Санкт-Петербургский Политехнический университет Петра Великого
Институт прикладной математики и механики
Кафердра «Прикладная математика и информатика»

Отчёт по проекту
Дисциплина: «Математическая статистика»
Лабораторные работы № 5-8

Выполнил студент гр. 3630102/70201 Н. А. Счастливцев
Преподаватель А. Н. Баженов

СПб 2020

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

Требуется:

1. Сгенерировать двумерные выборки размерами 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:

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

   Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

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

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

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

2 Теория

2.1 Корреляция двумерной случайной величины

2.1.1 Двумерное нормальное распределение

Определение. Двумерная случайная величина (X, Y) называется распределённой нормально (или просто нормальной), если её плотность вероятности определяется формулой: [1]

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

Двумерная нормальная плотность распределения содержит пять параметров: $m_1, m_2, \sigma_1, \sigma_2, \rho$. Из них первые два m_1, m_2 — определяют центр симметрии распределения (m_1, m_2). Это точка, в которую проектируется вершина поверхности, являющейся графиком нормальной плотности. Сама эта поверхность является холмообразной.

Сечения этой поверхности плоскостями, параллельными xOy, являются эллипсы с центрами, лежащими на оси симметрии. Построим проекции эллипса на xOy:

$$\frac{(x-\overline{x})^2}{\sigma_X^2} - 2\rho\frac{(x-\overline{x})(y-\overline{y})}{\sigma_X\sigma_Y} + \frac{(y-\overline{y})^2}{\sigma_Y^2} = \text{const} \tag{3}$$

Компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $\mathbb{E}x = m_1$, $\mathbb{E}y = m_2$ и средними квадратическими отклонениями $\sigma_X = \sigma_1$, $\sigma_Y = \sigma_2$ соответственно.

Параметр ρ называется коэффициентом корреляции. [1]

2.1.2 Коэффициент корреляции

Определение. Коррелляционным моментом, иначе ковариацией, двух случайных величин X и Y называется математическое ожидание произведения отклонений этих случайных величин от их математических ожиданий.

Обозначение корреляционного момента: K_{XY}, $\operatorname{cov}(X, Y)$.

Таким образом,

$$K_{XY} = \operatorname{cov}(X,Y) = \mathbb{E}\left[(X - m_X)(Y - m_Y)\right] \tag{4}$$

Очевидно, что $K_{XY} = K_{YX} (= K)$. [1]

Определение. Коэффициентом корреляции ρ_{XY} двух случайных величин X и Y называется отношение их корреляционного момента к произведению их средних квадратических отклонений:

$$\rho_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y} \tag{5}$$

Очевидно, что $\rho_{XY} = \rho_{YX}$. [1]

Коэффициент корреляции является нормированной величиной ($|\rho_{XY}| \leq 1$) и характеризует меру линейной зависимости случайных величин X и Y.

2.1.3 Выборочные коэффициенты корреляции

Пусть по выборке $\{x_i, y_i\}_1^n$ двумерной случайной величины (X, Y) требуется оценить генеральный коэффициент корреляции ρ_{XY}. [2]

Есть несколько вариантов оценки:

1. Выборочный коэффициент корреляции Пирсона:

$$r = r_{XY} = \frac{\overline{K_{XY}}}{s_X s_Y} \tag{6}$$

[2] Здесь:

- Выборочная ковариация:

$$\overline{K_{XY}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy} \tag{7}$$

- Выборочные дисперсии компонент:

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2; s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{8}$$

2. Знаковый коэффициент корреляции (квандрантный):

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}(x_i - \operatorname{med} x) \operatorname{sign}(y_i - \operatorname{med} y) \tag{9}$$

3. Ранговый коэффициент корреляции Спирмена:

$$r_s = 1 - \frac{6\sum\limits_{i=1}^{n}\left[\operatorname{rank}(x_i) - \operatorname{rank}(y_i)\right]^2}{n(n^2-1)} \tag{10}$$

Здесь $\operatorname{rank}(x_j) = i$, где $x_{(i)}$ — i-ая порядковая статистика.

2.1.4 Эллипс рассеивания

Из уравнения (3) видно, что если $\rho \neq 0$, то эллипс повёрнут.

Так как эллипсы, построенные по этому принципу, являются линиями уровня поверхности, образованной зависимостью $f(x, y)$, плотность распределения на этих эллипсах постоянна. Такие эллипсы называются эллипсами равной плотности или эллипсами рассеивания.

2.2 Линейная регрессия

2.2.1 Модель простой линейной регрессии

Определение. Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n, \tag{11}$$

где $x_1, \ldots, x_n$ — заданные числа (значения фактора); $y_1, \ldots, y_n$ — наблюдаемые значения отклика; $\varepsilon_1, \ldots, \varepsilon_n$ — независимые, нормально распределённые $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (наблюдаемые); β_0, β_1 — неизвестные параметры, подлежащие оцениванию. [1]

В модели (11) отклик y зависит от одного фактора x, и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика y. Погрешности результатов измерений x в этой модели полагают существенно меньшими погрешностей результатов измерений y, так что ими можно пренебречь. [1]

2.2.2 Метод наименьших квадратов и метод наименьших модулей

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распространённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определеяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0, \beta_1} \tag{12}$$

Определение. Задача минимизации квардратичного критерия (12) носит название задачи метода наименьших квадратов (МНК), а оценки $\overline{\beta_0}, \overline{\beta_1}$ параметров β_0, β_1, реализующие минимум критерия (12), называют МНК-оценками. [1]

Заметим, что сумма в критерии (12) представляет собой квадрат нормы «невязки» векторов $\overline{y} = (y_1, y_2, \ldots, y_n)^T$ и $\overline{y_\varepsilon}$ из пространства l^2, где $\overline{y_\varepsilon} = (\beta_0 + \beta_1 x_1, \beta_0 + \beta_1 x_2, \ldots, \beta_0 + \beta_1 x_n)$ и обе последовательности l^2 имеют конечное число ненулевых членов (не больше, чем n). Так как последовательности, соответствующие элементам $\overline{y}$ и $\overline{y_\varepsilon}$ обладают этим свойствам, то они являются элементами сразу всех пространств l^p, где $1 \leq p \leq \infty$, поэтому можно рассматривать другую норму невязки, например, из пространства l^1. Такой подход приведёт к выражению для метода наименьших модулей:

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

2.3 Рассчётные формулы для МНК-оценок

МНК-оценки параметров $\overline{\beta_0}$ и $\overline{\beta_1}$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум (функции: (12), (13)). [1]

Для нахождения МНК-оценок $\overline{\beta_0}$ и $\overline{\beta_1}$ выпишем необходимые условия экстремума (для уравнения (12)):

