Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

СВОДНЫЙ ОТЧЕТ

Тема: ***Многомерные распределения. Оценки характеристик распределений***

Направление: 01.03.02 Прикладная математика и информатика

Выполнил студент гр. 33631/4 Камалетдинова Ю.

Преподаватель Баженов А.

Санкт-Петербург
2019

Содержание

Постановка задачи

В данной группе лабораторных работ рассматриваются многомерные распределения и методы оценки характеристик распределений. Запишем постановки задач

- Сгенерировать двумерные выборки $x_n = (x_1, \ldots, x_n),\ y_n = (y_1, \ldots, y_n)$ размерами $n = 20,\ 60,\ 100$ из нормального двумерного распределения. Коэффициенты корреляции взять равными $\rho = 0,\ 0.5,\ 0.9$. Формула для плотности распределения приведена ниже

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

Каждая выборка генерируется $N = 1000$ раз, и для каждоый выборки вычисляются среднее значение, среднее значение квадрата и дисперсия следующих коэффициентов

$$r_p = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2 \sum\limits_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{cov(x, y)}{\sqrt{\sigma_x^2 \sigma_y^2}} \text{ — коэффициент корреляции Пирсона,} \tag{2}$$

где $\overline{x},\ \overline{y}$ — выборочные средние $x_n,\ y_n$, $\sigma_x^2,\ \sigma_y^2$ — выборочные дисперсии

$$r_s = \rho_{rg_x, rg_y} = \frac{cov(rg_x, rg_y)}{\sqrt{\sigma_x^2 \sigma_y^2}} = 1 - \frac{6\sum\limits_{i=1}^{n} d_i^2}{n(n^2-1)} \text{ — коэффициент корреляции Спирмена,} \tag{3}$$

где $rg_x,\ rg_y$ — ранговые переменные, $d_i = rg(x_i) - rg(y_i)$ — разность двух рангов

наблюдений. Формула для рассчета из источника [2]

$$r_q = \frac{(n_I + n_{III}) - (n_{II} + n_{IV})}{n} \text{ — квадрантный коэффициент корреляции,} \quad (4)$$

где $n_i,\ i = I,\ II,\ III,\ IV$ — число наблюдений, попавших в i - ый квадрант на плоскости

Приведем формулы для вычисления выборочного среднего, квадрата выборочного среднего и выборочной дисперсии в двумерном случае

$$\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{i_k},\ \ k = 1,\ 2 \quad (5)$$

$$\overline{x}_k^2 = \frac{1}{n}\sum_{i=1}^{n} x_{i_k}^2,\ \ k = 1,\ 2 \quad (6)$$

$$\overline{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{i_k} - \overline{x}_k)^2,\ \ k = 1,\ 2 \quad (7)$$

Требуется повторить вычисления характеристик (5), (6), (7) корреляционных коэффициентов (2), (3), (4) для смеси нормальных распределений

$$f(x, y) = 0.9N(x, y, 0, 0, 1, 1, 0.9) + 0.1N(x, y, 0, 0, 10, 10, -0.9) \quad (8)$$

Полученные выборки необходимо изобразить на плоскости и изобразить эллипс равновероятности

- Найти оценки коэффициентов линейной регрессии $y_i = a + bx_i + e_i$, используя $n = 20$ точек на отрезке $[-1.8; 2]$ с равномерным шагом равным 0.2. Ошибку e_i считать нормально распределенной с параметрами (0, 1). В качестве эталонной зависимости

взять функцию

$$y_i = 2 + 2x_i + e_i \tag{9}$$

При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей. Требуется проделать описанную работу для выборки, у которой в значения y_1 и y_{20} вносятся возмущения 10 и -10.

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

- Сгенерировать выборки объемами $n = 20,\ 100$ элементов для нормального распределения $N(x; 0, 1)$, затем для параметров положения и масштаба построить асимптотически нормальные интервальные оценки на основе точечных оценок метода максимального правдоподобия. Также необходимо оценить параметры распределения на основе статистик χ^2 и Стьюдента. В качестве параметра надежности взять $\gamma = 0.95$.

