Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра «Прикладная математика»

Математическая статистика

Сводный отчет по лабораторным работам №5-8

Работу выполнил:
Колесник В.Н.
Группа:
3630102/70201
Преподаватель:
к.ф.-м.н., доцент
Баженов
Александр
Николаевич

Санкт-Петербург
2020

Содержание

Список иллюстраций

Список таблиц

1. Постановка задачи

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 $N(x, \widehat{\mu}, \widehat{\sigma})$. Проверить основную гипотезу, используя критерий согласия χ^2. В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений χ^2.

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. Двумерное нормальное распределение

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

$$N(x, y, \overline{x}, \overline{y}, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \quad (2)$$

$$\times exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x-\overline{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\overline{x})(y-\overline{y})}{\sigma_x\sigma_y} + \frac{(y-\overline{y})^2}{\sigma_y^2}]\} \quad (3)$$

Компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями x, y и средними квадратическими отклонениями σ_x, σ_y соответственно. Параметр ρ называется коэффициентом корреляции.

2.2. Корреляционный момент (ковариация) и коэффи- циент корреляции

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

$$K = cov(X, Y) = M[(X - \overline{x})(Y - \overline{y})] \tag{4}$$

Коэффициентом корреляции ρ двух случайных величин X и Y называется отношение их корреляционного момента к произведению их средних квадратических отклонений:

$$\rho = \frac{K}{\sigma_x \sigma_y} \tag{5}$$

Коэффициент корреляции — это нормированная числовая характеристика, являющаяся мерой близости зависимости между случайными величинами к линейной.

2.3. Выборочные коэффициенты корреляции

2.3.1. Выборочный коэффициент корреляции Пирсона

Пусть по выборке значений $\{x_i, y_i\}_1^n$ двумерной с.в. (X, Y) требуется оценить коэффициент корреляции $\rho = \frac{cov(X,Y)}{\sqrt{DXDY}}$. Естественной оценкой для ρ служит его статистический аналог в виде выборочного коэффициента корреляции, предложенного К.Пирсоном, —

$$r = \frac{\frac{1}{n}\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \overline{x})^2 \frac{1}{n}\Sigma(y_i - \overline{y})^2}} = \frac{K}{s_X s_Y} \tag{6}$$

где K, s_X^2, s_Y^2 — выборочные ковариация и дисперсии с.в. X и Y.

2.3.2. Выборочный квадрантный коэффициент корреляции

Кроме выборочного коэффициента корреляции Пирсона, существуют и другие оценки степени взаимосвязи между случайными величинами. К ним относится выборочный квадрантный коэффициент корреляции

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \tag{7}$$

где n_1, n_2, n_3 и n_4 — количества точке с координатами (x_i, y_i), попавшими соответственно в I, II, III и IV квадранты декартовой системы с осями $x' = x - medx, y' = y - medy$ и с центром в точке с координатами $(medx, medy)$.

2.3.3. Выборочный коэффициент ранговой корреляции Спирмена

На практике нередко требуется оценить степень взаимодействия между качественными признаками изучаемого объекта. Качественным называется признак, который нельзя измерить точно, но который позволяет сравнивать изучаемые объекты между собой и располагать их в порядке убывания или возрастания их качества. Для этого объекты выстраиваются в определённом порядке в соответствии с рассматриваемым признаком. Процесс упорядочения называется ранжированием, и каждому члену упорядоченной последовательности объектов присваивается ранг, или порядковый номер. Например, объекту с наименьшим значением признака присваивается ранг 1, следующему за ним объекту

— ранг 2, и т.д. Таким образом, происходит сравнение каждого объекта со всеми объектами изучаемой выборки.
Если объект обладает не одним, а двумя качественными признаками — переменными X и Y, то для исследования их взаимосвязи используют выборочный коэффициент корреляции между двумя последовательностями рангов этих признаков.
Обозначим ранги, соотвествующие значениям переменной X, через u, а ранги, соотвествующие значениям переменной Y, — через v.
Выборочный коэффициент ранговой корреляции Спирмена определяется как выборочный коэффициент корреляции Пирсона между рангами u, v переменных X, Y:

$$r_S = \frac{\frac{1}{n}\Sigma(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\frac{1}{n}\Sigma(u_i - \overline{u})^2 \frac{1}{n}\Sigma(v_i - \overline{v})^2}} = \frac{K}{s_X s_Y} \tag{8}$$

где $\overline{u} = \overline{v} = \frac{1+2+...+n}{n} = \frac{n+1}{2}$ - среднее значение рангов.