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2 \sum\limits_{i=1}^{n} (y_i - \overline{\beta_0} - \beta_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2 \sum\limits_{i=1}^{n} (y_i - \overline{\beta_0} - \beta_1 x_i) x_i = 0 \end{cases} \tag{14}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы (14) получим:

$$\begin{cases} n\overline{\beta_0} + \overline{\beta_1} \sum x_i & = \sum y_i \\ \overline{\beta_0} \sum x_i + \overline{\beta_1} \sum x_i^2 & = \sum x_i y_i \end{cases} \tag{15}$$

Разделим оба уравнения на n:

$$\begin{cases} \overline{\beta_0} + \left(\dfrac{1}{n} \sum x_i \right) \overline{\beta_1} & = \dfrac{1}{n} \sum y_i \\ \left(\dfrac{1}{n} \sum x_i \right) \overline{\beta_0} + \left(\dfrac{1}{n} \sum x_i^2 \right) \overline{\beta_1} & = \dfrac{1}{n} \sum x_i y_i \end{cases} \tag{16}$$

и, используя известные статистические обозначения для выборочных первых и вторых начальных моментов

$$\overline{x} = \frac{1}{n}\sum x_i, \overline{y} = \frac{1}{n}\sum y_i, a_{2x} = \overline{x^2} = \frac{1}{n}\sum x_i^2, \overline{xy} = \frac{1}{n}\sum x_i y_i, \tag{17}$$

получим:

$$\begin{cases} \overline{\beta_0} + \overline{x}\overline{\beta_1} & = \overline{y} \\ \overline{x}\overline{\beta_0} + \overline{x^2\beta_1} & = \overline{xy} \end{cases} \tag{18}$$

откуда МНК-оценку $\overline{\beta_1}$ наклона прямой регрессии находим по формуле Крамера

$$\overline{\beta_1} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{19}$$

а МНК-оценку $\overline{\beta_0}$ определяем непосредственно из первого уравнения системы (18):

$$\overline{\beta_0} = \overline{y} - \overline{x}\overline{\beta_1} \tag{20}$$

Заметим, что определитель системы (18)

$$\overline{x^2} - (\overline{x})^2 = n^{-1}\sum (x_i - \overline{x})^2 = s_x^2 > 0, \tag{21}$$

если среди значений $x_1, \ldots, x_n$ есть различные, что и будем предполагать.

Доказательство минимальности функции $Q(\beta_0, \beta_1)$ в стационарной точке проведём с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2 = 2n\overline{x^2}, \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} = 2\sum x_i = 2n\overline{x} \tag{22}$$

$$\Delta = \frac{\partial^2 Q}{\partial \beta_0^2} \cdot \frac{\partial^2 Q}{\partial \beta_1^2} - \left(\frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1}\right)^2 = 4n^2\overline{x^2} - 4n^2(\overline{x})^2 = 4n^2\left[\overline{x^2} - (\overline{x})^2\right] =$$

$$= 4n^2\left[\frac{1}{n}\sum (x_i - \overline{x})^2\right] = 4n^2 s_x^2 > 0 \tag{23}$$

Этот результат вместе с условием $\dfrac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0$ означает, что в стационарной точке функция Q имеет минимум. [1]

2.3.1 Расчётные формулы для метода наименьших модулей

Аналогичным образом получаются формулы для метода наименьших модулей. Ищется минимум выражения (13). Для того, чтобы получилось проделать те же выкладки, доопределим производную от модуля в нуле нулём, то есть, положим:

$$|z| = \operatorname{sign} z = \begin{cases} 1 & \text{при } z > 0 \\ 0 & \text{при } z = 0 \\ -1 & \text{при } z < 0 \end{cases} \tag{24}$$

Если заново проделать выкладки, выборочные средние значения заменятся на выборочные медианы. Получим:

$$\overline{\beta_1} = r_Q \frac{q_y^*}{q_x^*}, \overline{\beta_0} = \operatorname{med} y - \overline{\beta_1} \operatorname{med} x, \tag{25}$$

где r_Q — знаковый коэффициент корреляции (9), q_y^* и q_x^* — интерквартильные широты.

2.4 Оценка параметров распределения

2.4.1 Метод максимального правдоподобия

[2]

Метод максимального правдоподобия, созданный Фишером, является достоточно универсальным и плодотворным методом оценивания.

Пусть имеется выборка $(x_1, \ldots, x_n)$ из генеральной совокупности с плотностью вероятности $f(x, \theta)$, содержащей один неизвестный параметр θ.

Выборка является n-мерной случайной величиной, компоненты x_i которой взаимно независимы, одинаково распределены с плотностью $f(x, \theta)$. Тогда плотность распределения n-мерной случайной величины $(x_1, x_2, \ldots, x_n)$ будет равна:

$$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta) \tag{26}$$

Эта функция называется функцией правдоподобия для рассматриваемой выборки.

Будем считать θ переменной неслучайной величиной, а элементы $x_1, x_2, \ldots, x_n$ выборки фиксированными, так как выборка фактически осуществлена. Если придавать θ различные значения, то естественно ожидать, что плотность $L(x_1, x_2, \ldots, x_n; \theta)$ примет максимальное значение в случае,

когда θ окажется равным истинному его значению, так как при других значениях θ менее вероятно за один раз получить именно данную выборку.

Эти интуитивные соображения приводят к тому, что за оценку θ берут такое значение $\bar{\theta}$, при котором функция правдоподобия достигает максимума.

Технически (так как L состоит из произведений) удобнее искать $\max \ln L$ (точка $\bar{\theta}$, дающая максимум $\ln L$, даёт и максимум L). Итак, для отыскания $\bar{\theta}$ имеем уравнение:

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

которое называется уравнением правдоподобия, а его решение $\bar{\theta} = \bar{\theta}(x_1, x_2, \ldots, x_n)$, зависящее от элементов выборки, оценкой максимального правдоподобия.

При выполнении достаточно общих условий оценки максимального правдоподобия являются состоятельными и асимптотически эффективными. В общем случае они являются смещёнными.