Описание алгоритма

Метод наименьших квадратов

Введем обозначение для уравнения прямой, полученного по тому или иному критерию рассогласованности отклика и регрессионной модели

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \tag{10}$$

где $\hat{a}, \hat{b}$ — оценки параметров a, b

Запишем минимизируемое выражение для случая критерия наименьших квадратов (МНК)

$$Q(a,b) = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min_{a,b} \tag{11}$$

Опустим запись необходимых условий экстремума и доказательства минимальности функции (11) в стационарной точке, описанных в [4], и приведем МНК-оценки коэффициентов

$$\hat{b} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{12}$$

$$\hat{a} = \overline{y} - \overline{x}\hat{b}, \tag{13}$$

где $\overline{x}$, $\overline{x^2}$, $\overline{y}$, $\overline{xy}$ — выборочные первые и вторые начальные моменты

Метод наименьших модулей

Одной из альтернатив МНК является метод наименьших модулей (МНМ)

$$A(a,b) = \sum_{i=1}^{n} |y_i - a - bx_i| \to \min_{a,b} \tag{14}$$

Запишем выражения для оценок (12), (13) в другом виде

$$\hat{b} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{k_{xy}}{s_x s_y} = \frac{k_{xy}}{s_y^2} \frac{s_y}{s_x} = r_{xy} \frac{s_y}{s_x} \tag{15}$$

$$\hat{a} = \overline{y} - \overline{x}\hat{b}, \tag{16}$$

В формулах (15), (16) заменим выборочные средние $\overline{x}$, $\overline{y}$ на выборочные медианы $med\, x$, $med\, y$, а среднеквадратические отклонения s_x, s_y на интерквартильные широты

IQR_x, IQR_y; выборочный коэффициент корреляции r_{xy} — на знаковый коэффициент корреляции r_Q

$$\hat{b}_R = r_q \frac{IQR_y}{IQR_x}, \tag{17}$$

$$\hat{a}_R = med\ y - \hat{b}_R\ med\ x, \tag{18}$$

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} sign(x_i - med\ x)\ sign(y_i - med\ y) \tag{19}$$

$$sign\ z = \begin{cases} 1,\ z > 0 \\ 0,\ z = 0 \\ -1,\ z < 0 \end{cases} \tag{20}$$

Формулы (15), (16), (17), (18), (19), (20) указаны в учебнике [4]. Уравнение регрессии примет вид

$$y = \hat{a}_R + \hat{b}_R x \tag{21}$$

Метод максимального правдоподобия

Пусть $x_1, \ldots, x_n$ — случайная выборка из распределения с плотностью вероятности $f(x;\theta)$. Функцией правдоподобия (ФП) назовем совместную плотность вероятности независимых случайных величин $x_1, \ldots, x_n$, рассматриваемую как функцию неизвестного параметра θ

$$L(x_1, \ldots, x_n; \theta) = f(x_1;\theta) f(x_2;\theta) \ldots f(x_n;\theta) \tag{22}$$

Оценкой максимального правдоподобия $\hat{\theta}_{\text{МП}}$ будем называть такое значение, для которого из множества допустимых значений параметра θ ФП имеет наибольшее значение

при заданных $x_1, \ldots, x_n$

$$\hat{\theta}_{\text{МП}} = arg\ max\ L(x_1, \ldots, x_n; \theta) \tag{23}$$

Если функция правдоподобия дважды дифференцируема, ее стационарные значения задаются корнями уравнения

$$\frac{\partial L(x_1, \ldots, x_n; \theta)}{\partial \theta} = 0 \tag{24}$$

Запишем условие локального максимума $\overline{\theta}$

$$\frac{\partial^2 L}{\partial \theta^2}(x_1, \ldots, x_n; \overline{\theta}) < 0 \tag{25}$$

Наибольший локальный максимум будет являться решением задачи (23).

Мы будем искать максимум логарифма функции правдоподобия в виду того, что он имеет максимум в одной точке с функцией правдоподобия

$$\frac{\partial \ln L}{\partial \theta}, \text{ если } L > 0, \tag{26}$$

и будем решать уравнение правдоподобия

$$\frac{\partial \ln L}{\partial \theta} = 0 \tag{27}$$

Для проверки гипотезы о характере распределения воспользуемся критерием χ^2 для случая, когда параметры распределения известны. Пусть H_0 — гипотеза о генеральном законе распределения, H_1 — гипотеза о справедливости одного из конкурирующих законов распределений. Разобьем генеральную совокупность на k непересекающихся подмножеств $\Delta_1, \ldots, \Delta_k$ при условиях

$$p_i = P(X \in \Delta_i),\ i = \overline{1, k}\ ;\ \ \sum_{i=1}^{k} p_i = 1 \tag{28}$$

Положим n_i — частота попадания выборочного элемента в подмножество Δ_i. За меру отклонения выборочного распределения от гипотетического примем величину

$$Z = \sum_{i=1}^{k} c_i \left(\frac{n_i}{n} - p_i \right)^2 , \tag{29}$$

где $\frac{n_i}{n}$ — относительные частоты, c_i — некие положительные числа (веса). В качестве весов К. Пирсоном были взяты числа $c_i = \frac{n}{p_i}$. Получаем статистику критерия хи-квадрат К. Пирсона

