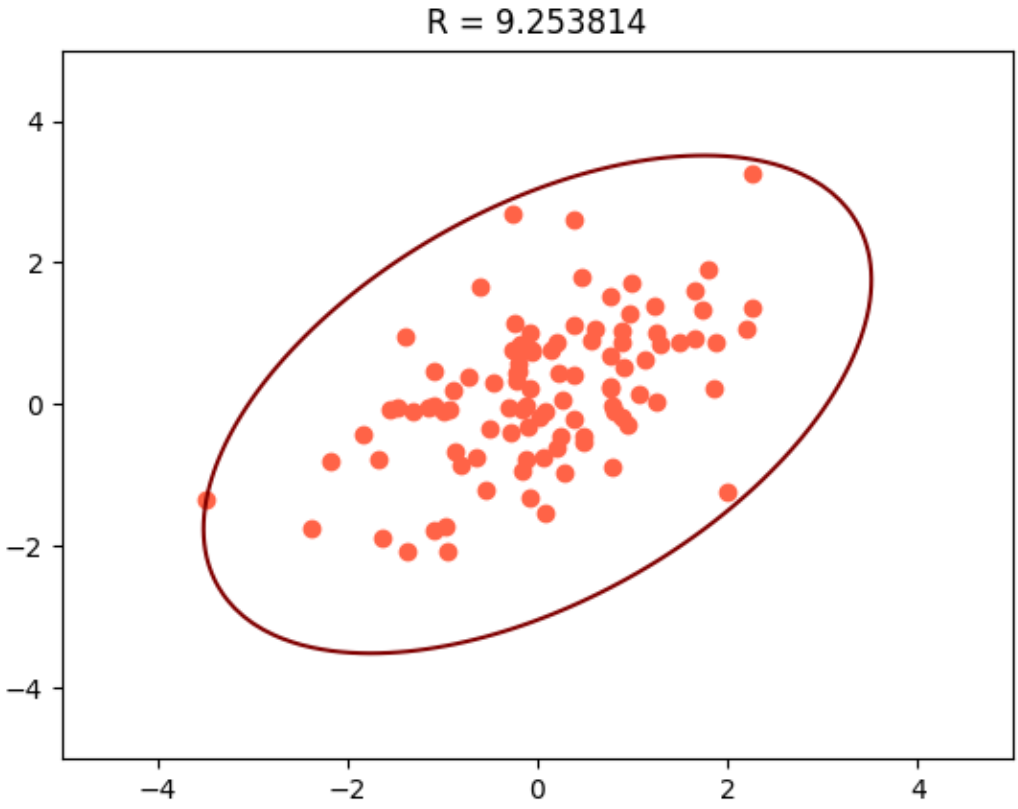

Рисунок 6: p = 0.5; n = 100

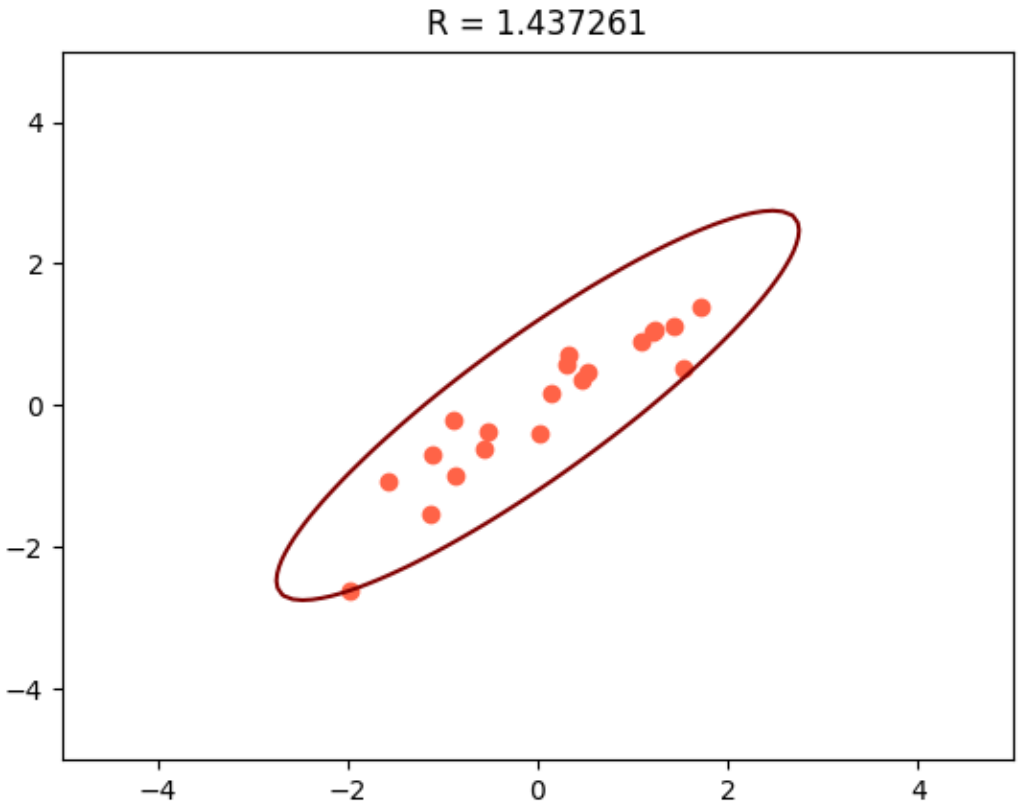

Рисунок 7: p = 0.9; n = 20

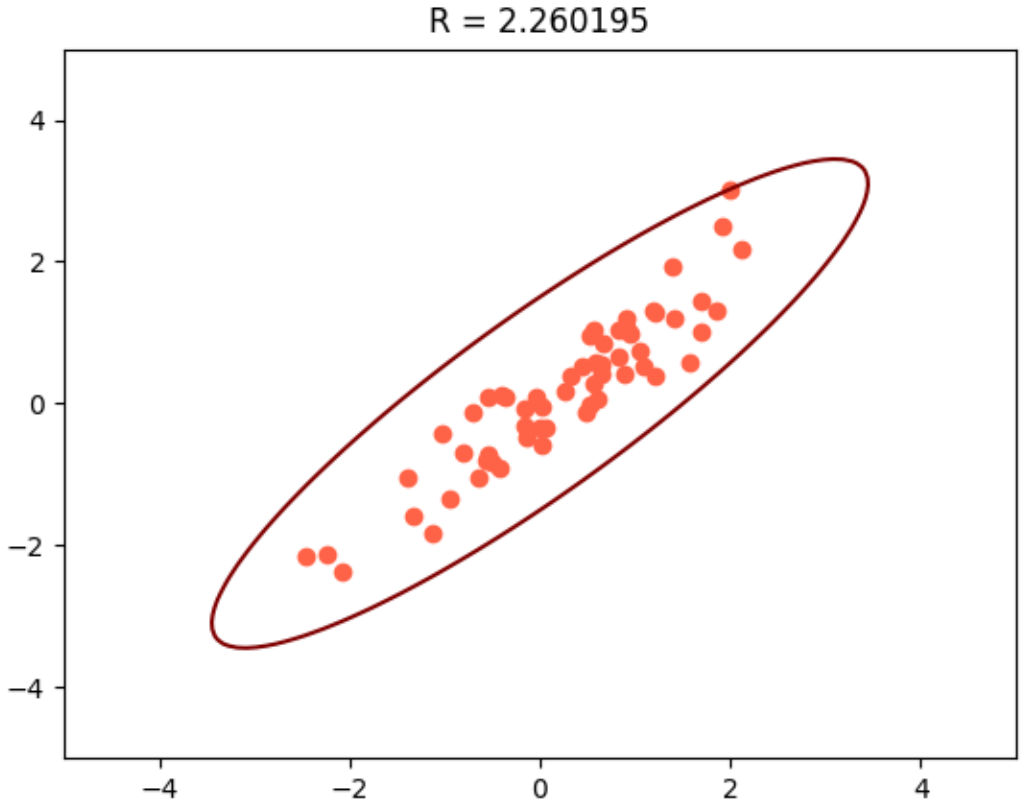

Рисунок 8: p = 0.9; n = 60

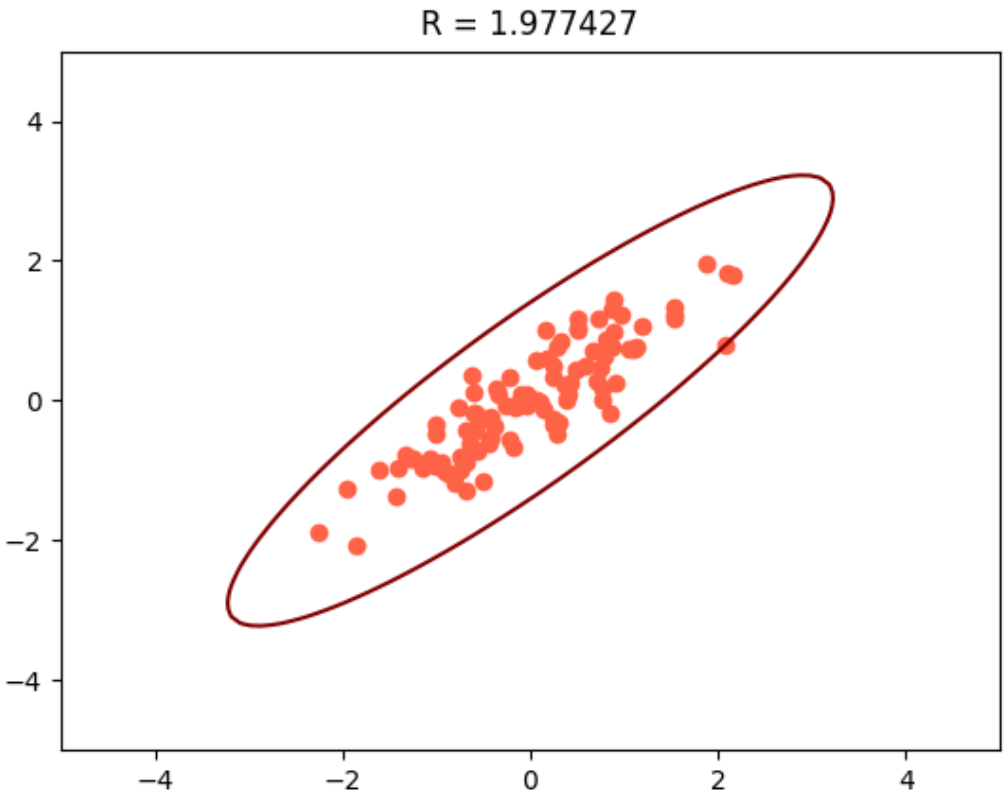

Рисунок 9: p = 0.9; n = 100

2. Лабораторная