В случае, когда генеральная плотность вероятности $f(x, \theta_1, \ldots, \theta_k)$ содержит k параметров, вместо одного уравнения правдоподобия решается система уравнений:

$$\frac{\partial \ln L}{\partial \theta_1} = 0, \ldots, \frac{\partial \ln L}{\partial \theta_k} = 0. \tag{28}$$

2.4.2 Оценка параметров нормальной величины

[2]

Для нормального закона $N(m, \sigma)$ плотность вероятности имеет вид:

$$f(x, m, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{29}$$

Удобно считать, что здесь два параметра m и σ^2. Следовательно, функция правдоподобия равна:

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{(x_1-m)^2}{2\sigma^2} - \ldots - \frac{(x_n-m)^2}{2\sigma^2}} \tag{30}$$

Тогда

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2. \tag{31}$$

Далее, дифференцируя $\ln L$ по m и σ^2, получаем систему уравнений правдоподобия:

$$\begin{cases} \dfrac{\partial \ln L}{\partial m} = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(x_i - m) = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(x_i - m)^2 = 0 \end{cases} \tag{32}$$

Из первого уравнения находим $\sum\limits_{i=1}^{n} x_i - nm = 0$. Отсюда

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}. \tag{33}$$

Из второго уравнения:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 = n \tag{34}$$

$$(\bar{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^2. \tag{35}$$

2.4.3 Критерий проверки. Метод хи-квадрат

Пусть есть гипотеза H_0 о виде закона распределения. Следующий алгоритм позволяет отвергнуть или принять гипотезу H_0.

Критерий проверки. [2]

1. Выбираем уровень значимости α.

2. С помощью гипотетической функции распределения $F(x)$ с l оценёнными параметрами вычисляем оценки вероятностей $\bar{p_i} = P\{X \in \Delta_i\}, (i = 1, 2, \ldots, k)$.

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

4. Находим частоты n_i попадания элементов выборки в подмножества Δ_i и вычисляем выборочное значение статистики критерия хи-квадрат:

$$\chi^2_B = \sum_{i=1}^{k} \frac{(n_i - n\bar{p_i})^2}{n\bar{p_i}}. \tag{36}$$

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

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

2.5 Интервальное оценивание

2.5.1 Задача интервального оценивания

[1]

Пусть требуется по заданной выборке $(x_1, x_2, \ldots, x_n)$ оценить числовую характеристику или параметр θ генерального распределения.

Определение. Доверительным интервалом (иначе, интервальной оценкой) числовой характеристики или параметра распределения θ генеральной совокупности с доверительной вероятностью γ называется интервал $(\theta_1; \theta_2)$ со случайными границами $\theta_1 = \theta_1(x_1, x_2, \ldots, x_n), \theta_2 = \theta_2(x_1, x_2, \ldots, x_n)$, который накрывает θ с вероятностью γ:

$$P(\theta_1 < \theta < \theta_2) = \gamma. \tag{37}$$

Часто вместо доверительной вероятности γ рассматривается вероятность

$$\alpha = 1 - \gamma, \tag{38}$$

называемая уровнем значимости.

2.5.2 Доверительные интервалы для параметров нормального распределения

[1]

Построим доверительный интервал для математического ожидания m нормального распределения.

Дана выборка $(x_1, x_2, \ldots, x_n)$ объёма n из нормальной генеральной совокупности. На её основе строим выборочное среднее $\overline{x}$ и выборочное среднее квадратическое отклонение s. Параметры m и σ нормального распределения неизвестны.

Доказано, что случайная величина

$$T = \sqrt{n-1} \cdot \frac{\overline{x} - m}{s}, \tag{39}$$

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

$$P(-x < \sqrt{n-1} \cdot \frac{\overline{x} - m}{s} < x) = P(-x < \sqrt{n-1} \cdot \frac{m - \overline{x}}{s} < x) =$$
$$= \int_{-x}^{x} f_T(t)dt = 2\int_{0}^{x} f_T(t)dt = 2\left(\int_{-\infty}^{x} f_T(t)dt - \frac{1}{2}\right) = 2F_T(x) - 1. \tag{40}$$

Здесь $F_T(x)$ — функция распределения Стьюдента с $n-1$ степенями свободы.

Полагаем $2F_T(x) - 1 = 1 - \alpha$, где α — выбранный уровень значимости. Тогда $F_T(x) = 1 - \alpha/2$. Пусть $t_{1-\alpha/2}(n-1)$ — квантиль распределения Стьюдента с $n-1$ степенями свободы и порядка $1-\alpha/2$. Из предыдущих равенств мы получаем

$$P\left(\overline{x} - \frac{sx}{\sqrt{n-1}} < m < \overline{x} + \frac{sx}{\sqrt{n-1}}\right) = 2F_T(x) - 1 = 1 - \alpha,$$
$$P\left(\overline{x} - \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}} < m < \overline{x} + \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}}\right) = 1 - \alpha, \tag{41}$$

что и даёт доверительный интервал для m с доверительной вероятностью $\gamma = 1 - \alpha$.

Построим доверительный интервал для среднеквадратического отклонения σ нормального распределения.

Дана выборка $(x_1, x_2, \ldots, x_n)$ объёма n из нормальной генеральной совокупности. На её основе строим выборочную дисперсию s^2. Параметры m и σ нормального распределения неизвестны. Доказано, что случайная величина ns^2/σ^2 распределена по закону χ^2 с $n-1$ степенями свободы.

Задаёмся уровнем значимости α и по таблице находим квантили $\chi^2_{\alpha/2}(n-1)$ и $\chi^2_{1-\alpha/2}(n-1)$. Это значит, что

$$P(\chi^2(n-1) < \chi^2_{\alpha/2}(n-1)) = \alpha/2;$$
$$P(\chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)) = 1-\alpha/2. \quad (42)$$

Тогда

$$P(\chi^2_{\alpha/2}(n-1) < \chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)) =$$
$$= P(\chi^2(n-1) < \chi^2_{1-\alpha/2}(n-1)) - P(\chi^2(n-1) < \chi^2_{\alpha/2}(n-1)) =$$
$$= 1-\alpha/2-\alpha/2 = 1-\alpha. \quad (43)$$

Отсюда

$$P\left(\chi^2_{\alpha/2} < \frac{ns^2}{\sigma^2} < \chi^2_{1-\alpha/2}\right) = P\left(\frac{1}{\chi^2_{\alpha/2}} < \frac{\sigma^2}{ns^2} < \frac{1}{\chi^2_{1-\alpha/2}}\right) =$$
$$= P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right). \quad (44)$$

Окончательно

$$P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right) = 1-\alpha, \quad (45)$$

что и даёт доверительный интервал для σ с доверительной вероятностью $\gamma = 1-\alpha$.

2.5.3 Доверительные интервалы для параметров произвольного распределения при большом объёме выборки. Асимптотический подход

[1]

Если закон распределения исследуемой генеральной совокупности неизвестен или он не является нормальным, то методы построения доверительных интервалов для m и σ, развитые выше, здесь неприменимы.

При большом объёме выборки для построения доверительных интервалов может быть использован асимптотический метод на основе центральной предельной теоремы.

Доверительный интервал для математического ожидания m произвольной генеральной совокупности при большом объёме выборки.