2.3.4. Эллипсы рассеивания

Рассмотрим поверхность распределения, изображающую функцию (1). Она имеет вид холма, вершина которого находится над точкой (x, y).
В сечении поверхности распределения плоскостями, параллельными оси $N(x, y, \overline{x}, \overline{y}, \sigma_x, \sigma_y, \rho)$, получаются кривые, подобные нормальным кривым распределения. В сечении поверхности распределения плоскостями, параллельными плоскости xOy, получаются эллипсы. Напишем уравнение проекции такого эллипса на плоскость xOy:

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

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

$$tg2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \tag{10}$$

Это уравнение дает два значения углов: α и α_1, различающиеся на $\frac{\pi}{2}$.
Таким образом, ориентация эллипса относительно координатных осей находится в прямой зависимости от коэффициента корреляции ρ системы (X, Y); если величины не коррелированны (т.е. в данном случае и независимы), то оси симметрии эллипса параллельны координатным осям; в противном случае они составляют с координатными осями некоторый угол.
Пересекая поверхность распределения плоскостями, параллельными плоскости xOy, и проектируя сечения на плоскость xOy мы получим целое семейство подобных и одинаково расположенных эллипсов с общим центром (x, y). Во всех точках каждого из таких эллипсов плотность распределения $N(x, y, \overline{x}, \overline{y}, \sigma_x, \sigma_y, \rho)$ постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче эллипсами рассеивания. Общие оси всех эллипсов рассеивания называются главными осями рассеивания.

2.4. Простая линейная регрессия

2.4.1. Модель простой линейной регрессии

Регрессионную модель описания данных называют простой линейной регрессией, если

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

где $x_1, ..., x_n$ — заданные числа (значения фактора); $y_1, ..., y_n$ — наблюдаемые значения отклика; $\varepsilon_1, ..., \varepsilon_n$ — независимые, нормально распределённые $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1 — неизвестные параметры, подлежащие оцениванию.
В модели (11) отклик y зависит 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.

2.4.2. Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распрстранённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются 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):

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

Задача минимизации квадратичного критерия (12) носит название задачи метода наименьших квадратов (МНК), а оценки $\widehat{\beta_0}, \widehat{\beta_1}$ параметров β_0, β_1, реализующие минимум критерия (12), называют МНК-оценками.

2.4.3. Расчётные формулы для МНК-оценок

МНК-оценки параметров $\widehat{\beta_0}$ и $\widehat{\beta_1}$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум. Они равны:

$$\widehat{\beta_1} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{13}$$

$$\widehat{\beta_0} = \overline{y} - \overline{x}\widehat{\beta_1} \tag{14}$$

2.5. Робастные оценки коэффициентов линейной регрессии

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

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

Напомним, что использование метода наименьших модулей в задаче оценивания параметра сдвига распределений приводит к оценке в виде выборочной медианы, обладающей робастными свойствами. В отличие от этого случая и от задач метода наименьших квадратов, на практике задача (15) решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу. Здесь мы рассмотрим простейшую в вычистлительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок (13) и (14) в другом виде:

$$\widehat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y} \frac{s_y}{s_x} = r_{xy} \frac{s_y}{s_x}, \tag{16}$$

$$\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1 \tag{17}$$

В формулах (16) и (17) заменим выборочные средние x и y соответственно на робастные выборочные медианы med x и med y, среднеквадратические отклонения s_x и s_y на робастные нормированные интерквартильные широты q_x^* и q_y^*, выборочный коэффициент корреляции r_{xy} — на знаковый коэффициент корреляции r_Q:

$$\widehat{\beta_{1R}} = r_Q \frac{q_y^*}{q_x^*}, \tag{18}$$

$$\widehat{\beta_{0R}} = medy - \widehat{\beta_{1R}} medx, \tag{19}$$

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - medx) sgn(y_i - medy), \tag{20}$$

