Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчёт
по лабораторным работам №5-8
по дисциплине
«Математическая статистика»

Выполнил студент:
Самутичев Евгений Романович
группа: 3630102/70201

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

Санкт-Петербург
2020 г.

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

1. Сгенерировать двумерные выборки размера $20, 60, 100$ 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:

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

   . Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

2. 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 $y_i = 2 + 2x_i + e_i$. 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_2 вносятся возмущения 10 и -10.

3. Сгенерировать выборку объёмом 100 элементов для нормального распределения $N(0, 1)$. 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 $N(\widehat{\mu}, \widehat{\sigma})$. Проверить основную гипотезу, используя критерий согласия χ^2. В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений χ^2.

   Дополнительное исследование: для проверки самого критерия, сгенерировать выборки объема 20, 100 для нормального распределения $U(-1, 1)$, после чего проверить их на «нормальность».

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

2 Теория

2.1 Двумерное нормальное распределение

Двумерная случайная величина (X, Y) называется *распределенной нормально* (или просто *нормальной*) если её плотность вероятности определена формулой

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

Можно показать [1, стр. 133-134] что компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $m_X = m_1, m_Y = m_2$ и среднеквадратическими отклонениями $\sigma_X = \sigma_1, \sigma_Y = \sigma_2$. В свою очередь параметр ρ - называют *коэффициентом корреляции*. Его значение будет раскрыто далее.

2.2 Ковариация и коэффициент корреляции

Ковариацией двух случайных величин X и Y называется величина:

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

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 величиныX и Y независимы, то $\rho_{XY} = 0$ т.к. в этом случае очевидно $K_{XY} = 0$.

2.3 Выборочные коэффициенты корреляции

2.3.1 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)*:

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

Важным для приложений свойством является то что при данной оценке гипотеза $\rho_{XY} \neq 0$ (о наличии зависимости между случайными) величинами может быть принята на уровне значимости 0.05 если выполнено:

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

это можно найти к примеру в [1, стр. 538]

2.3.2 Квадрантный

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 $(\operatorname{med} x, \operatorname{med} y)$ и осями $x_1 = x - \operatorname{med} x, y_1 = y - \operatorname{med} y$, где med - выборочная медиана.

Формулу (6) можно переписать эквивалентным образом:

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

Важным свойством этой оценки является робастность. Её мы можем проверить используя схему засорения (смесь нормальных распределений).

2.3.3 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, соотвествующие значениям переменной X, через u, а ранги, соотвествующие значениям переменной Y - через v. *Выборочный коэффициент ранговой корреляции Спирмена* определяется как выборочный коэффициент корреляции Пирсона между рангами u, v переменных X, Y:

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

2.4 Эллипс равновероятности

Рассмотрим выражение для плотности двумерного нормального распределения (1) несколько подробнее, а именно найдем линии уровня или что равносильно проекции сечения графика плотности плоскостями параллельными xOy на плоскость xOy:

$$N(x, y, m_1, m_2, \sigma_1, \sigma_2, \rho) = const$$

, или что равносильно:

$$\frac{(x - m_1)^2}{\sigma_1^2} - 2\rho\frac{(x - m_1)(y - m_2)}{\sigma_1\sigma_2} + \frac{(y - m_2)^2}{\sigma_2^2} = const \tag{9}$$

Во всех точках каждого из таких эллипсов плотность двумерного нормального распределения $N(x, y, m_1, m_2, \sigma_1, \sigma_2, \rho)$ постоянна. Поэтому они и называются *эллипсами равновероятности* [2, стр. 44-45]. Отметим что в предельном случае $\rho = 1$:

$$\left(\frac{x - m_1}{\sigma_1} - \frac{y - m_2}{\sigma_2}\right)^2 = const$$

, такое уравнение задает семейство прямых паралелельных прямой:

$$\frac{x - m_1}{\sigma_1} = \frac{y - m_2}{\sigma_2} \tag{10}$$

Аналогично рассматривается предельный случай $\rho = -1$.

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

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

2.5 Простая линейная регрессия

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{12}$$

, 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 $N(0, \sigma)$ случайные величины, а β_0, β_1 - оцениваемые параметры [1, стр. 507]. Для оценки применяются различные методы, в данной работе рассмотрен следующий подход: вводится критерий рассогласования отклика и регрессионной функции, после чего оценки параметров регресии выводятся из задачи минимизации критерия. Рассмотрим два таких критерия.

2.5.1 Критерий наименьших квадратов

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

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

Приведем сами расчетные формулы [1, стр. 509]:

$$\widehat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \quad \widehat{\beta}_0 = \bar{y} - \bar{x}\widehat{\beta}_1 \tag{14}$$

Важным свойством является несмещенность оценки, однако она чувствительна к выбросам и если нужна робастная оценка, то следует рассмотреть следующий критерий.

2.5.2 Критерий наименьших модулей

В отличие от задач метода наименьших квадратов, для этого критерия минимизацию на практике проводят численно, решая:

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

В данной работе был использован метод Нелдера-Мида [3], применимый к негладким функциям (в том числе к $M(\beta_0, \beta_1)$). Подробнее см. реализация.

2.6 Точечное оценивание

2.6.1 Основные понятия

Пусть имеется выборка $\{x_i\}_{i=1}^n$ из генеральной совокупности с плотностью распределения $f(x,\theta)$. Предполагается что функциональный вид зависимости задан с точностью до неизвестного параметра θ. Требуется по выборке наблюдений $\{x_i\}_{i=1}^n$ определить число $\widehat{\theta}_n$ которое можно принять за значение параметра θ. *Точечной оценкой* неизвестного параметра θ распределения называется борелевская функция наблюдений $\widehat{\theta}_n = \widehat{\theta}_n(x_1, ..., x_n)$, приближенно равная θ. Следует заметить что параметр может быть векторным, к примеру $\theta = (\mu, \sigma)$ для нормального распределения.