Выборочное среднее $\overline{x} = \sum\limits_{i=1}^{n} \frac{x_i}{n}$ при большом объёме выборки является суммой большого числа взаимно независимых одинаково распределённых случайных величин. Предполагаем, что исследуемое генеральное распределение имеет конечные математическое ожидание m и дисперсию σ^2. Тогда в силу центральной предельной теоремы центрированная и нормированная случайная величина $(\overline{x} - \mathbb{E}\overline{x})/\sqrt{\mathbb{D}\overline{x}} = \sqrt{n} \cdot (\overline{x} - m)/\sigma$ распределена приблизительно нормально с параметрами 0 и 1. Пусть

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2} dt \tag{46}$$

— функция Лапласа. Тогда

$$P\left(-x < \sqrt{n} \cdot \frac{\overline{x} - m}{\sigma} < x\right) = P\left(-x < \sqrt{n} \cdot \frac{m - \overline{x}}{\sigma}\right) \approx$$
$$\approx \Phi(x) - \Phi(-x) = 2\Phi(x) - 1. \tag{47}$$

Отсюда

$$P\left(\overline{x} - \frac{\sigma x}{\sqrt{n}} < m < \overline{x} + \frac{\sigma x}{\sqrt{n}}\right) \approx 2\Phi(x) - 1. \tag{48}$$

Полагаем $2\Phi(x) - 1 = \gamma = 1 - \alpha$; тогда $\Phi(x) = 1 - \alpha/2$. Пусть $u_{1-\alpha/2}$ — квантиль нормального распределения $N(0, 1)$ порядка $1 - \alpha/2$. Заменяя в равенстве (48) σ на s, запишем его в виде

$$P\left(\overline{x} - \frac{s u_{1-\alpha/2}}{\sqrt{n}} < m < \overline{x} + \frac{s u_{1-\alpha/2}}{\sqrt{n}}\right) \approx \gamma, \tag{49}$$

что и даёт доверительный интервал для m с доверительной вероятностью $\gamma = 1 - \alpha$.

Доверительный интервал для среднего квадратического отклонения σ произвольной генеральной совокупности при большом объёме выборки.

Выборочная дисперсия s^2 при большом объёме выборки является суммой большого числа практически взаимно независимых случайных величин (имеется одна связь $\sum_{i=1}^{n} x_i = n\overline{x}$, которой при большом n можно пренебречь). Предполагаем, что исследуемая генеральная совокупность имеет конечные первые четыре момента.

В силу центральной предельной теоремы центрированная и нормированная случайная величина $(s^2 - \mathbb{E}s^2)/\sqrt{\mathbb{D}s^2}$ при большом объёме выборки n распределена приблизительно нормально с параметрами 0 и 1. Пусть $\Phi(x)$ — функция Лапласа (46). Тогда

$$P\left(-x < \frac{s^2 - \mathbb{E}s^2}{\sqrt{\mathbb{D}s^2}} < x\right) \approx \Phi(x) - \Phi(-x) = 2\Phi(x) - 1. \qquad (50)$$

Положим $2\Phi(x) = \gamma = 1 - \alpha$. Тогда $\Phi(x) = 1 - \alpha/2$. Пусть $u_{1-\alpha/2}$ — корень этого уравнения — квантиль нормального распределения $N(0,1)$ порядка $1-\alpha/2$. Было получено: $\mathbb{E}s^2 \approx \sigma^2$ при большом n. Кроме того, известно, что $\mathbb{D}s^2 \approx \dfrac{\mu_4 - \mu_2^2}{n}$. Здесь μ_k — центральный момент k-го порядка генерального распределения; $\mu_2 = \sigma^2$; $\mu_4 = \mathbb{E}\left[(x - \mathbb{E}x)^4\right]$; $o(\frac{1}{n})$ — бесконечно малая высшего порядка, чем $1/n$, при $n \to \infty$. Итак, $\mathbb{D}s^2 \approx \frac{\mu_4 - \sigma^4}{n}$. Отсюда

$$\mathbb{D}s^2 \approx \frac{\sigma^4}{n}\left(\frac{\mu_4}{\sigma^4} - 1\right) = \frac{\sigma^4}{n}\left(\left(\frac{\mu_4}{\sigma^4} - 3\right) + 2\right) = \frac{\sigma^4}{n}(E+2) \approx \frac{\sigma^4}{n}(e+2), \quad (51)$$

где $E = \frac{\mu_4}{\sigma^4} - 3$ — эксцесс генерального распределения, $e = \frac{m_4}{s^4} - 3$ — выборочный эксцесс; $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$ — четвёртый выборочный центральный момент. Далее,

$$\sqrt{\mathbb{D}s^2} \approx \frac{\sigma^2}{\sqrt{n}}\sqrt{e+2}. \qquad (52)$$

Преобразуем неравенства, стоящие под знаком вероятности в формуле

$$P\left(-x < \frac{s^2 - \mathbb{E}s^2}{\sqrt{\mathbb{D}s^2}} < x\right) = \gamma;$$

$$-\sigma^2\mathbb{U} < s^2 - \sigma^2 < \sigma^2\mathbb{U};$$

$$\sigma^2(1-\mathbb{U}) < s^2 < \sigma^2(1+\mathbb{U});$$

$$1/[\sigma^2(1-\mathbb{U})] < 1/s^2 < 1/[\sigma^2(1+\mathbb{U})];$$

$$s^2/(1-\mathbb{U}) > \sigma^2 > s^2/(1+\mathbb{U});$$

$$s(1+\mathbb{U})^{-1/2} < \sigma < s(1-\mathbb{U})^{-1/2}, \quad (53)$$

где $\mathbb{U} = u_{1-\alpha/2}\sqrt{(e+2)/n}$ или

$$s\left(1+u_{1-\alpha/2}\sqrt{(e+2)/n}\right)^{-1/2} < \sigma < s\left(1-u_{1-\alpha/2}\sqrt{(e+2)/n}\right)^{-1/2}. \quad (54)$$

Разлагая функции в биномиальный ряд и оставляя первые два члена, получим

$$s(1-0.5\mathbb{U}) < \sigma < s(1+0.5\mathbb{U}). \quad (55)$$

или

$$s\left(1-0.5u_{1-\alpha/2}\sqrt{(e+2)}/\sqrt{n}\right) < \sigma < s\left(1+0.5u_{1-\alpha/2}\sqrt{(e+2)}/\sqrt{n}\right). \quad (56)$$

Формулы (53) и (55) дают доверительный интервал для σ с доверительной вероятностью $\gamma = 1 - \alpha$.

3 Реализация

Результаты лабораторной получены с использованием средств языка программирования R и среды R-studio. Ссылка на репозиторий с кодом. Использованные библиотеки: stringr, MASS, car, plotrix, L1pack, ggplot2, moments.

