Санкт-Петербургский политехнический университет имени
Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики и физики

Математическая статистика
Отчёт по лабораторным работам №5-8

Выполнил:

Студент: Парусов Владимир
Группа: 5030102/90201

Принял:

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

2022 г.

Содержание

Список иллюстраций

Список таблиц

1. Постановка задачи

1. Дано двумерное нормальное распределение $N(x, y, 0, 0, 1, 1, \rho)$.

   Требуется сгенерировать двумерные выборки размером 20, 60, 100 элементов и коэффициентами корреляции 0, 0.5, 0.9.

   Каждую выборку необходимо сгенерировать 1000 раз и вычислить для неё:

   - Среднее значение
   - Среднее значение квадрата
   - Дисперсию

   коэффициентов корреляции:

   - Пирсона
   - Спирмена
   - Квадрантного коэффициента корреляции

   Также требуется произвести эти 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 эллипсе равновероятности.

2. Требуется найти оценки коэффициентов линейной регрессии $y_i = a + bx_i + e_i$, используя 20 равномерно распределённых на отрезке $[-1.8; 2.0]$ точек. Ошибку e_i считаем нормально распределённой с матожиданием 0 и дисперсией 1. В качестве эталонной зависимости взять $y_i = 2 + 2x_i + e_i$.

   Оценку требуется произвести по двум критериям:

   - Критерий наименьших квадратов
   - Критерий наименьших модулей

   То же самое требуется проделать для выборки, у которой в крайние значения вносятся возмущения 10 и -10 соответственно.

3. Требуется сгенерировать выборку размером 100 элементов для нормального распределения $N(x, 0, 1)$. По сгенерированной выборке требуется определить параметры распределения μ, δ методом максимального правдоподобия.

   В качестве основной гипотезы H_0 положим, что сгенерированное распределение является нормальным с неизвестными матожиданием $\hat{\mu}$ и дисперсией $\hat{\sigma}$.

Требуется проверить основную гипотезу, используя критерий согласия χ^2. В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений χ^2.

То же самое требуется выполнить для выборки распределения Лапласа $L(0, \frac{1}{\sqrt{2}}, x)$, состоящей из 20 элементов.

4. Требуется для двух выборок размерами 20 и 100 элементов, сгенерированных по нормальному распределению $N(x, 0, 1)$ для параметров положения и масштаба построить:

   - асимптотически нормальные интервальные оценки на основе точечных оценок метода максимального правдоподобия
   - классические интервальные оценки на основе статистик χ^2 и Стьюдента.

   В качестве параметра надёжности взять $\gamma = 0.95$.

2. Теория

2.1. Двумерное нормальное распределение

Двумерная случайная величина называется нормально распределённой, если её плотность вероятности определена следующим образом:

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

При этом компоненты X и Y также распределены нормально со средним квадратичным отклонением соответственно σ_x и σ_y.

2.2. Корреляционный момент и коэффициент корреляции

Ковариацией или *корреляционным моментом* случайной величины называется матожидание произведения отклонений компонент случайной величины от её среднего:

$$K_{XY} = cov(X, Y) = M[(X - \overline{x})(Y - \overline{y})] \quad (2)$$

Коэффициентом корреляции является нормированный на единицу корреляционный момент. Показывает меру линейной зависимости между величинами.

$$\rho = \frac{K_{XY}}{\sigma_x\sigma_y} \quad (3)$$

2.3. Выборочные коэффициенты корреляции

2.3.1. Выборочный коэффициент корреляции Пирсона

Для выборки двумерной случайной величины $\{x_i, y_i\}_{i=\overline{1,n}}$ наиболее естественным приближением корреляционного коэффициента является соотношение:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}\frac{1}{n}(x_i - \overline{x})^2 \frac{1}{n}(y_i - \overline{y})^2}} = \frac{K_{XY}}{s_X s_Y}, \tag{4}$$

где K_{XY}, s_X^2, s_Y^2 – выборочная ковариация и дисперсии соответствующих случайных величин.

2.3.2. Выборочный квадрантный коэффициент корреляции

Альтернативным способом определения взаимосвязи является выборочный квадрантный коэффициент корреляции:

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \tag{5}$$

где n_i – количество точек выборки в соответствующим квадранте координатной плоскости, параллельно перенесённой относительно стандартной на вектор $(medx, medy)$.

2.3.3. Выборочный коэффициент ранговой корреляции Спирмена

Как правило, изучаемые на практике объекты обладают некоторым набором качественных признаков, то есть не ассоциированных с конкретными числовыми значениями, но позволяющих задать на множестве объектов отношение полного порядка.

Задав отношение полного порядка, мы тем самым присвоили объектам номера, иначе говоря, проранжировали объекты.

Каждый признак задаёт своё отношение порядка, соответственно, для двух признаков мы будем иметь две последовательности рангов – $\{u_i\}_{i\in\mathbb{N}}$ и $\{v_i\}_{i\in\mathbb{N}}$.