2.6.2 Метод максимального правдоподобия

Рассмотрим один общий метод построения точечных оценок. Для начала введем важное понятие, *функцией правдоподобия* (ФП) называется совместная плотность вероятности распределения n независимых с.в. $x_1, ..., x_n$:

$$L\left(x_1, \ldots, x_n, \theta\right) = f\left(x_1, \theta\right) f\left(x_2, \theta\right) \ldots f\left(x_n, \theta\right) \tag{16}$$

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

$$\widehat{\theta}_{\text{мп}} = \arg\max_{\theta} L\left(x_1, \ldots, x_n, \theta\right) \tag{17}$$

Легко обобщается на случай векторного параметра $\theta = (\theta_1, ..., \theta_m)$:

$$\widehat{\theta}_{\text{мп}} = \arg\max_{\theta_1, ..., \theta_m} L\left(x_1, \ldots, x_n, \theta_1, ..., \theta_m\right) \tag{18}$$

Известно [1, стр. 444] что о. м. п. нормального распределения являются выборочное среднее и выборочная дисперсия:

$$\widehat{\mu}_{\text{мп}} = \bar{x} \quad \widehat{\sigma}_{\text{мп}} = \sqrt{s^2} \tag{19}$$

2.7 Критерий согласия χ^2

Для проверки гипотезы о законе распределения применяются различные критерии согласия. В данный работе рассматривается наиболее обоснованный и наиболее часто используемый в практике - критерий χ^2 [1, стр. 482]. И так, выдвинута гипотеза H_0 о генеральном законе распределения с функцией распределения $F(x)$. Под конкурирующей гипотезой H_1 понимается гипотеза о справедливости одного из конкурирующих распределений.

Разобьем множество значений изучаемой случайной величины X на k непересекающихся подмножеств $\Delta_1, ..., \Delta_k$ и пусть $p_i = \mathbf{P}(X \in \Delta_k)$. Если множество значений представляет вещественную ось, то подмножества имеют вид:

$$\Delta_i = (a_{i-1}, a_i], i = 2, ..., k-1 \quad \Delta_1 = (-\infty, a_1] \quad \Delta_k = (a_{k-1}, +\infty) \tag{20}$$

Пусть $n_1, ..., n_k$ - частоты попадания выборочных элементов в подмножества $\Delta_1, ..., \Delta_k$ соответственно. В случае справедливости гипотезы относительные частоты $\frac{n_i}{n}$ должны быть близки к p_i при $i = 1, ..., k$. Поэтому за меру отклонения было предолжено (К. Пирсоном) [1, стр. 483] выбрать значение

$$\chi_B^2 = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i\right)^2}{np_i} \tag{21}$$

Существует **теорема**: *статистика критерия χ^2 асимптотически распределена по закону χ^2 с $k-1$ степенями свободы.* На основе этой теоремы формируется правило проверки гипотезы о законе распределения по методу χ^2: можно принять гипотезу H_0 на уровне значимости α если $\chi^2_B < \chi^2_{1-\alpha}$, в противном случае она отвергается.

В данной работе k и длины $\Delta_1, ..., \Delta_k$ выбирались по правилам, которые обычно используют при построении гистограмм [4]. Правило Райса для числа интервалов:

$$k = \lceil 1.72\sqrt[3]{n} \rceil \tag{22}$$

и правило Фридмена-Дайсона для ширины (считаем все интервалы кроме крайних одинаковой ширины)

$$a_i = \operatorname{med} N(\widehat{\mu}, \widehat{\sigma}) + \left(i - \frac{k-1}{2}\right) h, \text{где } h = 2\frac{\mathrm{IQR}(x_1, ..., x_n)}{\sqrt[3]{n}}, i = 2, ..., k-1 \tag{23}$$

, где $\mathrm{IQR}(x_1, ..., x_n)$ - выборочная интерквартильная широта, $\operatorname{med} N(\widehat{\mu}, \widehat{\sigma})$ - медиана гипотетического распределения (т.к. предполагается что именно в окрестности медианы будет большая часть элементов выборки).

2.8 Интервальное оценивание

Интервальной оценкой (или *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: $\theta_1 = \theta_1(x_1, ..., x_n), \theta_2 = \theta_2(x_1, ..., x_n)$, который накрывает θ с вероятностью γ:

$$\mathbf{P}(\theta_1 < \theta < \theta_2) = \gamma \tag{24}$$

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

$$\Delta = \frac{\theta_2 - \theta_1}{2} \tag{25}$$

Рассмотрим общий метод построения интервальных оценок [1, стр. 456- – 457]. Пусть известна статистика $Y(\widehat{\theta}, \theta)$, содержащая оцениваемый параметр θ и его точечную оценку $\widehat{\theta}$ со следующими свойствами:

- Функция распределения $F_Y(x)$ известна и не зависит от θ
- Функция $Y(\widehat{\theta}, \theta)$ непрерывна и строго монотонна (для определенности - строго возрастает) по θ

которые мы будем проверять при построении интервальных оценок нормального распределения. 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 соотв. порядков, тогда

$$\begin{aligned} \mathbf{P}\left(y_{\alpha/2} < Y(\widehat{\theta}, \theta) < y_{1-\alpha/2}\right) = F_Y\left(y_{1-\alpha/2}\right) - F_Y\left(y_{\alpha/2}\right) = \\ = 1 - \alpha/2 - \alpha/2 = 1 - \alpha = \gamma \end{aligned} \tag{26}$$