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{p} - p_i \right)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \tag{30}$$

По теореме К. Пирсона из пособия [4] статистика критерия χ^2 асимптотически распределена по закону χ^2 с $k-1$ степенями свободы

Формулы (22) — (30) и определения взяты из источника [4]

Оценка на основе статистик Стьюдента и хи-квадрат

Пусть $x_1, \ldots, x_n$ — заданная выборка из нормального распределения $N(x; \mu, \sigma)$, по которой требуется оценить параметры μ, σ, генерального распределения. Построим на ее основе выборочные среднее $\overline{x}$ и среднее квадратическое отклонение s. Параметры распределения μ, σ не известны. В источнике [4] показано, что статистика Стьюдента

$$T = \sqrt{n-1}\, \frac{\overline{x} - \mu}{s} \tag{31}$$

распределена по закону Стьюдента с $n-1$ степенями свободы. Пусть $f_T(x)$ — плотность вероятности данного распределения. Тогда

$$\begin{aligned} P(-x < \sqrt{n-1}\, \frac{\overline{x} - \mu}{s} < x) = P(-x < \sqrt{n-1}\, \frac{\mu - \overline{x}}{s} < x) = \\ = \int_{-x}^{x} f_T(t)dt = 2F_T(x) - 1 \, , \end{aligned} \tag{32}$$

где $F_T(t)$ — функция распределения Стьюдента с $n-1$ степенями свободы. Положим $2F_T(x) - 1 = 1 - \alpha$, где α — выбранный уровень значимости. Тогда $F_T(x) = 1 - \alpha/2$. Положим $t_{1-\alpha/2}(n-1)$ — квантиль распределения Стьюдента с $n-1$ степенями свободы и уровнем значимости $1 - \alpha/2$. Из (31), (32) получаем

$$P\left(\overline{x} - \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}} < \mu < \overline{x} + \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}}\right) = 1 - \alpha\ , \tag{33}$$

что дает доверительный интервал для μ с вероятностью $\gamma = 1 - \alpha$

Для поиска оценки параметра σ воспользуемся источником [4], где показано, что случайная величина ns^2/σ^2 распределена по закону χ^2 с $n-1$ степенями свободы. Найдем квантили $\chi^2_{\alpha/2}(n-1)$, $\chi^2_{1-\alpha/2}(n-1)$ и приведем выражение для доверительного интервала для σ с доверительной вероятностью $\gamma = 1 - \alpha$

$$P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right) = 1 - \alpha \tag{34}$$

Асимптотический подход при построении оценок

Данный метод оценивания параметров применяется в случае неизвестности закона распределения, или когда он не является нормальным. Асимптотический метод построения доверительных интервалов основан на центральный предельной теореме.

Пусть $\overline{x}$ — выборочное среднее из выборки большого объема n независимых одинаково распределенных случайных величин. Тогда в силу центральной предельной теоремы случайная величина $(\overline{x} - M\overline{x})/\sqrt{D\overline{x}} = \sqrt{n}(\overline{x} - \mu)/\sigma$ распределена приблизительно нормально с параметрами 0, 1. Из данных рассуждений получим выражение для доверительного интервала для μ с доверительной вероятностью $\gamma = 1 - \alpha$

$$P\left(\overline{x} - \frac{su_{1-\alpha/2}}{\sqrt{n}} < \mu < \overline{x} + \frac{su_{1-\alpha/2}}{\sqrt{n}}\right) \approx \gamma\ , \tag{35}$$

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

Приведем выражение для доверительного интервала для σ с доверительной вероятностью $\gamma = 1 - \alpha$

$$s\,(1 - 0.5U) < \sigma < s\,(1 + 0.5U)\ , \tag{36}$$

где $U = u_{1-\alpha/2}\sqrt{(e+2)/n}$; e — выборочный эксцесс; m_4 — четвертый выборочный центральный момент.