Коэффициентом ранговой корреляции мы будем называть коэффициент корреляции Пирсона для двумерной выборки (u_i, v_i):

$$r_S = \frac{K_{UV}}{s_U s_V} \tag{6}$$

2.4. Эллипсы рассеивания

Рассмотрим линии уровня плотности вероятности. Они удовлетворяют условию:

$$\frac{(x-\overline{x})^2}{\sigma_x^2} - 2\rho(x-\overline{x})(y-\overline{y}) + \frac{(y-\overline{y})^2}{\sigma_y^2} = k^2 \tag{7}$$

Как мы знаем из аналитической геометрии, это уравнение семейства концентрических эллипсов, где параметр k определяет размер эллипса.

Данное уравнение представляет из себя квадратичную форму, приведя которую к каноническому виду и отнормировав на константу, стоящую справа от знака равенства, мы получим каноническое уравнение эллипса.

Матрица преобразований квадратичной формы подскажет нам расположение системы координат, в которой фокусы эллипса будут лежать на абсциссе и, соответственно, смещение и угол поворота этой системы координат относительно старой. Таким образом, мы узнаем, как направлены полуоси данного эллипса относительно исходной системы координат.

Произведя соответствующие выкладки, получаем, что большая полуось эллипса составляет с абсциссой угол, выраженный по следующей формуле:

$$tg2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \tag{8}$$

Посмотрев на данное соотношение, можно понять, что угол поворота эллипса даёт нам качественное представление о степени коррелированности данных.

Контур эллипса является линией равной вероятности, поэтому такой эллипс называется эллипсом равной плотности либо эллипсом рассеивания.

В результатах рассматриваются "полные"эллипсы рассеивания, в которые с вероятностью 0.99 укладывается всё рассеивание. Для таких эллипсов константа k выбирается равной 3.

3. Простая линейная регрессия

3.1. Описание модели

Регрессионную модель описания данных называют простой линейной, если заданный набор данных аппроксимируется прямой с внесённой добавкой в виде некоторой нормально распределённой ошибки:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i \in \overline{1,n} \tag{9}$$

где
$\{x_n\}_{n\in\mathbb{N}}$ – заданные значения,
$\{y_n\}_{n\in\mathbb{N}}$ – параметры отклика,

$\{\varepsilon_n\}_{n\in\mathbb{N}}$ – независимые, центрированные, нормально распределённые случайные величины с неизвестной дисперсией δ, суть предполагаемые погрешности,

β_0, β_1 – параметры, подлежащие оцениванию.

В данной модели мы считаем, что у заданных значений нет погрешности (пренебрегаем ей). Полагаем, что основная погрешность получается при измерении $\{y_n\}_{n\in\mathbb{N}}$.

3.2. Метод наименьших квадратов

Данный метод, вообще говоря, может применяться для аппроксимации заданного набора экспериментальных данных линейной комбинацией линейно независимых функций, размер которой не превосходит мощности множества данных (в случае равенства получаем интерполяцию).

Критерием оптимальности подобранной аппроксимации является l^2-норма, точнее, для простоты вычисления, её квадрат:

$$\| \sum_{i=1}^{m} \lambda_i f_i(\{x_n\}) - \{y_n\} \|_{l^2}^2 \underset{\{\lambda_i\}}{\longrightarrow} min \tag{10}$$

Минимум ищется по коэффициентам линейной комбинации, исходя из критерия равенства нулю градиента и положительной определённости якобиана.

Для того, чтобы градиент был равен нулю, необходимо решить СЛАУ с матрицей:

$$\begin{pmatrix} [f_1(\{x_n\}), f_1(\{x_n\})] & \cdots & [f_1(\{x_n\}), f_m(\{x_n\})] \\ \vdots & \ddots & \vdots \\ [f_m(\{x_n\}), f_1(\{x_n\})] & \cdots & [f_m(\{x_n\}), f_m(\{x_n\})] \end{pmatrix} \tag{11}$$

и правым столбцом:

$$\begin{pmatrix} [f_1(\{x_n\}), \{y_n\}] \\ \vdots \\ [f_m(\{x_n\}), \{y_n\}] \end{pmatrix} \tag{12}$$

где $[f, g]$ – стандартное скалярное произведение векторов в ортнормированном базисе в $\mathbb{R}^n$.

Матрица скалярных произведений является матрицей Грама, а значит она невырожденная. Якобиан здесь также является положительной полуопределённой матрицей, значит найденное решение будет точкой минимума.

В нашем случае мы аппроксимируем одной линейной функцией $y(x) = \beta_0 + \beta_1 x$, а получаемый в результате решения задачи минимизации (10) результат будет являться суммой квадратов ε_i, иными словами, нормировочным множителем для распределения $\{\varepsilon_i\}$.