Т.к. $Y(\widehat{\theta}, \theta)$ – строго возрастает по θ, то у неё есть обратная функция $Y^{-1}(y)$ относительно θ и она также строго возрастает, а значит:

$$\begin{gathered} y_{\alpha/2} < Y(\widehat{\theta}, \theta) < y_{1-\alpha/2} \\ Y^{-1}(y_{\alpha/2}) < \theta < Y^{-1}(y_{1-\alpha/2}) \end{gathered} \tag{27}$$

итого $\theta_1 = Y^{-1}(y_{\alpha/2})$ и $\theta_2 = Y^{-1}(y_{1-\alpha/2})$ – мы построили границы интервала. Применим это для построения интервальных оценок нормального распределения по выборке $(x_1, ..., x_n)$.

2.9 Классические оценки

2.9.1 Для математического ожидания m

Доказано что случайная величина $T = \sqrt{n-1} \cdot \frac{\bar{x}-m}{s}$ называемая статистикой Стьюдента, распределена по закону Стьюдента с $n-1$ степенями свободы, применяя с некоторыми деталями [1, стр. 457 – 458] выкладки, получаем оценки границ интервала:

$$\begin{aligned} m_1 &= \bar{x} - \frac{xt_{1-\alpha/2}(n-1)}{\sqrt{n-1}} \\ m_2 &= \bar{x} + \frac{xt_{1-\alpha/2}(n-1)}{\sqrt{n-1}} \end{aligned} \tag{28}$$

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

2.9.2 Для среднего квадратичного отклонения σ

Доказано что случайная величина ns^2/σ^2 распределена по закону χ^2 с $n-1$ степенями свободы. Применяя общий метод построения интервальных оценок получаем оценки границ интервала:

$$\begin{aligned} \sigma_1 &= \frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} \\ \sigma_2 &= \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}} \end{aligned} \tag{29}$$

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

2.10 Асимптотически нормальные оценки

2.10.1 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 [1, стр. 460] получаем оценку:

$$\begin{aligned} m_1 &= \bar{x} - \frac{su_{1-\alpha/2}}{\sqrt{n}} \\ m_2 &= \bar{x} + \frac{su_{1-\alpha/2}}{\sqrt{n}} \end{aligned} \tag{30}$$

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

2.10.2 Для среднего квадратичного отклонения σ

Аналогично, в силу центральной предельной теоремы центрированная и нормированная случайная величина $(s^2 - \mathbf{M}s^2)/\sqrt{\mathbf{D}s^2}$ при большом объеме выборки n распределена приблизительно нормально с параметрами 0 и 1. Исходя из этого [1, стр. 461] получаем оценку:

$$\begin{aligned} \sigma_1 &= s\left(1 + u_{1-\alpha/2}\sqrt{(e+2)/n}\right)^{-1/2} \\ \sigma_2 &= s\left(1 - u_{1-\alpha/2}\sqrt{(e+2)/n}\right)^{-1/2} \end{aligned} \tag{31}$$

, где e - выборочный эксцесс, определяемый как

$$e = \frac{m_4}{s^4} - 3 \tag{32}$$

, где m_4 - четвертый выборочный центральный момент, определяемый как

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \tag{33}$$

3 Реализация

Работа выполнена с использованием языка **Python** в интегрированной среде разработки **PyCharm**, были задействованы библиотеки:

- **NumPy** - векторизация вычислений, работа с массивами данных, вычисление выборочных характеристик
- **SciPy** - модуль **stats** для генерации данных по распределениям и эталонной зависимости, вычисления коэффициентов корелляции, оценок МНК, оценки методом максимального правдоподобия, модуль **optimize** для метода Нелдера-Мида
- **Matplotlib** - построение эллипсов рассеяния, построение графиков

Исходный код лабораторных работ приведен в приложении.

4 Результаты

4.1 Коэффициенты корреляции

$n = 20$	$r(4)$	$r_S(8)$	$r_Q(7)$
$E(z)$	0.0	0.0	0.0
$E(z^2)$	0.1	0.1	0.1
$D(z)$	0.053556	0.053729	0.054264
$n = 60$	r	r_S	r_Q
$E(z)$	0.0	0.0	0.0
$E(z^2)$	0.02	0.02	0.02
$D(z)$	0.016997	0.017097	0.018564
$n = 100$	r	r_S	r_Q
$E(z)$	0.0	0.0	0.0
$E(z^2)$	0.01	0.01	0.01
$D(z)$	0.010095	0.010177	0.010416

Таблица 1: $\rho = 0$

$n = 20$	r	r_S	r_Q
$E(z)$	0.5	0.5	0.3
$E(z^2)$	0.3	0.2	0.2
$D(z)$	0.03309	0.036427	0.046232
$n = 60$	r	r_S	r_Q
$E(z)$	0.49	0.47	0.33
$E(z^2)$	0.25	0.23	0.12
$D(z)$	0.009512	0.010584	0.014097
$n = 100$	r	r_S	r_Q
$E(z)$	0.5	0.48	0.33
$E(z^2)$	0.25	0.23	0.12
$D(z)$	0.006043	0.00652	0.008687

Таблица 2: $\rho = 0.5$

$n = 20$	r	r_S	r_Q
$E(z)$	0.891	0.86	0.7
$E(z^2)$	0.797	0.75	0.52
$D(z)$	0.002823	0.005097	0.02752
$n = 60$	r	r_S	r_Q
$E(z)$	0.899	0.882	0.71
$E(z^2)$	0.809	0.78	0.51
$D(z)$	0.000641	0.001065	0.008925
$n = 100$	r	r_S	r_Q
$E(z)$	0.898	0.885	0.71
$E(z^2)$	0.807	0.784	0.5
$D(z)$	0.000417	0.000637	0.004951