Формулы (31) — (36) и определения взяты из источника [4]

Реализация

Для выполнения поставленных задач будем пользоваться библиотеками для языка Python: *numpy, scipy* – расчеты, законы распределения вероятностей; *matplotlib, seaborn* – визуализация результатов. Ход работы:

- Двумерное распределение
  - Задаем распределение с заданными параметрами
  - Формируем двойной цикл: внешний — по объемам выборок n, внутренний — по корреляционным коэффициентам ρ
  - На каждой итерации цикла для выборки строим 99% доверительный эллипс, теоретическое описание которого находится по ссылкам [1], [3]; изображаем выборки и эллипс в одних осях
  - Генерируем выборку и вычисляем корреляционные коэффициенты по формулам (2), (3), (4) 1000 раз
  - Находим среднее, среднее квадрата и дисперсию корреляций по формулам (5), (6), (7)
- Линейная регрессия
  - Задаем вектор точек $x_n = [-1.8, -1.6, \ldots, 2.0]$ с шагом 0.2, $n = 20$

- Вычисляем вектор значений функции (9)
- Рассчитываем оценки коэффициентов линейной регрессии по формулам (12), (13), (17), (18)
- Вносим возмущения $+10$ и -10 в первое и последнее значения регрессионной функции соответсвтенно и повторяем шаги 2, 3
- Изображаем полученные результаты на графике и сравниваем коэффициенты, рассчитанные по разным критериям

- Точечные оценки
  - Генерируем выборку из распределения $N(x; 0, 1)$ объемом $n = 100$
  - Запишем выражение для логарифма функции правдоподобия:

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{37}$$

  - Получим два уравнения правдоподобия:

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i - \hat{\mu}) = \dfrac{n}{\sigma^2}(\overline{x} - \hat{m}) = 0 \\ \dfrac{\partial \ln L}{\partial (\sigma^2)} = -\dfrac{n}{\hat{\sigma}^2} + \dfrac{1}{2(\hat{\sigma}^2)^2}\displaystyle\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \dfrac{n}{2(\hat{\sigma}^2)^2}\left[\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \hat{\mu})^2 - \hat{\sigma}^2\right] = 0 \end{cases} \tag{38}$$

  - Из уравнений (38) получили, что выборочное среднее $\overline{x}$ — оценка максимума правдоподобия математического ожидания: $\hat{\mu}_{\text{МП}} = \overline{x}$, а выборочная дисперсия $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}^2)$ — оценка максимума правдоподобия генеральной дисперсии $\hat{\sigma}^2_{\text{МП}} = s^2$

- Интервальные оценки
  - Генерируем выборки из распределения $N(0, 1)$ объемами $n = 20,\ 100$

– Вычисляем выборочные среднее, дисперсию, четвертый центральный момент, эксцесс по приведенным ниже формулам

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{39}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{40}$$

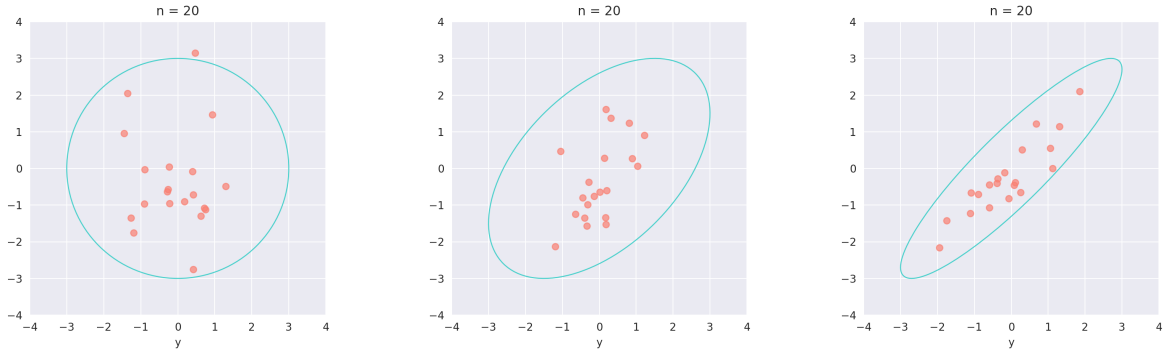
$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4 \tag{41}$$

$$e = \frac{m_4}{s^4} - 3 \tag{42}$$

– Вычисляем границы доверительных интервалов по формулам (33), (34), (35), (36)