Для данной задачи имеем:

$$\begin{cases} \hat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \\ \hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1 \end{cases} \tag{13}$$

4. Метод наименьших модулей

Данный метод основан на минимизации l^1-нормы разности последовательностей полученных экспериментальных данных $\{y_n\}$ и значений аппроксимирующей функции $f(\{x_n\})$. Увы, автору данного отчёта неизвестно метода, позволяющего решить, как в случае МНК, данную задачу минимизации для линейной комбинации заданного количества базисных функций, действующих на $\mathbb{R}$, однако метод позволяет решать задачу для линейной функции любой размерности:

$$\|[\boldsymbol{a}, \{x_n\}] - \{y_n\}\|_{l^1} \underset{\{\lambda_i\}}{\longrightarrow} min \tag{14}$$

Данную задачу минимизации можно решать точно, например, используя алгоритм спуска по узловым направлениям. Метод основан на теореме о том, что точка минимума искомой функции лежит в одной из точек нарушения дифференцируемости минимизируемой функции (в точке, где какой-либо модуль обращается в ноль), заданного данными и реализует направленный перебор всех таких точек [2].

Кроме того, можно решать численно, методом Вейсфельда [3]. Суть метода в том, что вместо решения негладкой задачи мы на каждой итерации минимизируем взвешенную l^2-норму разности, где вес равен величине, обратной невязке на предыдущем шаге (таким образом, мы делим квадрат невязки на текущем шаге на невязку на предыдущем, и получаем "почти невязку" в первой степени, что соответствует l^1-норме).

Нетрудно понять, что такую задачу можно решать с помощью МНК – отличие лишь в том, что скалярные произведения в соответствующей матрице (11) будут взвешенными.

5. Метод максимального правдоподобия

Пусть $\{x_n\}$ – случайная выборка из генеральной совокупности с плотностью вероятности $f(x, \theta)$, где $\theta \in \mathbb{R}^m, m \in \mathbb{N}$ – совокупность параметров плотности вероятности.

Определение 1. Функцией правдоподобия назовём совместную плотность вероятность независимых случайных величин $x_1, ..., x_n$ с *одним и тем же* параметром распределения θ:

$$L(x_1, ..., x_n, \theta) = f(x_1, \theta) \cdot ... \cdot f(x_n, \theta) \tag{15}$$

Определение 2. Оценкой максимального правдоподобия $\hat{\theta}$ назовём такое значение параметра, при котором функция правдоподобия достигает своего максимума.

Замечание 1. L – функция **одной переменной** θ. Будем далее иметь в виду, что любые операции дифференцирования ведутся исключительно по θ.

Если L дважды дифференцируема, то достаточным условием максимума будет равенство нулю градиента и отрицательная определённость якобиана.

Также можно рассматривать не саму функцию правдоподобия, а её натуральный логарифм, поскольку применение внешней монотонной функции не изменяет точек экстремума. Например, это удобно, когда мы имеем дело с нормальным распределением, так как в нём фигурирует экспоненциальная зависимость, которая может быть линеаризована посредством применения натурального логарифма.

6. Проверка гипотезы о законе генеральной совокупности. Метод χ^2

Для проверки гипотезы о функции распределения часто используется критерий согласия χ^2.

Для одномерного случая, когда функция распределения не содержит неизвестных параметров, методика следующая.

Сделаем разбиение вещественной оси на k полуинтервалов Δ_i. Обозначим через n_i количество событий, попавших в интервал Δ_i.

Для нормального распределения получаем, что выборочное среднее – о.м.п. матожидания, а выборочная дисперсия – о.м.п. генеральной дисперсии.

Если гипотеза справедлива, то относительные частоты должны быть близки к вероятности. Проверка такой близости производится по l^2-норме с весами $c_i = \frac{n}{p_i}$:

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

Данное обозначение критерия близости, называемого *статистикой критерия* χ^2 неслучайно, поскольку имеет место следующая

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

То есть, какую бы мы гипотезу ни проверяли, функция распределения статистики стремится к истинной функции распределения случайной величины с плотностью вероятности:

$$f_{k-1}(x) = \begin{cases} 0, & x \le 0 \\ \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}, & x > 0 \end{cases} \tag{17}$$

Также необходимо понять, что какую гипотезу мы будем считать достоверной, а какую – нет. Для этого необходимо ввести *уровень значимости* α. С его помощью проверка будем производиться следующим образом:

- Если $\chi^2_{\text{В}} < \chi^2_{1-\alpha}(k-1)$, то гипотеза принимается
- Иначе гипотеза считается несостоятельной

7. Доверительные оценки для параметров нормального распределения

7.1. Доверительный интервал для матожидания m нормального распределения

Для выборки $(x_1, ..., x_n)$ из нормальной генеральной совокупности найдём среднее $\overline{x}$ и среднее квадратичное отклонение s.

Тогда величина

$$T = \sqrt{n-1} \cdot \frac{\overline{x} - m}{s}, \tag{18}$$

называемая статистикой Стьюдента, распределена по закону Стьюдента с $n-1$ степенями свободы.

Произведя несложные преобразования, получим, что:

$$P\left(-x < T < x\right) = 2F_T(x) - 1, \tag{19}$$

где F_T – функция распределения Стьюдента с $n-1$ степенями свободы.