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)}, q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)}. \tag{21}$$

$$l = \begin{cases} [n/4] + 1 \text{ при } n/4 \text{ дробном}, \\ n/4 \text{ при } n/4 \text{ целом}. \end{cases} \tag{22}$$

$$j = n - l + 1 \tag{23}$$

$$sgnz = \begin{cases} 1 \text{ при } z > 0 \\ 0 \text{ при } z = 0 \\ -1 \text{ при } z < 0 \end{cases} \tag{24}$$

Уравнение регрессии здесь имеет вид

$$y = \widehat{\beta_{0R}} + \widehat{\beta_{1R}} x. \tag{25}$$

Статистики выборочной медианы и интерквартильной широты обладают робастными свойствами в силу того, что основаны на центральных порядковых статистиках, малочувствительных к большим по величине выбросам в данных. Статистика выборочного знакового коэффициента корреляции робастна, так как знаковая функция sgn z чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии (25) обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба.

2.6. Метод максимального правдоподобия

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

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

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

$$\widehat{\theta}_{\text{мп}} = \arg \max_{\theta} L(x_1, ..., x_n, \theta) \tag{27}$$

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

$$\frac{\partial L(x_1, ..., x_n, \theta)}{\partial \theta} = 0 \tag{28}$$

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

$$\frac{\partial^2 L(x_1, ..., x_n, \tilde{\theta})}{\partial \theta^2} < 0 \tag{29}$$

Определив точки локальных максимумов ФП (если их несколько), находят наибольший, который и даёт решение задачи (26).
Часто проще искать максимум логарифма ФП, так как он имеет максимум в одной точке с ФП:

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{L} \frac{\partial L}{\partial \theta}, \text{ если } L > 0, \tag{30}$$

и соответственно решать уравнение

$$\frac{\partial \ln L}{\partial \theta} = 0, \tag{31}$$

которое называют *уравнением правдоподобия.*
В задаче оценивания векторного параметра $\theta = (\theta_1, ...\theta_m)$ аналогично (27) находится максимум ФП нескольких аргументов:

$$\widehat{\theta}_{\text{мп}} = \arg \max_{\theta_1, ...\theta_m} L(x_1, ..., x_n, \theta_1, ...\theta_m) \tag{32}$$

и в случае дифференцируемости ФП выписывается система уравнений правдоподобия

$$\frac{\partial L}{\partial \theta_k} = 0 \text{ или } \frac{\partial \ln L}{\partial \theta_k} = 0, k = 1, ..., m. \tag{33}$$

2.7. Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Исчерпывающей характеристикой изучаемой случайной величины является её закон распределения. Поэтому естественно стремление исследователей построить этот закон приближённо на основе статистических данных.

Сначала выдвигается гипотеза о виде закона распределения.

После того как выбран вид закона, возникает задача оценивания его параметров и проверки (тестирования) закона в целом.

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

Мы ограничимся рассмотрением случая одномерного распределения.

Итак, выдвинута гипотеза H_0 о генеральном законе распределения с функцией распределения $F(x)$.

Рассматриваем случай, когда гипотетическая функция распределения $F(x)$ не содержит неизвестных параметров.

Разобьём генеральную совокупность, т.е. множество значений изучаемой случайной величины X на k непересекающихся подмножеств $\Delta_1, \Delta_2, ..., \Delta_k$.

Пусть $p_i = P(X \in \Delta_i), i = 1, ..., k$.

Если генеральная совокупность — вся вещественная ось, то подмножества $\Delta_i = (a_{i-1}, a_i]$ — полуоткрытые промежутки $(i = 2, ..., k-1)$. Крайние промежутки будут полубесконечными: $\Delta_i = (-\infty, a_1], \Delta_k = (a_{k-1}, +\infty)$. В этом случае $p_i = F(a_i) - F(a_{i-1})$; $a_0 = -\infty, a_k = +\infty (i = 1, ..., k)$.

Отметим, что $\sum_{i=1}^{k} p_i = 1$. Будем предполагать, что все $p_i > 0 (i = 1, ..., k)$.

Пусть, далее, $n_1, n_2, ..., n_k$ — частоты попадания выборочных элементов в подмножества $\Delta_1, \Delta_2, ..., \Delta_k$ соответственно.