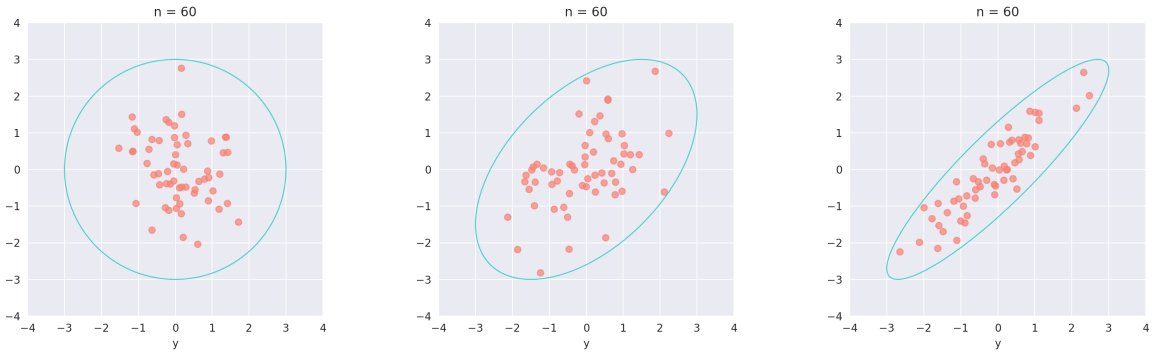
Результат

Выборки из двумерного нормального распределения и 99%-эллипс



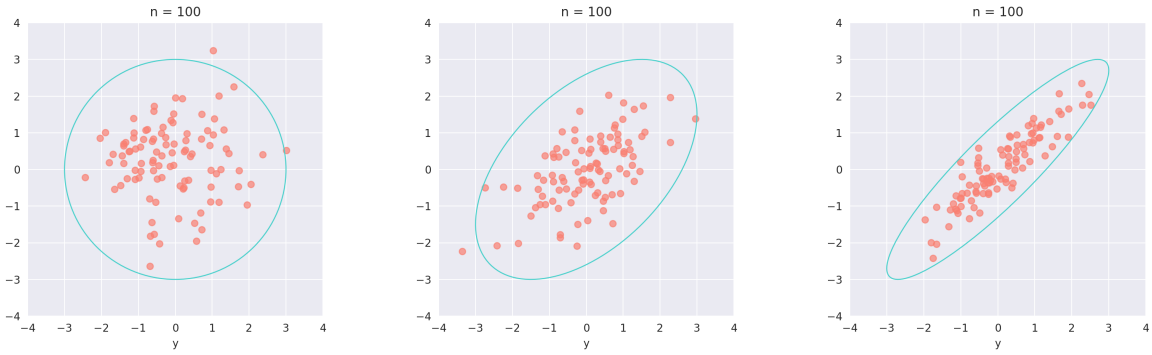
(a) (b) (c)

Рис. 1: ρ: (a) 0.0; (b) 0.5; (c) 0.9



(a) (b) (c)

Рис. 2: ρ: (a) 0.0; (b) 0.5; (c) 0.9



(a) (b) (c)

Рис. 3: ρ: (a) 0.0; (b) 0.5; (c) 0.9

Характеристики для двумерного нормального распределения

$n = 20, \rho = 0.0$	r_p	r_s	r_q
$E(z)$	0.0132	0.0086	0.0073
$E(z^2)$	0.0505	0.0517	0.0489
$D(z)$	0.0503	0.0517	0.0488

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_p^2) < E(r_s^2)$$
$$D(r_q) < D(r_p) < D(r_s)$$

$n = 20, \rho = 0.5$	r_p	r_s	r_q
$E(z)$	0.4893	0.4628	0.3263
$E(z^2)$	0.2712	0.2484	0.1530
$D(z)$	0.0318	0.0343	0.0465

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

$n = 20, \rho = 0.9$	r_p	r_s	r_q
$E(z)$	0.8942	0.8645	0.7170
$E(z^2)$	0.8019	0.7518	0.5384
$D(z)$	0.0024	0.0044	0.0243

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

$n = 60, \rho = 0.0$	r_p	r_s	r_q
$E(z)$	0.0024	0.0012	-0.0028
$E(z^2)$	0.0172	0.0175	0.0167
$D(z)$	0.0172	0.0175	0.0167

$$E(r_s) < E(r_p) < E(r_q)$$
$$E(r_q^2) < E(r_p^2) < E(r_s^2)$$
$$D(r_q) < D(r_p) < D(r_s)$$

$n = 60, \rho = 0.5$	r_p	r_s	r_q
$E(z)$	0.4942	0.4725	0.3372
$E(z^2)$	0.2544	0.2344	0.1287
$D(z)$	0.0101	0.0111	0.0150

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

$n = 60, \rho = 0.9$	r_p	r_s	r_q
$E(z)$	0.8991	0.8832	0.7103
$E(z^2)$	0.8090	0.7811	0.5122
$D(z)$	0.0007	0.0010	0.0077

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

$n = 100, \rho = 0.0$	r_p	r_s	r_q
$E(z)$	-0.0081	-0.0064	-0.0062
$E(z^2)$	0.0107	0.0108	0.0101
$D(z)$	0.0106	0.0107	0.0101