Полагая $2F_T(x) - 1 = 1 - \alpha$, где α – уровень значимости, имеем:

$$P\left(\overline{x} - \frac{sx}{\sqrt{n-1}} < m < \overline{x} + \frac{sx}{\sqrt{n-1}}\right) = \\ = P\left(\overline{x} - \frac{st_{1-\frac{\alpha}{2}}(n-1)}{\sqrt{n-1}} < m < \overline{x} + \frac{st_{1-\frac{\alpha}{2}}(n-1)}{\sqrt{n-1}}\right) = 1 - \alpha \tag{20}$$

И таким образом получаем доверительный интервал для матожидания с вероятностью $1 - \alpha$.

7.2. Доверительный интервал для среднего квадратического отклонения σ нормального распределения

Доказано, что случайная величина $\frac{ns^2}{\sigma^2}$ распределена по закону χ^2 с $n-1$ степенями свободы.

После ряда преобразований, получаем:

$$P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right) \tag{21}$$

8. Доверительные оценки для параметров произвольного распределения. Асимптотический подход

8.1. Доверительные оценки для матожидания при большом размере выборки

Если исследуемое распределение имеет конечное матожидание и дисперсию, то имеет место центральная предельная теорема:

$$\frac{\overline{x} - \mathbf{M}x}{\sqrt{\mathbf{Dx}}} = \sqrt{n} \cdot \frac{\overline{x} - m}{\sigma} \xrightarrow{F} N(x, 0, 1) \tag{22}$$

Отсюда получаем, что

$$P\left(-x < \sqrt{n} \cdot \frac{\overline{x} - m}{\sigma} < x\right) \approx 2\Phi(x),$$

где $\Phi(x)$ – функция Лапласа.

Полагая $u_{1-\alpha/2}$ за соответствующий квантиль центрированного нормального распределения с единичной дисперсией, получаем:

$$P\left(\overline{x} - \frac{su_{1-\alpha/2}}{\sqrt{n}} < m < \overline{x} + \frac{su_{1-\alpha/2}}{\sqrt{n}}\right) \approx \gamma, \tag{23}$$

что и даёт доверительный интервал для матожидания m с доверительной вероятностью γ.

8.2. Доверительные оценки для дисперсии при большом размере выборки

Используя ЦПТ и разложение в ряд Тейлора для характеристической функции Лапласа, получим, что:

$$s(1+U)^{-1/2} < \sigma < s(1-U)^{-1/2}, \tag{24}$$

где $U = u_{1-\alpha/2}\sqrt{\frac{e+2}{n}}$

9. Реализация

Данная работа реализована на языке программирования Python с использованием редактора VIM и библиотек NumPy, MatPlotLib, Statsmodels, Scipy в ОС Ubuntu 19.04.

Отчёт подготовлен с помощью компилятора pdflatex и среды разработки TeXStudio.

10. Результаты

10.1. Выборочные коэффициенты корреляции

$\rho = 0$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.0	0.0	0.0
$E(z^2)$	0.0	0.0	0.0
$D(z)$	0.2	0.2	0.2
$\rho = 0.5$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.5	0.5	0.3
$E(z^2)$	0.3	0.2	0.2
$D(z)$	0.2	0.2	0.2
$\rho = 0.9$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.90	0.87	0.7
$E(z^2)$	0.81	0.76	0.5
$D(z)$	0.05	0.07	0.2

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

$\rho = 0$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.0	0.0	0.00
$E(z^2)$	0.0	0.0	0.01
$D(z)$	0.1	0.1	0.12
$\rho = 0.5$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.50	0.5	0.33
$E(z^2)$	0.26	0.2	0.13
$D(z)$	0.10	0.1	0.13
$\rho = 0.9$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.90	0.88	0.71
$E(z^2)$	0.81	0.78	0.52
$D(z)$	0.03	0.03	0.09

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

$\rho = 0$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.00	0.00	0.00
$E(z^2)$	0.0	0.01	0.01
$D(z)$	0.1	0.10	0.10
$\rho = 0.5$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.50	0.48	0.34
$E(z^2)$	0.26	0.24	0.12
$D(z)$	0.08	0.08	0.09
$\rho = 0.9$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.90	0.89	0.71
$E(z^2)$	0.81	0.79	0.51
$D(z)$	0.02	0.02	0.07

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

$n = 20$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.80	0.76	0.60
$E(z^2)$	0.65	0.60	0.39
$D(z)$	0.09	0.11	0.18
$n = 60$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.80	0.79	0.60
$E(z^2)$	0.66	0.62	0.37
$D(z)$	0.05	0.05	0.11
$n = 100$ (3)	r (4)	r_Q (5)	r_S (6)
$E(z)$	0.81	0.80	0.60
$E(z^2)$	0.66	0.63	0.37
$D(z)$	0.03	0.04	0.08