4 Результаты

4.1 Корреляция

4.1.1 Выборочные коэффициенты корреляции

Список таблиц с результатами представлен в таблице 1:

ρ	n = 20	n = 60	n = 100
0	Таблица 2	Таблица 5	Таблица 8
0.5	Таблица 3	Таблица 6	Таблица 9
0.9	Таблица 4	Таблица 7	Таблица 10
Смесь (1)	Таблица 11	Таблица 12	Таблица 13

Таблица 1: Эксперименты, коэффициенты корреляции

n = 20, $\rho = 0$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.007	0.050	0.05
Пирсона (6)	-0.002	0.05	0.05
Спирмена (10)	-0.01	0.05	0.05

Таблица 2: Выборочные коэффициенты корреляции, n = 20, $\rho = 0$

n = 20, $\rho = 0.5$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.34	0.16	0.045
Пирсона (6)	0.490	0.27	0.03
Спирмена (10)	0.46	0.25	0.040

Таблица 3: Выборочные коэффициенты корреляции, n = 20, $\rho = 0.5$

n = 20, $\rho = 0.9$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.7	0.520	0.024
Пирсона (6)	0.894	0.803	0.0026
Спирмена (10)	0.866	0.755	0.005

Таблица 4: Выборочные коэффициенты корреляции, n = 20, $\rho = 0.9$

n = 60, $\rho = 0$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	-0.001	0.016	0.016
Пирсона (6)	0.003	0.017	0.017
Спирмена (10)	0.004	0.017	0.017

Таблица 5: Выборочные коэффициенты корреляции, n = 60, $\rho = 0$

n = 60, $\rho = 0.5$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.34	0.131	0.015
Пирсона (6)	0.50	0.260	0.010
Спирмена (10)	0.48	0.24	0.01

Таблица 6: Выборочные коэффициенты корреляции, n = 60, $\rho = 0.5$

n = 60, $\rho = 0.9$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.712	0.516	0.009
Пирсона (6)	0.8990	0.8083	0.0007
Спирмена (10)	0.882	0.779	0.001

Таблица 7: Выборочные коэффициенты корреляции, n = 60, $\rho = 0.9$

n = 100, $\rho = 0$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	-0.005	0.010	0.010
Пирсона (6)	-0.0001	0.01	0.01
Спирмена (10)	0.001	0.01	0.01

Таблица 8: Выборочные коэффициенты корреляции, n = 100, $\rho = 0$

n = 100, $\rho = 0.5$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.333	0.120	0.009
Пирсона (6)	0.5	0.257	0.006
Спирмена (10)	0.476	0.233	0.007

Таблица 9: Выборочные коэффициенты корреляции, n = 100, $\rho = 0.5$

n = 100, $\rho = 0.9$	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.712	0.512	0.005
Пирсона (6)	0.8993	0.809	0.0004
Спирмена (10)	0.886	0.786	0.0007

Таблица 10: Выборочные коэффициенты корреляции, n = 100, $\rho = 0.9$

n = 20, смесь (1)	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.680	0.490	0.027
Пирсона (6)	0.68	0.47	0.01
Спирмена (10)	0.680	0.472	0.005

Таблица 11: Выборочные коэффициенты корреляции, n = 20, смесь (1)

n = 60, смесь (1)	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.872	0.764	0.003
Пирсона (6)	0.876	0.768	0.001
Спирмена (10)	0.8770	0.7690	0.0006

Таблица 12: Выборочные коэффициенты корреляции, n = 60, смесь (1)

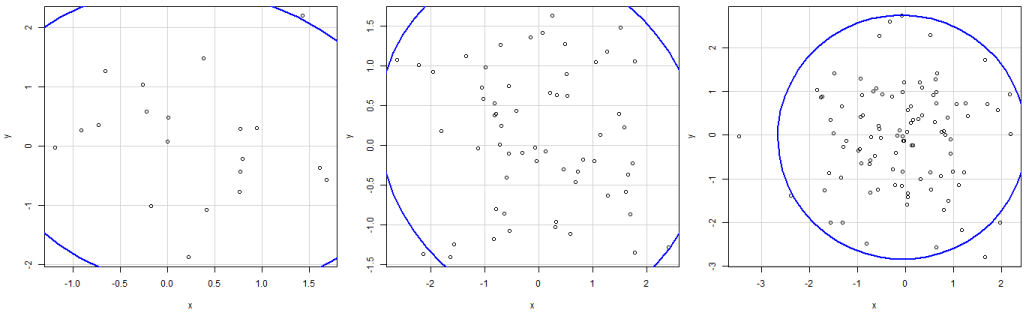
n = 100, смесь (1)	$\mathbb{E}z$	$\mathbb{E}z^2$	$\mathbb{D}z$
Знаковый (9)	0.836	0.71	0.007
Пирсона (6)	0.86	0.742	0.0014
Спирмена (10)	0.863	0.745	0.001

Таблица 13: Выборочные коэффициенты корреляции, n = 100, смесь (1)

4.1.2 Эллипсы рассеивания

Эллипсы рассеивания:

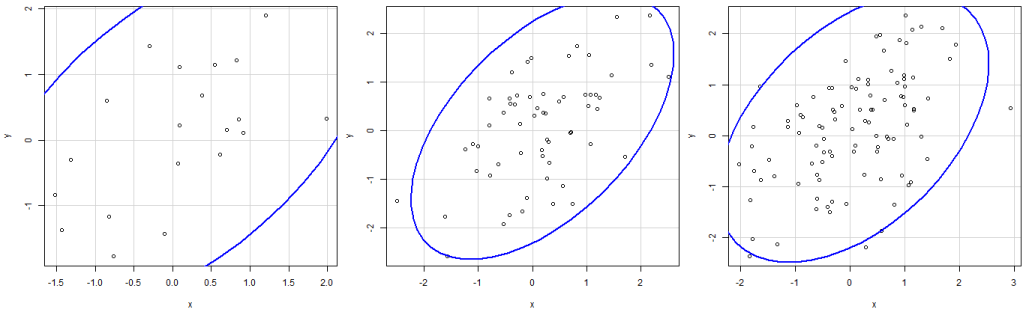
- $\rho = 0$: рис. 1



(a) n = 20 (b) n = 60 (c) n = 100

Рис. 1: Эллипсы рассеивания, $\rho = 0$

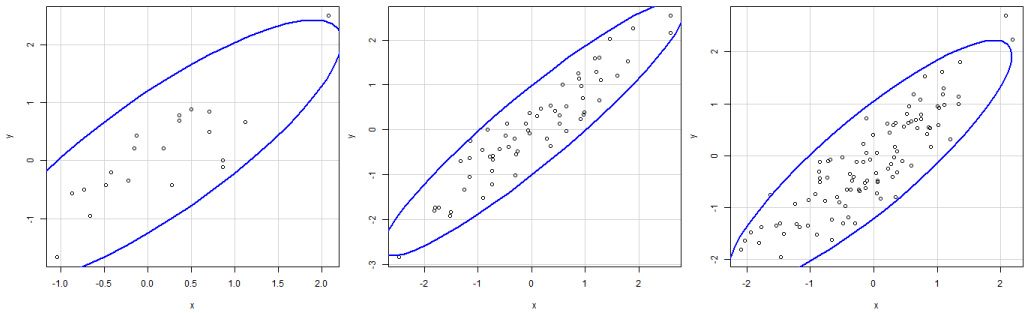
- $\rho = 0.5$: рис. 2



(a) n = 20 (b) n = 60 (c) n = 100

Рис. 2: Эллипсы рассеивания, $\rho = 0.5$

- $\rho = 0.9$: рис. 3



(a) n = 20 (b) n = 60 (c) n = 100

Рис. 3: Эллипсы рассеивания, $\rho = 0.9$

4.2 Линейная регрессия

Полученные аппроксимации коэффициентов представлены в таблице 14. Под выборкой с выбросами подразумевается выборка, где $y_1 := y_1 + 10$ и $y_{20} := y_{20} - 10$.

Метод		β_0	β_1
Исходная выборка	МНК	1.30	2
	МНМ	1.28	2
Выборка с выбросами	МНК	0.91	2
	МНМ	1.29	2

Таблица 14: Линейная регрессия

4.3 Оценка параметров нормального распределения

Полученные для выборки оценки приведены в таблице 15.

μ	σ^2
0.03	0.92

Таблица 15: Оценки методом максимального правдоподобия

Проверим гипотезу H_0 о том, что выборка взята из нормального распределения. Построим таблицу 16 для вычисления χ^2_B. Также по этому конкретному распределению построена гистограмма (рис. 4).

i	Δ_i	$\bar{p_i}$	n_i	$n\bar{p_i}$	$n_i - n\bar{p_i}$	$\frac{(n_i - n\bar{p_i})^2}{n\bar{p_i}}$
1	-2.0	0.017	1	1.654	-0.654	0.259
2	-1.5	0.044	6	4.406	1.594	0.577
3	-1.0	0.092	9	9.185	-0.185	0.004
4	-0.5	0.150	9	14.988	-5.988	2.392
5	0.0	0.191	25	19.146	5.854	1.790
6	0.5	0.191	23	19.146	3.854	0.776
7	1.0	0.150	13	14.988	-1.988	0.264
8	1.5	0.092	6	9.185	-3.185	1.104
9	2.0	0.044	4	4.406	-0.406	0.037
10	2.5	0.017	3	1.654	1.346	1.095
11	3.0	0.005	1	0.486	0.514	0.544
$\sum$	—	1	100	100	0	8.842

Таблица 16: Проверка гипотезы H_0 критерием χ^2

Из таблицы квантилей для $\alpha = 0.05$ находим $\chi^2_{1-\alpha}(r) = 19.68$. Полученное значение 8.842 меньше 19.68, поэтому гипотеза H_0 подтверждена.

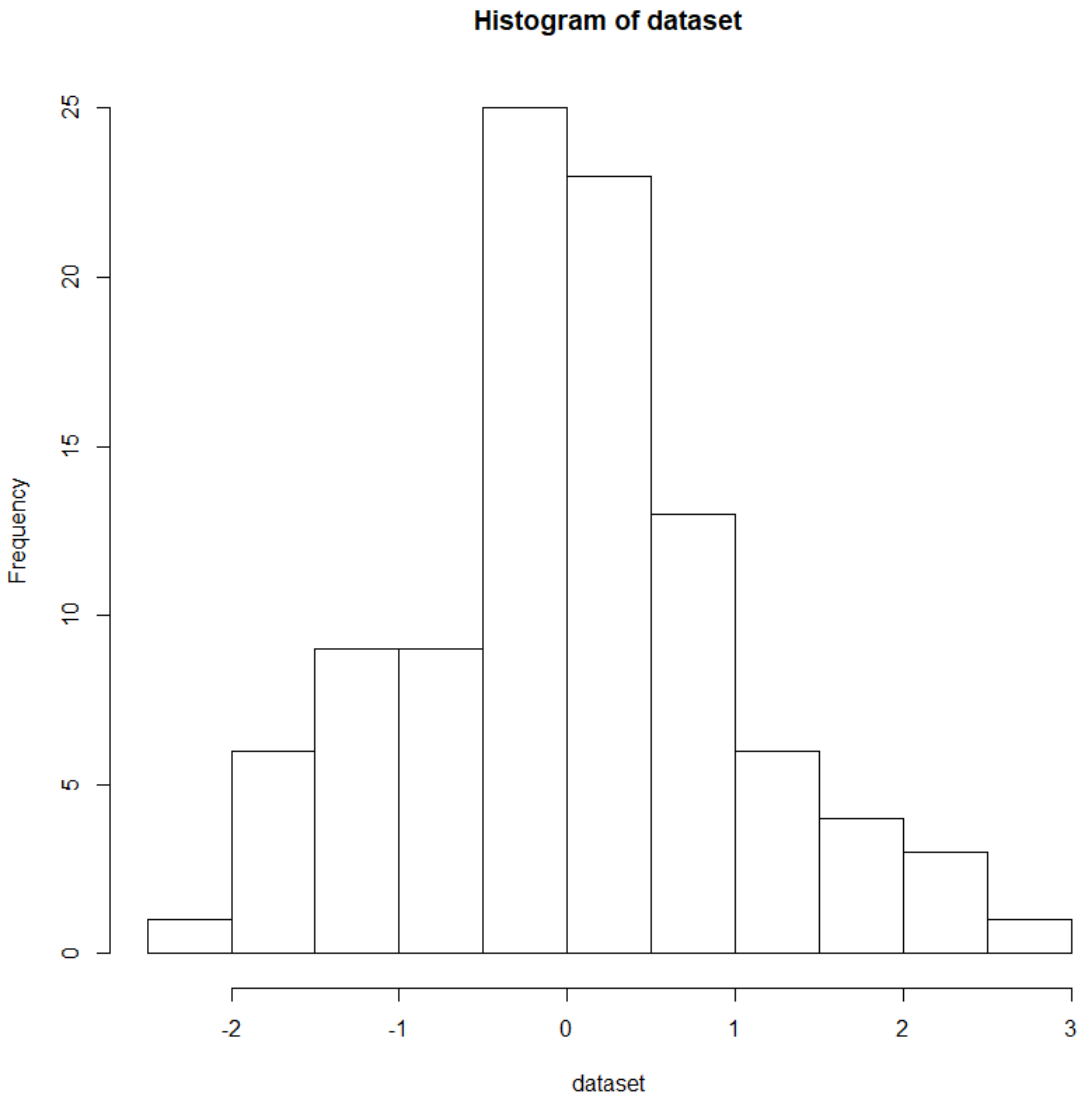


Рис. 4: Гистограмма распределения для оценки параметров

4.3.1 Анализ малой выборки из равномерного распределения

Проведём тот же эксперимент, сгенерировав выборку размером $n = 5$ не из нормального распределения, а из равномерного распределения $U(-2, 2)$. Гипотезу H_0 оставим прежней.