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_p^2) < E(r_s^2)$$
$$D(r_q) < D(r_p) < D(r_s)$$

$n = 100, \rho = 0.5$	r_p	r_s	r_q
$E(z)$	0.4980	0.4775	0.3341
$E(z^2)$	0.2538	0.2344	0.1206
$D(z)$	0.0058	0.0064	0.0090

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

$n = 100, \rho = 0.9$	r_p	r_s	r_q
$E(z)$	0.8994	0.8868	0.7134
$E(z^2)$	0.8093	0.7869	0.5137
$D(z)$	0.0004	0.0006	0.0048

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

Характеристики для смеси двумерных нормальных распределений

$n = 20$	r_p	r_s	r_q
$E(z)$	0.6886	0.6547	0.4866
$E(z^2)$	0.4904	0.4485	0.2738
$D(z)$	0.0162	0.0200	0.0370

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

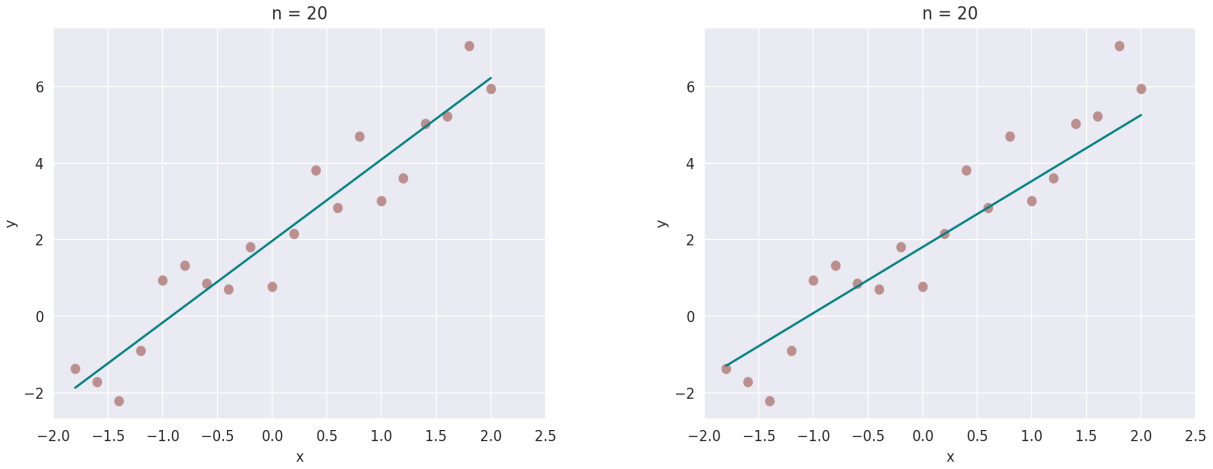
$n = 60$	r_p	r_s	r_q
$E(z)$	0.6948	0.6702	0.4885
$E(z^2)$	0.4874	0.4555	0.2506
$D(z)$	0.0047	0.0063	0.0119

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

$n = 100$	r_p	r_s	r_q
$E(z)$	0.7003	0.6796	0.4952
$E(z^2)$	0.4932	0.4653	0.2525
$D(z)$	0.0028	0.0035	0.0073

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

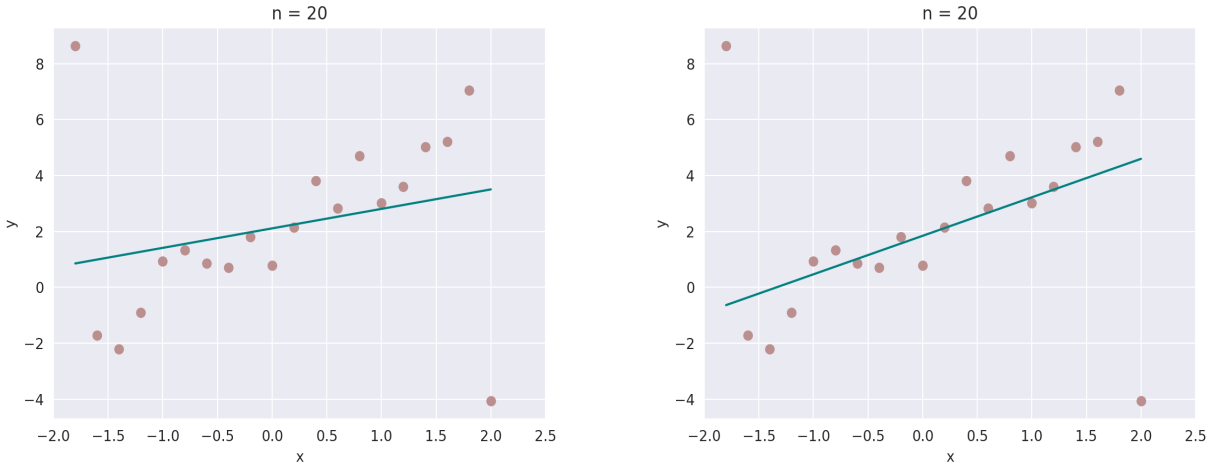
Линейная регрессия



(a) (b)