Таблица 4. Смесь нормальных распределений

10.2. Эллипсы рассеивания

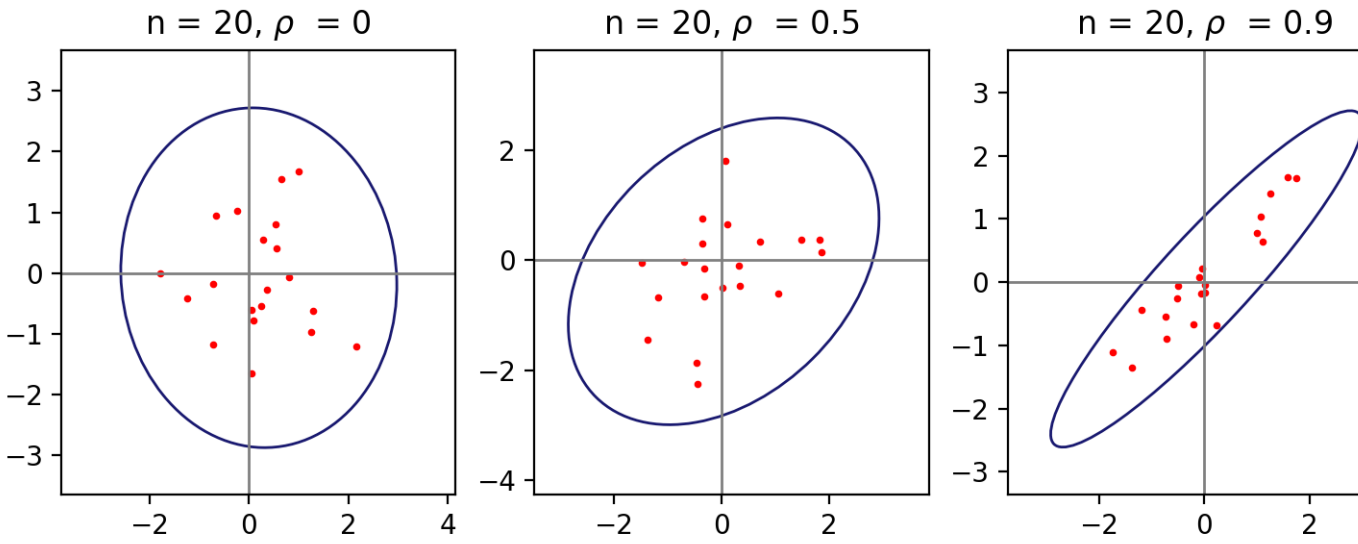

Рис. 1. Двумерное нормальное распределение, $n = 20$

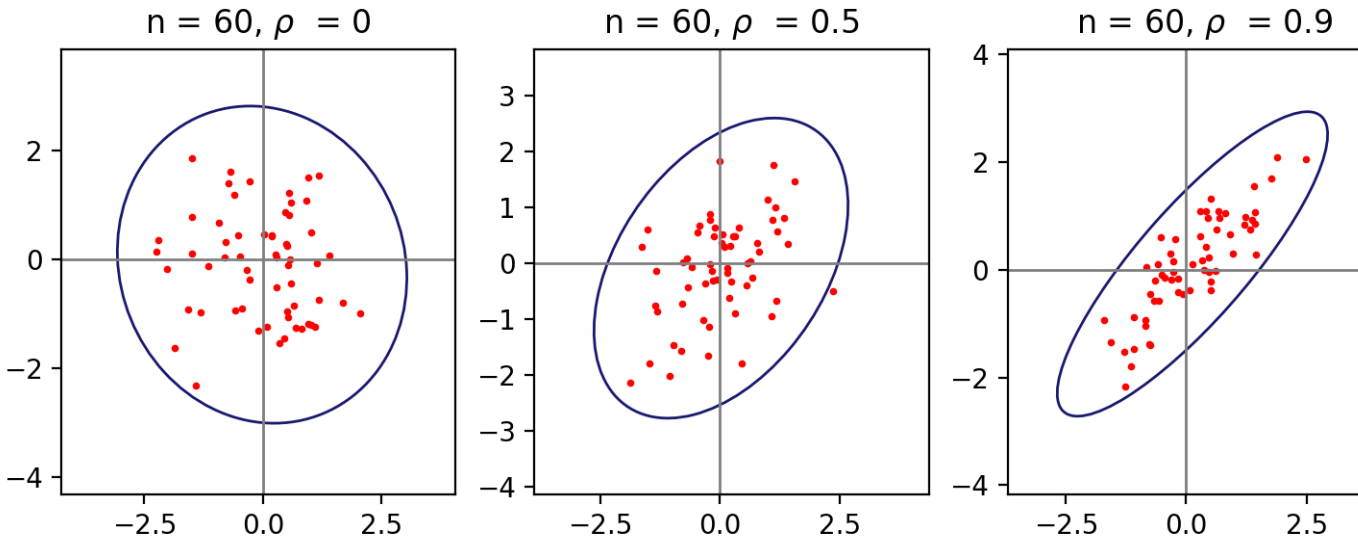

Рис. 2. Двумерное нормальное распределение, $n = 60$

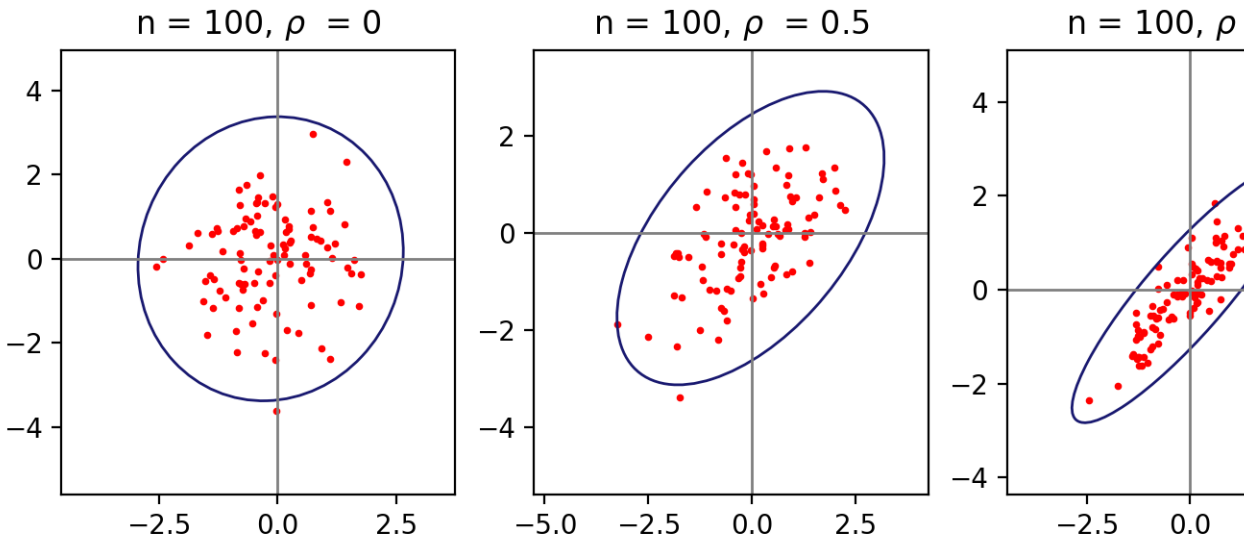