Полученные данные для оценки выборки приведены в таблице 17.

μ	σ^2
-0.52	2.57

Таблица 17: Оценки методом максимального правдоподобия 2

Проверим гипотезу H_0 для этой выборки тем же хи-квадрат критерием, построив таблицу 18.

i	Δ_i	$\bar{p_i}$	n_i	$n\bar{p_i}$	$n_i - n\bar{p_i}$	$\frac{(n_i - n\bar{p_i})^2}{n\bar{p_i}}$
1	-1	0.136	2	0.680	1.320	2.566
2	0	0.341	1	1.707	-0.707	0.292
3	1	0.341	1	1.707	-0.707	0.292
4	2	0.136	1	0.680	0.320	0.151
$\sum$	—	1	100	100	0	3.302

Таблица 18: Проверка гипотезы H_0 критерием χ^2

Для данной выборки $\chi^2_{1-\alpha}(1) = 3.84 > 3.302 = \chi^2_B \Rightarrow$ гипотеза H_0 принимается.

Гистограмма исследуемой выборки изображена на рис. 5.

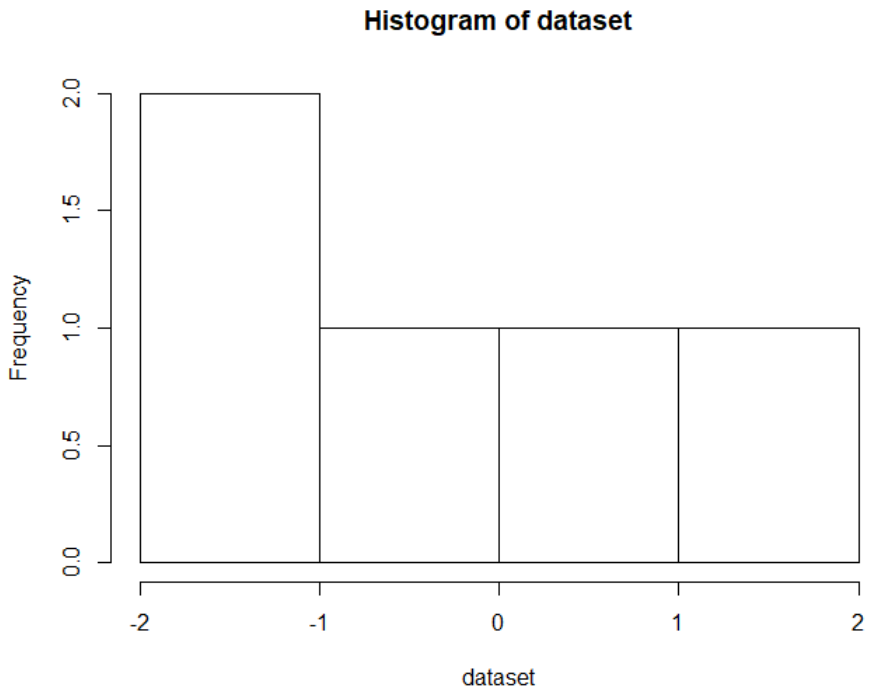


Рис. 5: Гистограмма распределения для оценки параметров

4.4 Интервальное оценивание

Результаты интервального оценивания приведены в таблице 19. В таблице параметр положения кратко обозначен как $\mathbb{E}$, а параметр масштаба как $\mathbb{D}$.

Параметр	Метод	n	Точечная оценка	Граница	
				Правая	Левая
$\mathbb{E}$	Классический	20	0.01	-0.19	0.20
		100	0.15	-0.05	0.35
	Асимптотический	20	0.01	-0.17	0.19
		100	0.15	-0.05	0.34
$\mathbb{D}$	Классический	20	0.41	0.32	0.62
		100	0.99	0.88	1.16
	Асимптотический	20	0.41	0.23	0.60
		100	0.99	0.79	1.20

Таблица 19: Интервальное оценивание

5 Обсуждение

5.1 Корреляция и эллипсы рассеивания

5.1.1 Оценка корреляции

Из результатов (таблица 1) следует, что для выборок исследуемых размеров оценки коэффициента корреляции в случае, когда истинный коэффициент корреляции $\rho = 0$, близки к нулю, но нулю не равны, у всех порядок $\approx 10^{-3}$. Однако, в этом случае дисперсия разворачивается в мат. ожидание квадрата случайной величины (другой член равен нулю, так как $\rho = 0 \Rightarrow K_{XY} = 0 \Rightarrow \mathrm{cov}(X, Y) = 0$), пользуемся формулой (4). Этим объясняется равенство столбцов $\mathbb{E}z^2$ и $\mathbb{D}z$.

Если обратить внимание на эксперименты, где $\rho \neq 0$, видно, что наименее точным является знаковый коэффициент корреляции, дальше идёт коэффициент Спирмена, наиболее точным является коэффициент Пирсона. Этот результат объясняется тем, что знаковый коэффициент корреляции при подсчёте наиболее прост, так как использует мало информации о выборке. Тем не менее, он обладает меньшей чувствительностью к выбросам, чем коэффициент Пирсона, так как в коэффициенте Пирсона используются выборочное среднее и выборочное стандартное отклонение — характеристики, не устойчивые к выбросам. В знаковом коэффициенте, если произойдёт выброс, это будет или не заметно (так как точка (x_i, y_i) может остаться в четверти, куда попало большинство других точек), или почти не заметно. Когда в выброке присутствует выброс, он может сильно исказить приближение эллипса рассеивания, построенного на основе выборки (выброс «растянет» полуось эллипса). Коэффициент Спирмена к выбросам устойчив, но имеет очень хорошую точность по сравнению со знаковым коэффициентом корреляции (так как строится на основе устойчивой к выбросам характеристике — медиане).

Коэффициент Пирсона в данном случае показывает большую точность, так как в нормальном распределении выбросы редки (их вероятность мала, пользуемся правилом «трёх сигм»).