Таблица 3: $\rho = 0.9$

$n = 20$	r	r_S	r_Q
$E(z)$	-0.0	0.5	0.5
$E(z^2)$	1.0	0.3	0.3
$D(z)$	0.448101	0.078135	0.039344
$n = 60$	r	r_S	r_Q
$E(z)$	-0.6	0.48	0.56
$E(z^2)$	0.5	0.26	0.33
$D(z)$	0.079885	0.027009	0.01148
$n = 100$	r	r_S	r_Q
$E(z)$	-0.7	0.47	0.56
$E(z^2)$	0.51	0.24	0.33
$D(z)$	0.029483	0.015814	0.006452

Таблица 4: Смесь нормальных распределений

4.2 Эллипсы равновероятности

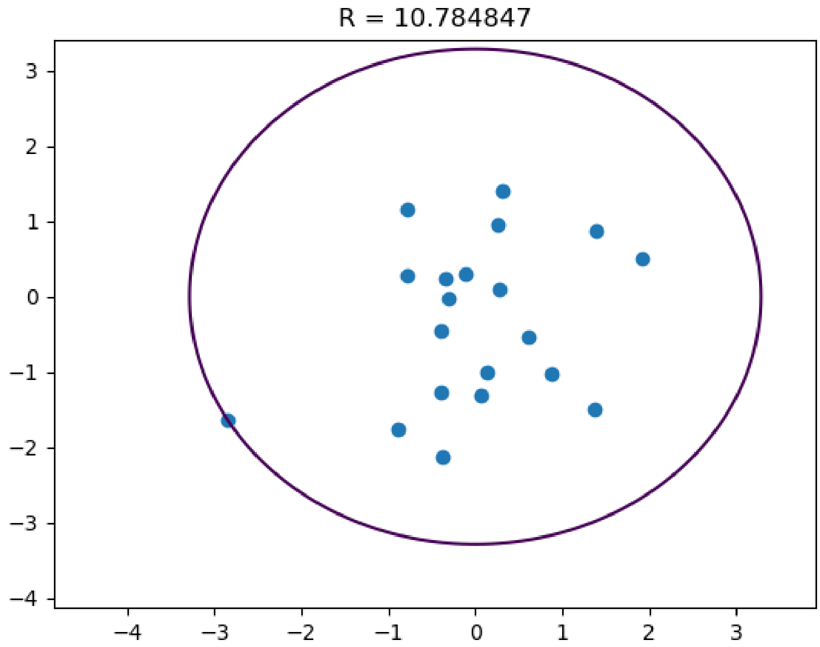

Рис. 1: $\rho = 0.0, n = 20$

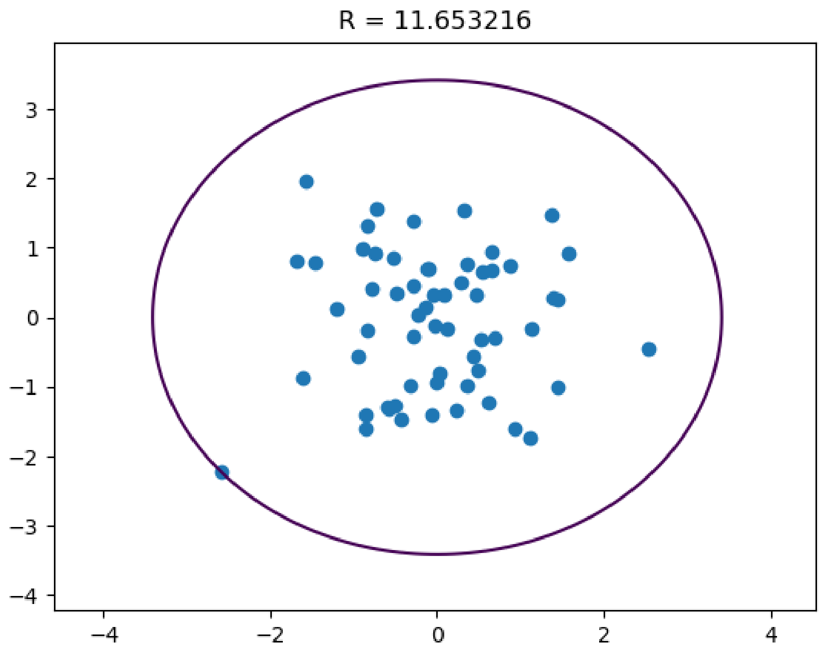

Рис. 2: $\rho = 0.0, n = 60$

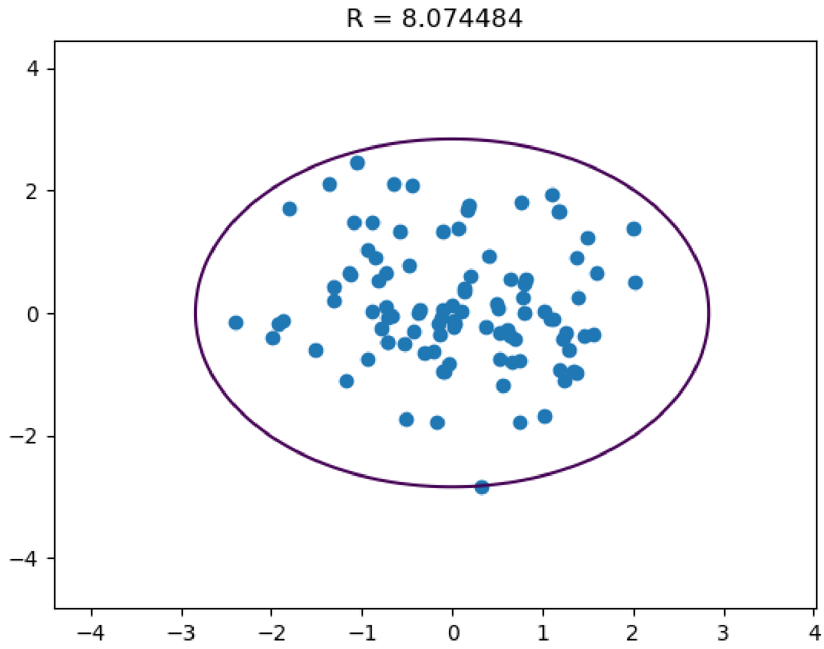