Рис. 3. Двумерное нормальное распределение, $n = 100$

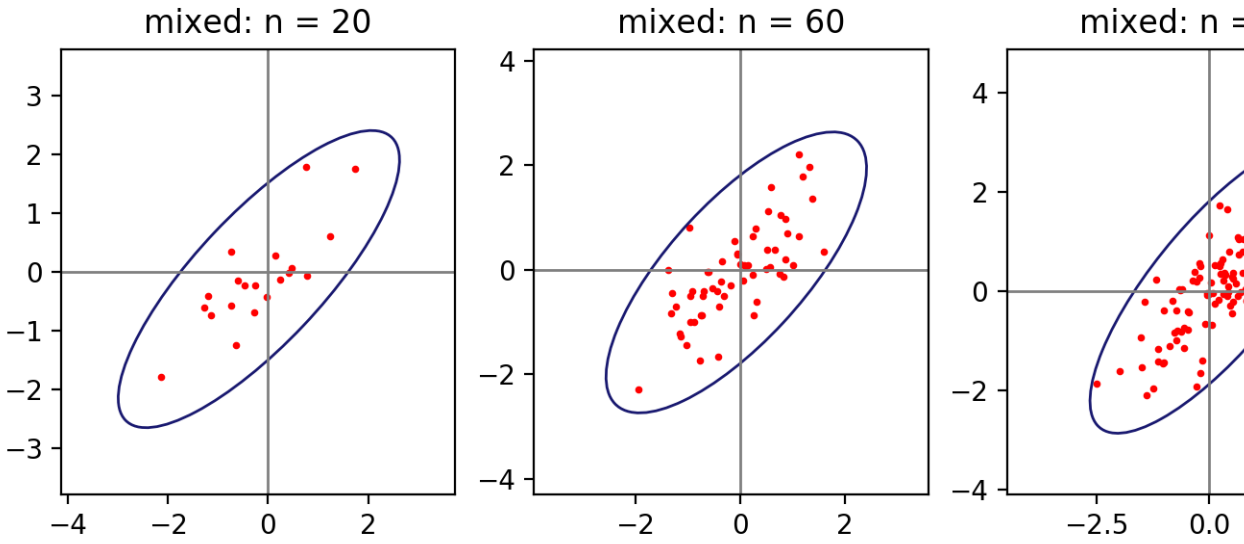

Рис. 4. Смесь нормальных распределений

10.3. Оценки коэффициентов линейной регрессии

10.3.1. Выборка без возмущений

- Критерий наименьших квадратов:

$$\hat{a} = 2.23 \quad \hat{b} = 2.05$$

- Критерий наименьших модулей:

$$\hat{a} = 2.45 \quad \hat{b} = 2.12$$

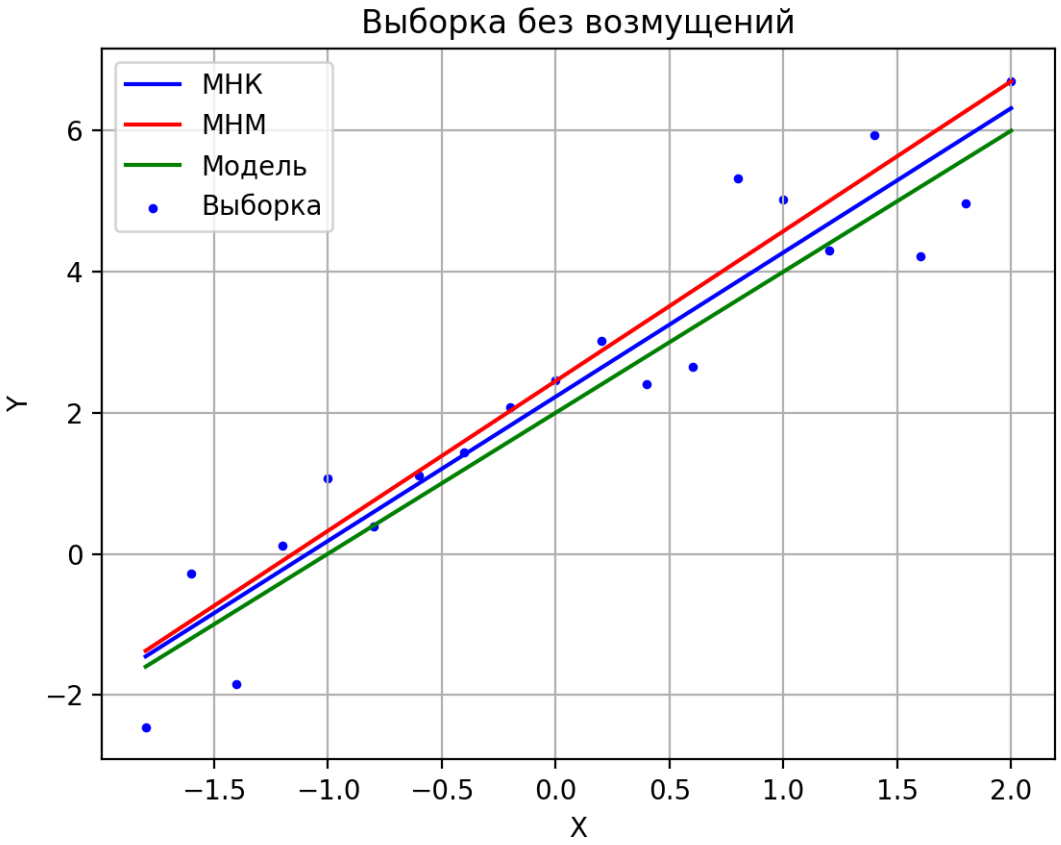

Рис. 5. Выборка без возмущений

10.3.2. Выборка с возмущениями

- Критерий наименьших квадратов:

$$\hat{a} = 2.37 \quad \hat{b} = 0.62$$

- Критерий наименьших модулей:

$$\hat{a} = 2.19 \quad \hat{b} = 1.54$$

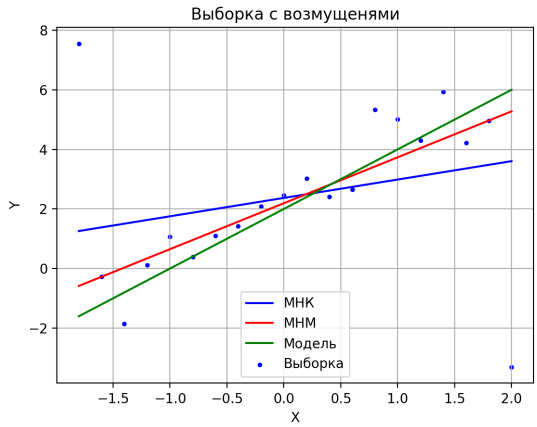

Рис. 6. Выборка с возмущениями

10.4. Проверка гипотезы о законе генеральной совокупности. Метод χ^2

10.4.1. Метод максимального правдоподобия

Для сгенерированной выборки были получены результаты:

$$\hat{\mu} = 0.12, \hat{\sigma} = 0.97$$

10.4.2. Критерий согласия χ^2

i	Границы Δ_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
1	-1.69, -0.49	0.00	0.0008	0.08	-0.08	0.08
2	-0.49, 0.71	5.00	0.0282	2.82	2.18	1.67
3	0.71, 1.91	17.00	0.2348	23.48	-6.48	1.79
4	1.91, 3.11	54.00	0.4724	47.24	6.76	0.97
5	3.11, -2.89	22.00	0.2348	23.48	-1.48	0.09
6	-2.89, -1.69	2.00	0.0282	2.82	-0.82	0.24
7	-1.69, -0.49	0.00	0.0008	0.08	-0.08	0.08
Σ		100	1.0000	100.00	0.00	$\chi^2_{\text{B}} = 2.34$

Таблица 5. Вычисление χ^2_{B} при проверке закона о нормальном распределении для выборки нормального распределения

Видим, что $\chi^2_{\text{B}} < \chi^2_{0.95} \approx 14.1$, следовательно, гипотезу полагаем верной.

i	Границы Δ_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
1	-1.60, -0.40	0.00	0.0000	0.00	-0.00	0.00
2	-0.40, 0.80	0.00	0.0007	0.01	-0.01	0.01
3	0.80, 2.00	2.00	0.1434	2.87	-0.87	0.26
4	2.00, 3.20	15.00	0.7117	14.23	0.77	0.04
5	3.20, -2.80	3.00	0.1434	2.87	0.13	0.01
6	-2.80, -1.60	0.00	0.0007	0.01	-0.01	0.01
7	-1.60, -0.40	0.00	0.0000	0.00	-0.00	0.00
Σ		20	1.0000	100.00	0.00	$\chi^2_{\text{B}} = 13.92$

Таблица 6. Вычисление χ^2_{B} при проверке закона о нормальном распределении для выборки распределения Лапласа

Получили, что $\chi^2_{\text{B}} < \chi^2_{0.95} \approx 14.1$, следовательно гипотеза о нормальности принимается.