Рис. 4: График прямой, исходные данные без возмущений (a) МНК; (b) МНМ

4(a) $\hat{a} = 1.9677,\ \hat{b} = 2.1254$ 4(b) $\hat{a} = 1.8099,\ \hat{b} = 1.7207$



(a) (b)

Рис. 5: График прямой, исходные данные с возмущениями (a) МНК; (b) МНМ

5(a) $\hat{a} = 2.1106,\ \hat{b} = 0.6968$

5(b) $\hat{a} = 1.8443,\ \hat{b} = 1.3766$

Точечные оценки параметров

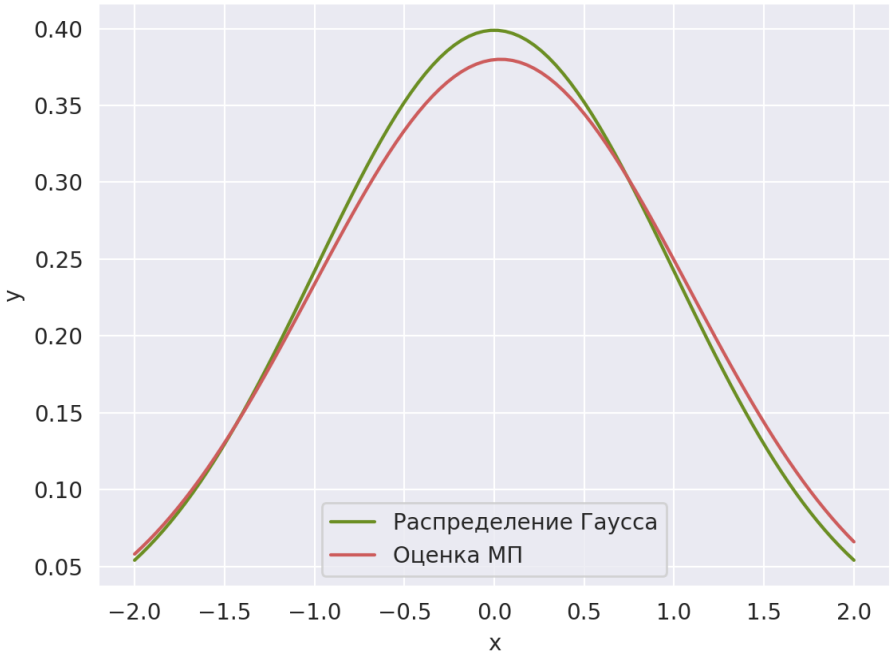


Рис. 6: Графики стандартного нормального распределения и распределения с параметрами, полученными методом МП

Представим табулированное значение квантиля хи-квадрат, $\alpha = 0.05$

- $\chi^2_{0.95}(15) = 26.296$

$k = 16$	a_{i-1} ; a_i	n_i	$\hat{p}_i$	$n_i - n\hat{p}_i$	$(n_i - n\hat{p}_i)^2/(n\hat{p}_i)$
1	-inf ; -3.50	0	0.0004	-0.04	0.04
2	-3.50 ; -2.96	0	0.0018	-0.18	0.18
3	-2.96 ; -2.42	1	0.0074	0.26	0.09
4	-2.42 ; -1.88	2	0.0241	-0.41	0.07
5	-1.88 ; -1.35	3	0.0604	-3.04	1.53
6	-1.35 ; -0.81	14	0.1169	2.31	0.46
7	-0.81 ; -0.27	22	0.1749	4.51	1.16
8	-0.27 ; 0.27	20	0.2023	-0.23	0.00
9	0.27 ; 0.81	15	0.1809	-3.09	0.53
10	0.81 ; 1.35	11	0.1251	-1.51	0.18
11	1.35 ; 1.88	7	0.0668	0.32	0.01
12	1.88 ; 2.42	4	0.0276	1.24	0.56
13	2.42 ; 2.96	1	0.0088	0.12	0.02
14	2.96 ; 3.50	0	0.0022	-0.22	0.22
15	3.50 ; inf	0	0.0005	-0.05	0.05
$\sum$	————	100	1	0	$\chi^2_B = 5.09$