2.1. Выборка без выбросов

- Критерий наименьших квадратов

  $\widehat{\beta_0}$ (14) = 2.16

  $\widehat{\beta_1}$ (14) = 2.11

  Q (13) = 6.9024

  M (15) = 9.8735

- Критерий наименьших модулей

  $\widehat{\beta_0}$ (14) = 1.93

  $\widehat{\beta_1}$ (14) = 2.03

  Q (13) = 8.0054

  M (15) = 9.3798

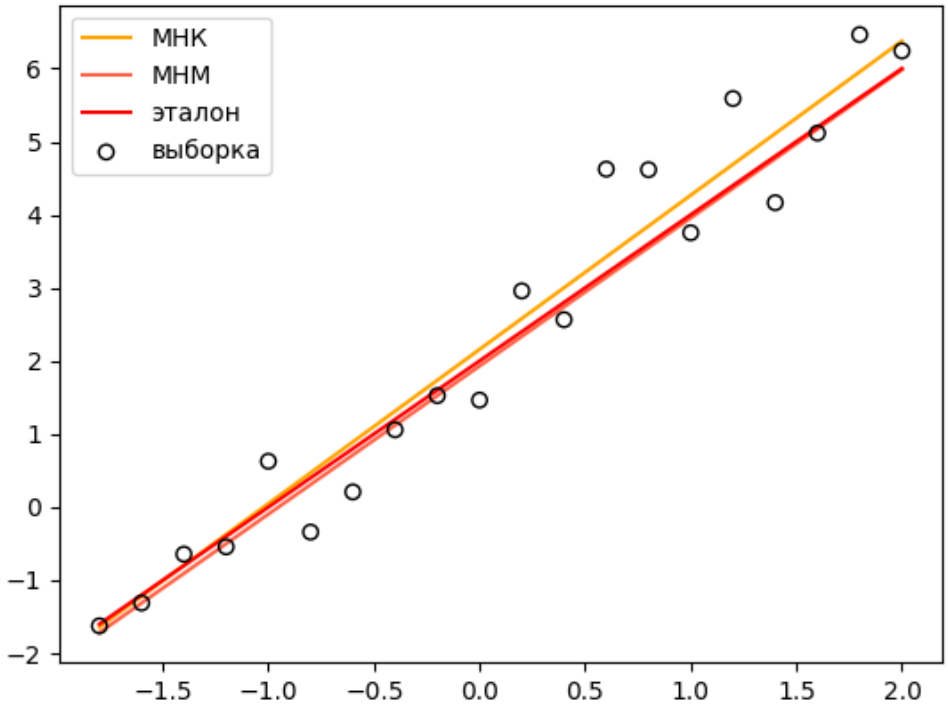

Рисунок 10: без выбросов

2.2. Выборка с выбросами

- Критерий наименьших квадратов

  $\widehat{\beta_0}$ (14) = 0.73

  $\widehat{\beta_1}$ (14) = 2.25

  Q (13) = 258.5346

  M (15) = 45.4088

- Критерий наименьших модулей

  $\widehat{\beta_0}$ (14) = 1.9