10.5. Доверительные интервалы матожидания и дисперсии для нормального распределения

$n = 20$	m	σ
	$-0.20 < m < 0.69$	$0.72 < \sigma < 1.39$
$n = 100$	m	σ
	$-0.20 < m < 0.25$	$1.00 < \sigma < 1.33$

Таблица 7. Доверительные интервалы для параметров нормального распределения

10.6. Доверительные интервалы для параметров произвольного распределения. Асимптотический подход

$n = 20$	m	σ
	$-0.16 < m < 0.65$	$0.67 < \sigma < 2.98$
$n = 100$	m	σ
	$-0.20 < m < 0.25$	$0.95 < \sigma < 1.53$

Таблица 8. Доверительные интервалы для параметров произвольного распределения

11. Обсуждение

11.1. Выборочные коэффициенты корреляции

Выборочные коэффициенты корреляции Пирсона и Спирмена достаточно точно описывают истинный коэффициент корреляции двумерной случайной величины: он всегда попадает в доверительный интервал (с центром в соответствующем выборочном коэффициенте и радиусом, равным дисперсии), то есть 1000 вычислительных экспериментов дали правильный результат, при чём вне зависимости от размера выборки. Более точным был всё же коэффициент Пирсона, однако коэффициент Спирмена, как бы знаем, является более универсальным и позволяет оценивать степень коррелированности любых упорядоченных множеств, не обязательно ассоциированных с подмножеством вещественных чисел, что, безусловно, даёт куда большую область применения коэффициента Спирмена. Наше исследование показало его состоятельность.

Квадрантный коэффициент всюду, кроме нулевого коэффициента корреляции давал результат, далёкий от истинного.

- Для двумерного нормального распределения выборочные коэффициенты корреляции находятся в отношении: $r_S < r_Q \leq r$
- Для смеси двумерных нормальных распределений: $r_S < r_Q \leq r$

11.2. Оценки коэффициентов линейной регрессии

На выборке без возмущений критерии наименьших квадратов и наименьших модулей показали схожие результаты, хотя невязка по l^2-критерию оказалась значительно меньше: 4.01 по сравнению с 15.93.

На выборке с возмущениями метод наименьших модулей дал более точный результат, что подтверждает его робастность. И невязка по l^2-критерию оказалась намного меньше: 12.45 и 31.88.

11.3. Проверка гипотезы о законе генеральной совокупности. Метод χ^2

Исходя из полученных результатов, заключаем, что гипотеза H_0 о нормальности распределения $N(x, \hat{\mu}, \hat{\sigma})$ на уровне значимости $\alpha = 0.05$ верна.

Также мы видим, что гипотеза H_0 подтвердилась и для выборки распределения Лапласа, что не является верным. Это обусловлено тем, что теорема Пирсона говорит про асимптотическое распределение, а при малых размерах выборки результат не будет получаться достоверным.

11.4. Доверительные интервалы для матожидания и дисперсии

Для обоих подходов характерно, что при увеличении размера выборки доверительный интервал сужается, что является логичным, учитывая характер методов, с помощью которых были получены оценки.

Также видно, что асимптотический подход даёт значительно более широкий доверительный интервал для дисперсии на выборке из 20 элементов, что обусловлено собственно природой подхода: мы производим асимптотические оценки там, где они таковыми не являются.

При этом результаты для матожидания получились похожими в обоих подходах, потому что идея оценок в подходах абсолютно идентична. Отличаются лишь только функции распределения, фигурирующие в соответствующей оценке. При этом вид этих функций очень похож для любых распределений.

Таким образом, асимптотический подход можно рекомендовать для оценки матожидания любого распределения даже на сравнительно небольших выборках, чего, однако, нельзя сказать про дисперсию.

На выборке из 100 элементов методы дали схожие результаты, из чего можно сделать вывод, что асимптотический подход пригоден для применения на выборках такого размера.

12. Литература

[1] Максимов Ю. Д. Математическая статистика //СПб.: СПбГПУ. – 2004.

[2] А.Н. Тырсин, А. А. Азарян. Точное оценивание линейных регрессионных моделей МНМ на основе спуска по узловым прямым

[3] П. А. Акимов, А. И. Матасов, Уровни неоптимальности алгоритма Вейсфельда в МНМ, *Автомат. и телемех.*, 2010, выпуск 2, 4–16

13. Приложения

1. Репозиторий с кодом программыа:

   https://github.com/sairsey/MathStats