Таблица 1: Таблица вычислений χ^2_B при проверке гипотезы о нормальности распределения

$\chi^2_B = 5.09 < 26.296 \approx \chi^2_{1-\alpha}(k-1)$ — гипотеза принимается

Интервальные оценки параметров

Представим значения квантилей распределений необходимых порядков, взятых из таблиц, $\alpha = 0.05$

- $t_{0.95}(19) = 1.72,\ t_{0.95}(99) = 1.66$ — квантили распределения Стьюдента
- $\chi^2_{0.025}(19) = 8.91,\ \chi^2_{0.975}(19) = 32.85,\ \chi^2_{0.025}(99) = 73.12,\ \chi^2_{0.975}(99) = 128.4$ — квантили распределения хи-квадрат
- $u_{0.975} = 1.96$ — квантиль стандартного нормального распределения

n	Интервал для μ	Интервал для σ
20	$(-0.239; 0.686)$	$(0.891; 1.711)$
100	$(-0.067; 0.266)$	$(0.876; 1.160)$

Таблица 2: Таблица оценок на основе статистик Стьюдента и хи-квадрат

n	Интервал для μ	Интервал для σ
20	$(-0.290; 0.737)$	$(0.873; 1.471)$
100	$(-0.096; 0.295)$	$(0.868; 1.126)$

Таблица 3: Таблица оценок на основе на основе асимптотического подхода

Вывод

Рассмотрим полученные соотношения для двумерного распределения. На некоррелированных данных квадрантный коэффициент корреляции имеет наименьшую дисперсию. Также дисперсия коэффициента Пирсона всегда меньше дисперсии коэффициента Спирмена вне зависимости от объема выборки или корреляции двумерного нормального

распределения (1). Можно полагать, что в случае такого распределения лучше рассчитывать коэффициент корреляции Пирсона.

Неравенства для смеси двух нормальных распределений (8) совпадают с неравенствами для двумерного нормального распределения с корреляциями $\rho = 0.5,\ 0.9$.

Решение задачи нахождения коэффициентов линейной регрессии показало, что наиболее устойчивым критерием к выбросам является метод наименьших модулей. Выборочная медиана и интерквартильные широты менее чувствительны к выбросам, что и объясняет полученные результаты.

Также можно заметить, что использование метода наименьших квадратов в случае отсутствия наблюдений, не свойственных данной выборке, дает лучшие результаты. Применение МНК при наличии больших по величине выбросов имеет смысл после предварительной отбраковки значений.

Оценка параметров распределения методом максимума правдоподобия в случае нормального распределения эффективна, состоятельна, асимптотически нормальна. В ходе проверки гипотезы было получено значение $\chi^2_B = 5.09$, являющееся очень малым, и соответствующий ему уровень значимости равен $\alpha = 0.99$, что говорит об очень хорошем согласии гипотезы H_0 и полученных данных.

По полученным интервальным оценкам можно говорить о том, что асимптотический подход не имеет преимущества по обоим параметрам сразу в случае малой выборки ($n = 20$) при условиях, что закон распределения известен и является нормальным. При объеме выборки $n = 100$ можно заметить сокращение длин доверительных интервалов ($0.333 < 0.391$ для μ, $0.258 < 0.284$ для σ), что является преимуществом асимптотического подхода в оценке параметров распределения.

Недостатком интервальных оценок на основе статистик Стьюдента и хи-квадрат может выступать сложность получения точечных оценок параметра распределения, если оно не является нормальным, что в с свою очередь усложнит вычисления.

Список литературы

[1] *Eisele, R.* (2018). How to plot a covariance error ellipse. URL: https://www.xarg.org/2018/04/how-to-plot-a-covariance-error-ellipse/

[2] *Zwillinger, D. and Kokoska, S.* (2000). CRC Standard Probability and Statistics Tables and Formulae. Chapman & Hall: New York. 2000.

[3] *Ллойд Э., Ледерман У.* Справочник по прикладной статистике. Том 1. М.: Финансы и статистика, 1989. - 510 с.

[4] *Амосова Н.Н., Куклин Б.А., Макарова С.Б., Максимов Ю.Д., Митрофанова Н.М., Полищук В.И., Шевляков Г.Л.* Вероятностные разделы математики. Учебник для бакалавров технических направлений. — СПб.: Иван Федоров, 2001. — 592 с.: илл. — ISBN 5-81940-050-X.