Рис. 3: $\rho = 0.0, n = 100$

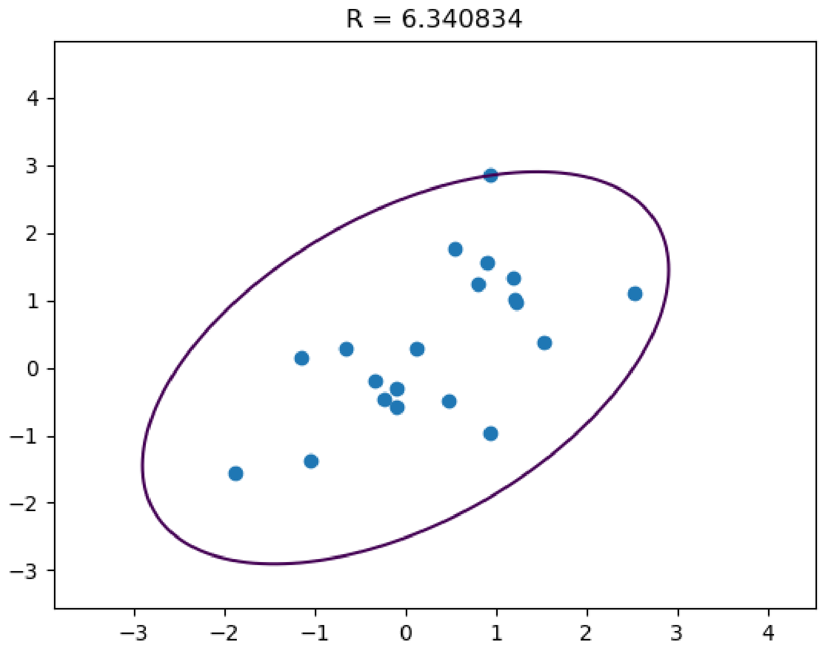

Рис. 4: $\rho = 0.5, n = 20$

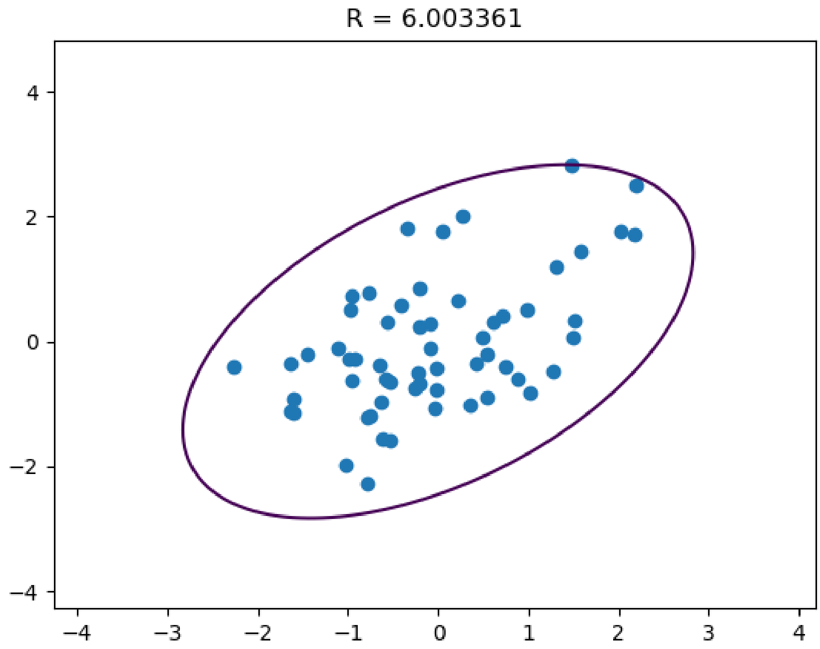

Рис. 5: $\rho = 0.5, n = 60$

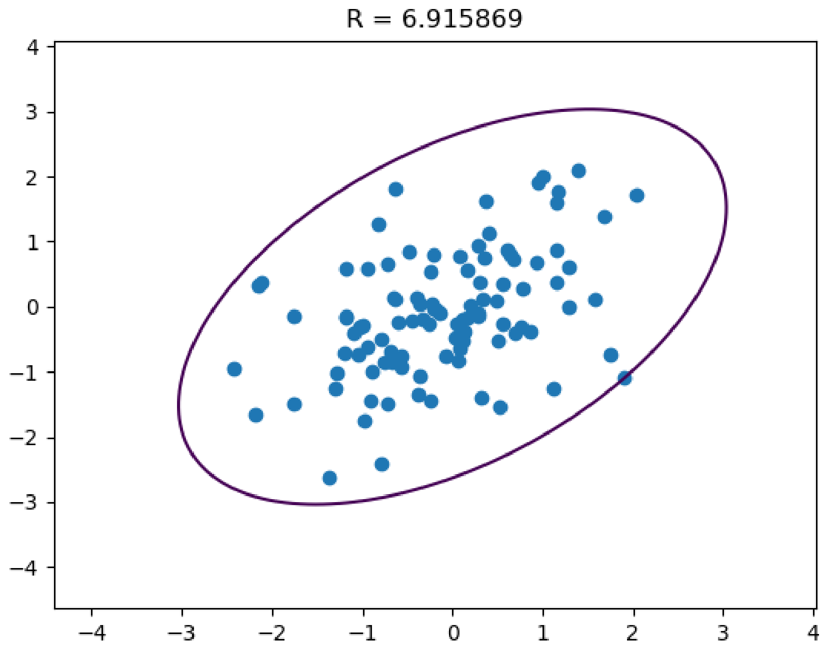

Рис. 6: $\rho = 0.5, n = 100$

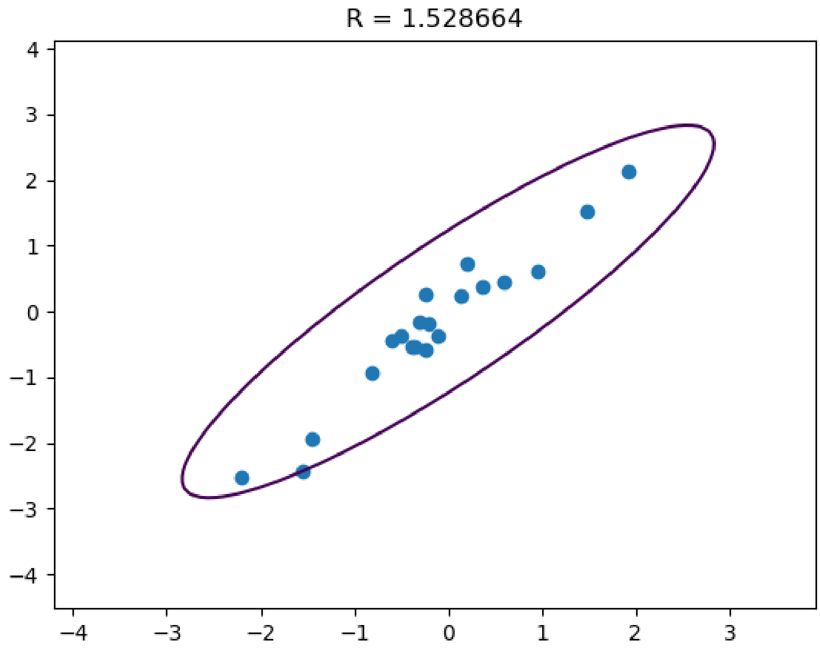

Рис. 7: $\rho = 0.9, n = 20$

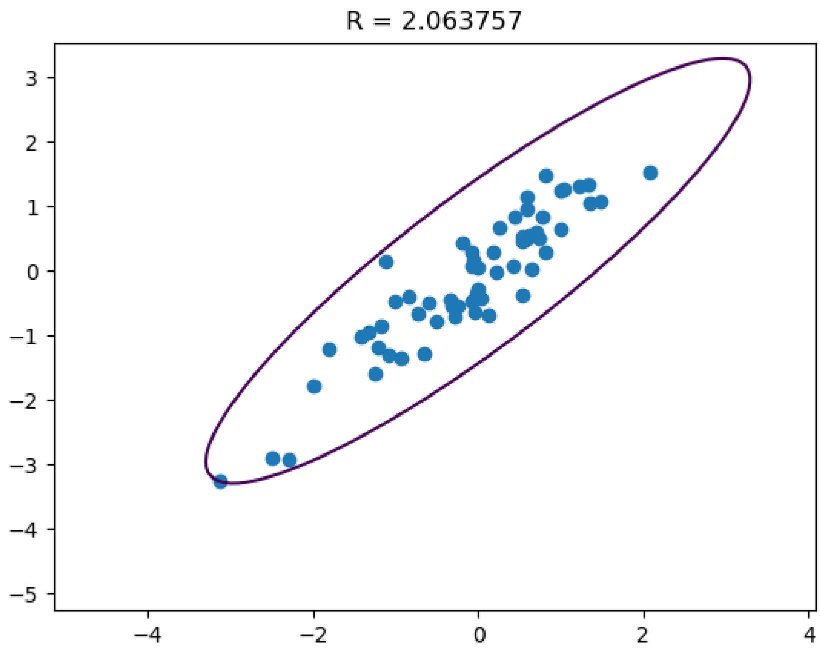

Рис. 8: $\rho = 0.9, n = 60$

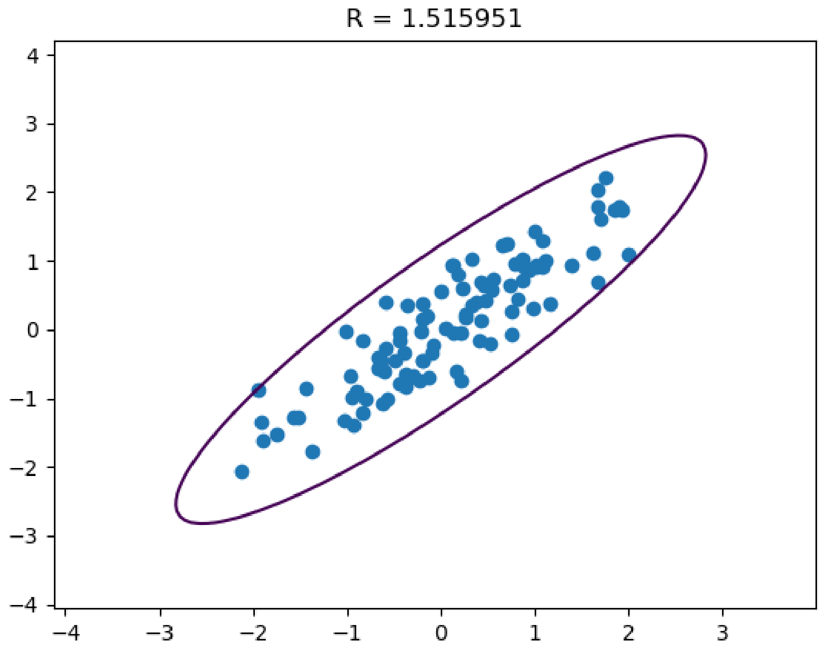

Рис. 9: $\rho = 0.9, n = 100$

4.3 Выборка без выбросов

- Критерий наименьших квадратов:

$$\widehat{\beta}_0 = 2.47 \quad \widehat{\beta}_1 = 1.95 \quad Q(13) = 13.9637 \quad M(15) = 13.9182$$

- Критерий наименьших модулей:

$$\widehat{\beta}_0 = 2.49 \quad \widehat{\beta}_1 = 1.68 \quad Q = 15.9356 \quad M = 13.3737$$

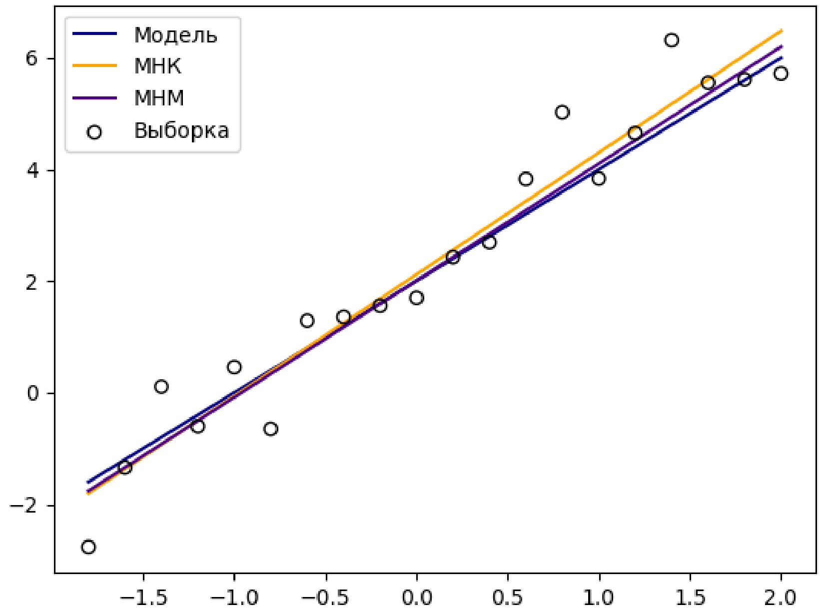

Рис. 10: Без выбросов

4.4 Выборка с выбросами

- Критерий наименьших квадратов:

$$\widehat{\beta}_0 = 2.61 \quad \widehat{\beta}_1 = 0.52 \quad Q = 154.2302 \quad M = 37.381$$

- Критерий наименьших модулей:

$$\widehat{\beta}_0 = 2.67 \quad \widehat{\beta}_1 = 1.35 \quad Q = 172.7536 \quad M = 29.9906$$