В случае справедливости гипотезы H_0 относительные частоты n_i/n при большом n должны быть близки к вероятностям $p_i (i = 1, ..., k)$, поэтому за меру отклонения выборочного распределения от гипотетического с функцией $F(x)$ естественно выбрать величину

$$Z = \sum_{i=1}^{k} c_i (\frac{n_i}{n} - p_i)^2, \tag{34}$$

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

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i} (\frac{n_i}{n} - p_i)^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \tag{35}$$

которая обозначена тем же символом, что и закон распределения хи-квадрат.

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

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

2.8. Доверительные интервалы для параметров нормального распределения

2.8.1. Доверительный интервал для математического ожидания m нормального распределения

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

Доказано, что случайная величина

$$T = \sqrt{n-1}\frac{\overline{x}-m}{s}, \tag{36}$$

называемая статистикой Стьюдента, распределена по закону Стьюдента с $n-1$ степенями свободы.
Пусть $f_T(x)$ — плотность вероятности этого распределения.
Пусть $t_{1-\alpha/2}(n-1)$ — квантиль распределения Стьюдента с $n-1$ степенями свободы и порядка $1-\alpha/2$.
Тогда доверительный интервал для m с доверительной вероятностью $\gamma = 1-\alpha$ можно получить из равенства:

$$P\left(\overline{x} - \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}} < m < \overline{x} + \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n-1}}\right) = 1-\alpha. \tag{37}$$

2.8.2. Доверительный интервал для среднего квадратического отклонения σ нормального распределения

Дана выборка $(x_1, x_2, ..., x_n)$ 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.
Тогда доверительный интервал для σ с доверительной вероятностью $\gamma = 1-\alpha$ можно получить из равенства:

$$P\left(\frac{s\sqrt{n}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} < \sigma < \frac{s\sqrt{n}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right) = 1-\alpha. \tag{38}$$

2.9. Доверительные интервалы для математического ожидания m и среднего квадратического отклонения σ произвольного распределения при большом объёме выборки. Асимптотический подход

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

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

Выборочное среднее $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ 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.
Пусть $u_{1-\alpha/2}$ — квантиль нормального распределения $N(x, 0, 1)$ порядка $1 - \alpha/2$.
Тогда доверительный интервал для m с доверительной вероятностью $\gamma = 1 - \alpha$ можно получить из равенства:

$$P\left(\overline{x} - \frac{su_{1-\alpha/2}}{\sqrt{n}} < m < \overline{x} + \frac{su_{1-\alpha/2}}{\sqrt{n}}\right) \approx \gamma, \tag{39}$$

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

Выборочная дисперсия $s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}$ 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.
Пусть $m_4 = \frac{1}{n} \sum_{i=1}^{n} (x_i = \overline{x})^4$ — четвёртый выборочный центральный момент. Тогда доверительный интервал для σ с доверительной вероятностью $\gamma = 1 - \alpha$ можно получить из равенства

$$s(1 + U)^{-1/2} < \sigma < s(1 - U)^{-1/2}, \tag{40}$$

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

$$s(1 - 0.5U) < \sigma < s(1 + 0.5U). \tag{41}$$

3. Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования R в среде разработки RStudio. Исходный код лабораторной работы приведён в приложении.

4. Результаты

4.1. Выборочные коэффициенты корреляции

$\rho = 0$	r	r_S	r_Q
$E(z)$	-0.005	-0.005	-0.008
$E(z^2)$	0.057	0.056	0.050
$D(z)$	0.057	0.056	0.050
$\rho = 0.5$	r	r_S	r_Q
$E(z)$	0.489	0.462	0.331
$E(z^2)$	0.270	0.249	0.156
$D(z)$	0.031	0.036	0.046
$\rho = 0.9$	r	r_S	r_Q
$E(z)$	0.896	0.867	0.711
$E(z^2)$	0.805	0.756	0.529
$D(z)$	0.002	0.004	0.024

Таблица 4.1: Двумерное нормальное распределение, $n = 20$

$\rho = 0$	r	r_S	r_Q
$E(z)$	-0.004	-0.005	-0.007
$E(z^2)$	0.017	0.017	0.016
$D(z)$	0.017	0.017	0.016
$\rho = 0.5$	r	r_S	r_Q
$E(z)$	0.503	0.483	0.338
$E(z^2)$	0.262	0.244	0.128
$D(z)$	0.009	0.010	0.014
$\rho = 0.9$	r	r_S	r_Q
$E(z)$	0.897	0.881	0.711
$E(z^2)$	0.806	0.777	0.513
$D(z)$	0.0006	0.001	0.008