  $\widehat{\beta_1}$ (14) = 2.01

  Q (13) = 203.7506

  M (15) = 28.7808

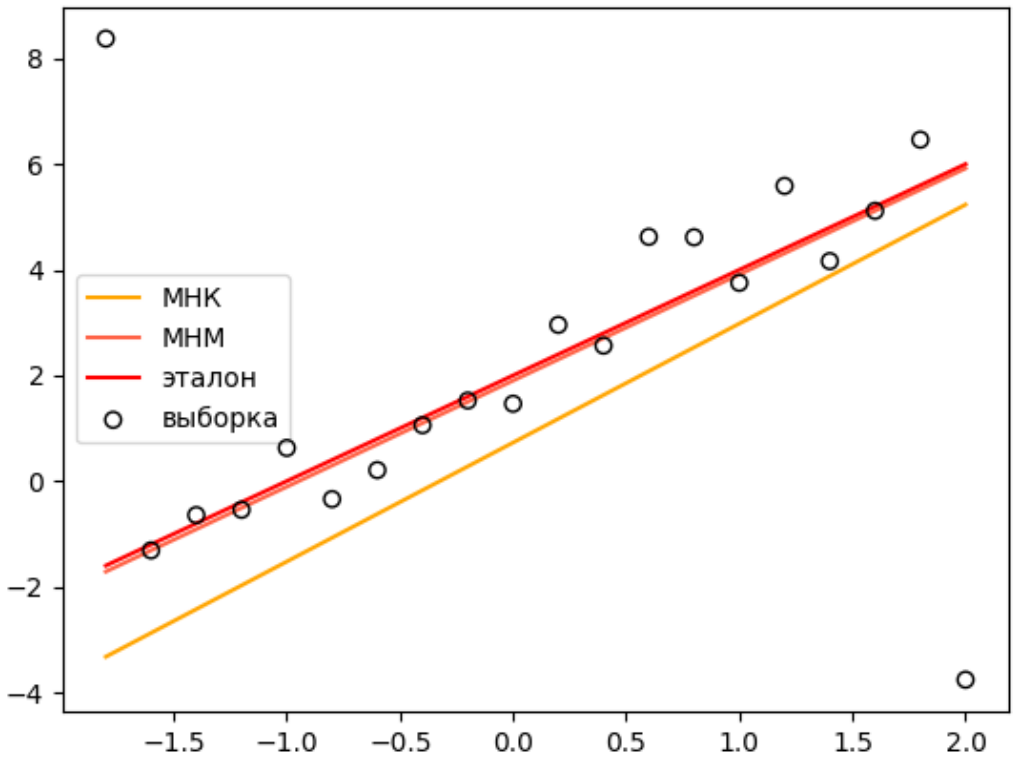

Рисунок 11: с выбросами

3. Лабораторная

$\widehat{\mu} = -0.09$ $\widehat{\sigma} = 0.9$ $\alpha = 0.5$

Число промежутков: $1.72\ \sqrt[3]{100} = 8$

Таблица вычислений χ^2:

i	Δ_i	n_i	p_i	$n_i - np_i$
1	$(-\infty;\ -2.48]$	0	0.0042	-0.4249
2	$(-2.48;\ -1.8]$	2	0.0258	-0.5823
3	$(-1.8;\ -1.12]$	15	0.0996	5.0379
4	$(-1.12;\ -0.44]$	16	0.2238	-6.3785
5	$(-0.44; 0.25]$	34	0.293	4.6955
6	$(0.25;\ 0.93]$	18	0.2238	-4.3785
7	$(0.93;\ 1.61]$	13	0.0996	3.0379
8	$(1.61;\ \infty)$	2	0.0301	-1.0071

Таблица 5: таблица вычислений χ^2

Имеем $\chi^2_{1-\alpha}(k-1) \approx 14.0671$, а вычисленное $\chi^2_{\text{в}} = 7.7947$. Видно, что $\chi^2_{\text{в}} < \chi^2_{1-\alpha}(k-1)$.

Дополнительное задание:

Для выборки в 20 элементов, сгенерированной по распределению Лапласа $\mathcal{L}(0,\ \frac{1}{\sqrt{2}})$ имеем $\chi^2_{1-\alpha}(k-1) \approx 9.4877$, а вычисленное $\chi^2_{\text{в}} = 1.1365$. Видно, что $\chi^2_{\text{в}} < \chi^2_{1-\alpha}(k-1)$. Критерий дает вывод, что генеральное распределение является нормальным $\mathcal{N}(-0.01, 0.65)$.

4. Лабораторная

4.1. Классические оценки

	m (35)	σ (36)
n= 20	-0.57 < m < 0.25	0.67 < σ < 1.28
n = 100	-0.12 < m < 0.27	0.86 < σ < 1.14

Таблица 6: Классические оценки

4.2. Асимптотически нормальные оценки

	m (37)	σ (38)
n= 20	-0.54 < m < 0.21	0.69 < σ < 1.23
n = 100	-0.12 < m < 0.27	0.86 < σ < 1.14

Таблица 7: Асимптотически нормальные оценки

Обсуждение

1. Лабораторная

1.1. Коэффициенты корреляции

Из таблиц 1, 2 и 3 видно, что r, r_s являются состоятельными оценками ρ_{XY} т.к. они все ближе к нему с ростом n.

Из таблицы 4 видим, что r_Q устойчивая к выбросам оценка. Квадрантный коэффициент корреляции показывает лучшие результаты в устойчивости.

1.2. Эллипсы равновероятности

Видно, что чем ближе ρ к 1, тем эллипс равновероятности становится все больше похож на прямую. Т.е. наглядно показано как между с.в. X и Y возникает линейная.

2. Лабораторная

Графики показали, что оценка по критерию наименьших модулей значительно лучше приближает эталонную зависимость при наличии выбросов.
С другой стороны, критерий наименьших квадратов лучше в случае отсутствия выбросов.

Полученные значения M, Q упорядочены, для оценки МНК значение Q меньше, чем для любой другой, аналогично для оценки МНМ и значения M.

3. Лабораторная

Эксперимент показал, что заданное распределение $\mathcal{N}(\widehat{\mu}, \widehat{\sigma})$ является генеральным законом, по которому построена выборка с уровнем $\alpha = 0.5$, значит оценки максимального правдоподобия состоятельны.

4. Лабораторная

Результаты показывают, что значения $m = 0, \sigma = 1$ лежат в соответствующих интервалах с вероятностью 0.95. Интервалы действительно накрывают значения параметров, причем при увеличении n асимптотические оценки почти совпадают с классическими.

Список литературы

1. Конспекты лекции
2. Википедия: https://ru.wikipedia.org/wiki

Ссылка на github: https://github.com/KateZabolotskih/MathStatLabs