Из экспериментов видно, что среднее коэффициентов корреляции уже достаточно хорошо приближает истинный коэффициент корреляции, начиная с выборок размером $n = 20$. При возрастании размера выборки сильнее уменьшается стандартное отклонение приближения (то есть, точность в некотором смысле всё равно увеличивается): чем больше выборка, тем больше вероятность оценить для неё коэффициент корреляции наиболее точно.

Касательно смеси нормальных распределений, на опыте оказалось,

что размер выборки сказывается сильнее: в случае $n = 20$ средние значения приближений между собой почти совпадают и далеки от результатов для $n = 60$ и $n = 100$.

В случае смеси распределений коэффициент Спирмена выигрывает в точности за счёт меньшего отклонения (оно на порядок меньше, чем у коэффициента Пирсона). Это следует из характера самого распределения — можно предположить, что «смесь» распределена нормально с коэффициентом корреляции $\rho \approx 0.9$ в общей массе, так как «вес» второго слагаемого с другой корреляцией мал, но есть вероятность получить «выброс», если наблюдение от первого слагаемого окажется малым, а от второго — большим; вероятность такого «выброса» больше, чем у выброса в нормальном распределении за счёт грубых весов.

5.1.2 Эллипсы рассеивания

По выборке можно построить приближение эллипса рассеивания. Из теории (формула (3)) следует, что по виду эллипса рассеивания можно судить о корреляции величин. В случае $\rho \approx 0$ построенные эллипсы будут близки к окружностям, корреляции нет: рис. 1. Когда корреляция есть, эллипс будет повёрнут и вытянут (рис. 2, 3).

Формула (3) подсказывает, что можно подобрать такие распределения для ρ, что полуоси эллипса будут близки, поэтому в некоторых случаях имеет смысл стандартизовать выборку перед построением выборочного эллипса рассеивания.

Из рассуждений выше следует, что о корреляции величин лучше судить по повороту выборочного эллипса рассеивания.

5.2 Линейная регрессия

Из анализа таблицы 14 следует, что оценка параметра β_1 (множителя при x) рассмотренными методами наиболее точна. Порядок дисперсии оценок β_1 при 1000 экспериментах составил 10^{-13}, что говорит о применимости методов для оценки этого параметра.

Оценки параметра положения β_0 уже гораздо менее точные. В выборки был добавлен нормальный шум, поэтому в результате 1000 повторов эксперимента дисперсия оценки β_0 составила примерно $0.93 \approx 1$ (параметр, соответствующий дисперсии шума). Делаем вывод, что шум серьёзно влияет на оценку параметра положения.

Однако, если «забыть» про шум и взять полученные результаты для исходной выборки за β_0, то при анализе выборки с выбросами видно, что МНК сильнее отреагировал на выбросы, а МНМ дал результат близкий

к полученному для выборки без выбросов. Это можно объяснить построением методов: в МНМ используются такие оценки положения как медиана, в отличие от МНК, где используется выборочное среднее. Таким образом, МНМ даёт более устойчивый результат.

5.3 Оценка параметров нормального распределения

5.3.1 Метод максимального правдоподобия

Метод максимального правдоподобия достаточно удобен в использовании. Он позволяет записать систему уравнений и решить её аналитически, чтобы получить выражения для оценок параметров. Часто в выражении плотности распределения есть экспонента, поэтому минимизация логарифма функции правдоподобия наиболее удобна.

Оценки, полученные методом максимального правдоподобия, являются состоятельными и асимптотически эффективными (но смещёнными в общем случае), что позволяет применять эти оценки на практике.

5.3.2 Метод хи-квадрат

С помощью достаточно удобного критерия хи-квадрат можно судить о характере генерального распределения. Сама проверка «состоятельности» гипотезы производится с помощью последовательности простых действий, заключающихся в построении суммы и её сравнения с табличным значением. Из результатов видно, что построение оцениваемой суммы эквивалентно разбору гистограммы распределения (можно построить некоторую аналогию: эксперт смотрит на диаграмму и по её характеру может подтвердить гипотезу).

Исследовав выборку из равномерного распределения размером $n = 5$, мы увидели, что гипотеза H_0 принимается для этой выборки согласно критерию хи-квадрат. Однако, при уровне значимости $\alpha = 0.1$ соответствующий квантиль хи-квадрат распределения равен $2.71 < \chi^2_B$. Это означает, что можно было взять другой уровень значимости α и получить для него другой ответ.

Таким образом, метод хи-квадрат применим при достаточно большом объёме выборки.

5.4 Интервальное оценивание

Проанализируем полученные оценки параметра положения.

Все полученные интервалы для параметра положения содержат в себе оцениваемый параметр.

При увеличении размера выборки в данном эксперименте увеличения точности оценки не происходит. Это может быть связано с конкретными выборками: в эксперименте получилось так, что у выборки меньшей мощности точечная оценка оказалась на порядок точнее, когда для выборки мощностью $n = 100$ эта оценка сместилась.

Если сравнить подходы к вычислению доверительных интервалов, можно заметить, что в случае $n = 20$ классический метод даёт чуть более меньший интервал, чем асимптотический, так как в классическом подходе используются допущения о том, что генеральное распределение известно (неизвестны лишь параметры). Асимптотический же подход стоит рассматривать, когда мощность выборки велика. В случае $n = 100$ доверительный интервал, полученный асимптотическим методом, оказался немного точнее.

Проанализируем полученные оценки параметра рассеяния.

Не все полученные интервалы содержат оцениваемый параметр. Генеральную дисперсию содержат лишь интервалы для выборок мощности $n = 100$.

При увеличении размера выборки произошло уменьшение интервала только для оценки, полученной классическим методом.

Классический метод дал меньший доверительный интервал для выборки мощностью $n = 20$, чем асимптотический метод для этой же выборки. Это объясняется характером асимптотического метода: он эффективен только для достаточно больших выборок. Даже на выборке $n = 100$ асимптотический метод дал больший интервал, что можно объяснить недостаточным объёмом выборки для применения асимптотического метода для наиболее точной оценки рассеяния. Из формулы (56) видно, что требуется вычислить такую характеристику, как выборочный эксцесс, он вычисляется с помощью четвёртых выборочных центральных моментов. Оценка генерального мат. недостаточно близка к истинному значению, поэтому в выборочном эксцессе накапливается эта неточность.

6 Литература

Список литературы

[1] Ю. Д. Максимов. “Вероятностные разделы математики”. в: *СПб.: Иван Федоров* (2001).

[2] Ю. Д. Максимов. “Математическая статистика”. в: *СПб.: СПбГПУ* (2004).