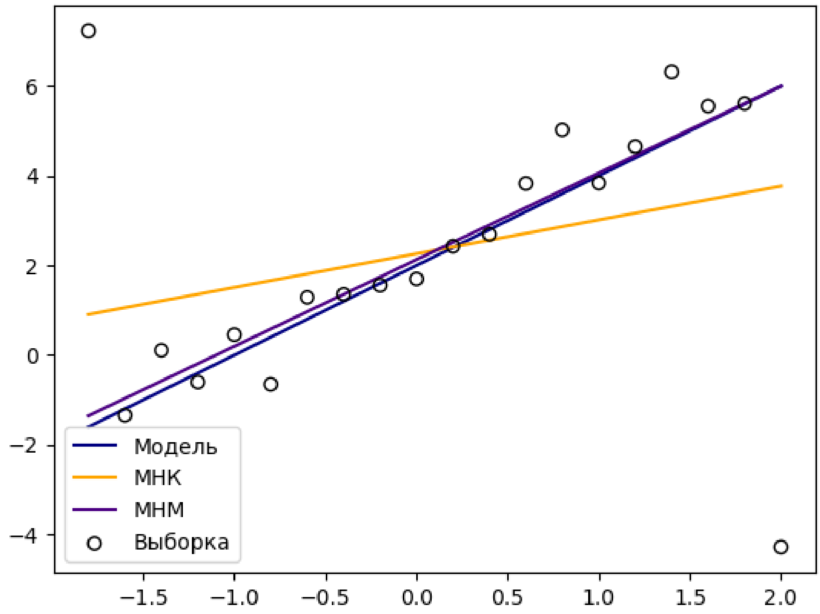

Рис. 11: С выбросами

4.5 Критерий согласия χ^2

Оценки:

$$\widehat{\mu} = 0.03 \quad \widehat{\sigma} = 1.01 \tag{34}$$

Число промежутков: $k = \lceil\ 1.72 \cdot \sqrt[3]{100}\ \rceil = 8$

Таблица вычислений χ^2:

i	Δ_i	n_i	p_i	$n_i - np_i$
1	$(-\infty, -1.79]$	2	0.0366	-1.6609
2	$(-1.79, -1.27]$	7	0.0637	0.627
3	$(-1.27, -0.75]$	12	0.121	-0.0972
4	$(-0.75, -0.23]$	22	0.1777	4.2306
5	$(-0.23, 0.29]$	22	0.202	1.8009
6	$(0.29, 0.81]$	15	0.1777	-2.7694
7	$(0.81, 1.33]$	8	0.121	-4.0972
8	$(1.33, +\infty)$	12	0.1003	1.9661

Таблица 5: Таблица вычислений χ^2

При $\alpha = 0.05: \ \chi^2_{1-\alpha}(k-1) \approx 14.0671$, а вычисленное $\chi^2_{\text{B}} = 4.1883$, видно что $\chi^2_{\text{B}} < \chi^2_{1-\alpha}(k-1)$

В результате доп. исследования, было получено что при $n = 20$ критерий дает вывод что генеральное распределение является нормальным $N(0.024, 0.59)$, в результате вычислений $\chi^2_B = 4.8612 < 4.8784 = \chi^2_{1-\alpha}$, а при $n = 100$ уже $\chi^2_B = 19.2086 \geq 8.3834 = \chi^2_{1-\alpha}$ т.е. установлено что генеральное распределение не является нормальным (и это соответствует тому что оно задано как равномерное)

4.6 Классические оценки

	$m(28)$	$\sigma(29)$
$n = 20$	$-0.6 < m < 0.27$	$0.71 < \sigma < 1.36$
$n = 100$	$-0.04 < m < 0.34$	$0.84 < \sigma < 1.12$

4.7 Асимптотически нормальные оценки

	$m(30)$	$\sigma(31)$
$n = 20$	$-0.56 < m < 0.23$	$0.71 < \sigma < 1.46$
$n = 100$	$-0.03 < m < 0.34$	$0.84 < \sigma < 1.13$

5 Обсуждение

5.1 Коэффициенты корреляции

Для начала воспользуемся (5) для анализа экспериментов по которым были получены таблицы 1, 2. Выясним можно ли принять гипотезу о зависимости между случайными величинами на уровне значимости $\alpha = 0.05$ для $n = 100$ по коэффициенту Пирсона.

$$0\sqrt{100-1} \leq 2.5, 4.98 \approx 0.5\sqrt{100-1} > \cdot 2.5$$

В эксперименте 1 эту гипотезу принять нельзя, а в эксперименте 2 можно. При этом в эксперименте 1 с.в. заведомо независимы, а в эксперименте 2 зависимы, так что все согласуется с теорией.

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 (робастная) оценка. Квадрантный коэффициент корреляции показывает лучшие результаты в устойчивости.

5.2 Эллипсы равновероятности

Видно что чем ближе ρ к 1, тем эллипс равновероятности становится все больше похож на прямую, заданную как (10). Т.е. наглядно показано как между с.в. X и Y возникает линейная зависимость.

5.3 Линейная регрессия

Из графиков видно, что оценка по критерию наименьших модулей значительно лучше приближает эталонную зависимость при наличии выбросов и это согласуется с теорией т.к. она является робастной. В тоже время, критерий наименьших квадратов дает более точное приближение в отсутствие выбросов и, к тому же, проще для вычислений. Полученные значения M, Q упорядочены как и ожидалось, для оценки МНК значение Q меньше, чем для любой другой, аналогично для оценки МНМ и значения M

5.4 Критерий согласия χ^2

Согласно результатам эксперимента, заданное по оценкам (34) распределение $N(\widehat{\mu}, \widehat{\sigma})$ является генеральным законом по которому построена выборка с уровнем значимости 0.05. Теоретически это обосновывается тем что оценки максимального правдоподобия состоятельны. Было установлено что при небольших объемах выборки уверенности в полученных результатах нет, ведь статистика критерия χ^2 лишь асимптотически распределена по закону $\chi^2(k-1)$ т.е. n предполагается достаточно большим.

5.5 Интервальное оценивание

Полученные интервальные оценки говорят о том что с вероятностью 0.95 значения $m = 0$ и $\sigma = 1$ лежат в соответствующих интервалах. По постановке эксперимента, интервалы действительно накрывают истинные значения параметров. Следует заметить что при большом объеме n выборки - асимптотические оценки практически совпадают с классическими.

6 Приложения

1. Исходный код лабораторной 5 `https://github.com/zhenyatos/statlabs/tree/master/Lab5`
2. Исходный код лабораторной 6 `https://github.com/zhenyatos/statlabs/tree/master/Lab6`
3. Исходный код лабораторной 7 `https://github.com/zhenyatos/statlabs/tree/master/Lab7`
4. Исходный код лабораторной 8 `https://github.com/zhenyatos/statlabs/tree/master/Lab8`

Список литературы

[1] **Вероятностные разделы математики.** Учебник для бакалавров технических направлений. // Под ред. Максимова Ю.Д. - СПб «Иван Федоров», 2001. - 592 с., илл

[2] Вентцель Е.С. *Теория вероятностей: Учеб. для вузов.* — 6-е изд. стер. — М.: Высш. шк., 1999.— 576 с.

[3] Метод Нелдера — Мида // Википедия. [2019—2019]. Дата обновления: 11.09.2019. URL: https://ru.wikipedia.org/?oldid=102111276 (дата обращения: 11.09.2019).

[4] Wikipedia contributors. (2020, March 19). Histogram. In Wikipedia, The Free Encyclopedia. Retrieved 18:27, May 14, 2020, from `https://en.wikipedia.org/w/index.php?title=Histogram&oldid=946321806`