Таблица 4.2: Двумерное нормальное распределение, $n = 60$

$\rho = 0$	r	r_S	r_Q
$E(z)$	0.004	0.004	0.004
$E(z^2)$	0.010	0.010	0.010
$D(z)$	0.010	0.010	0.010
$\rho = 0.5$	r	r_S	r_Q
$E(z)$	0.493	0.473	0.329
$E(z^2)$	0.248	0.230	0.117
$D(z)$	0.006	0.006	0.008
$\rho = 0.9$	r	r_S	r_Q
$E(z)$	0.898	0.885	0.710
$E(z^2)$	0.808	0.784	0.509
$D(z)$	0.0003	0.0006	0.004

Таблица 4.3: Двумерное нормальное распределение, $n = 100$

$n = 20$	r	r_S	r_Q
$E(z)$	0.785	0.753	0.592
$E(z^2)$	0.625	0.579	0.381
$D(z)$	0.008	0.011	0.030
$n = 60$	r	r_S	r_Q
$E(z)$	0.787	0.766	0.580
$E(z^2)$	0.622	0.590	0.347
$D(z)$	0.002	0.003	0.010
$n = 100$	r	r_S	r_Q
$E(z)$	0.790	0.771	0.579
$E(z^2)$	0.625	0.596	0.342
$D(z)$	0.001	0.001	0.006

Таблица 4.4: Смесь нормальных распределений

4.2. Эллипсы рассеивания

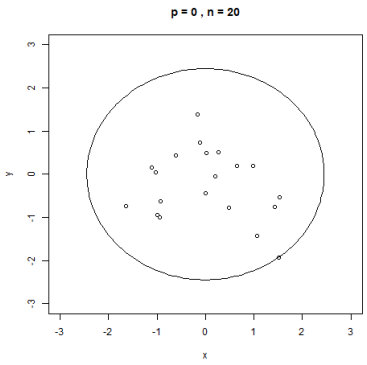

Рисунок 4.1. $\rho = 0, n = 20$

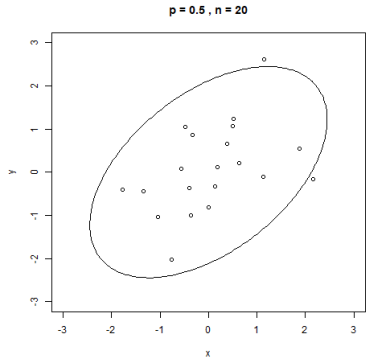

Рисунок 4.2. $\rho = 0.5, n = 20$

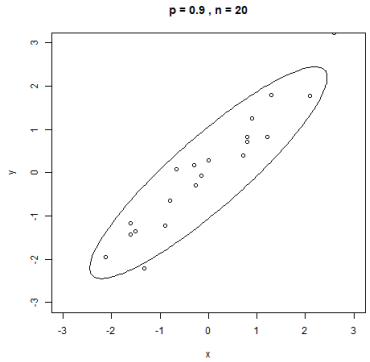

Рисунок 4.3. $\rho = 0.9, n = 20$

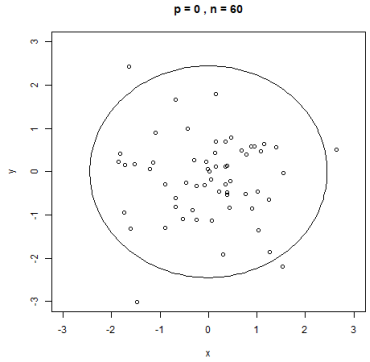

Рисунок 4.4. $\rho = 0, n = 60$

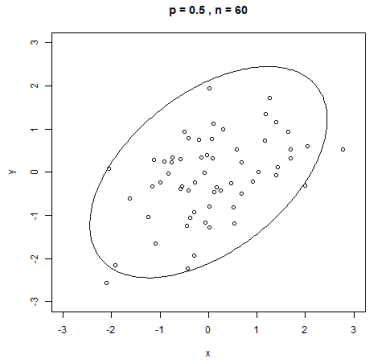

Рисунок 4.5. $\rho = 0.5, n = 60$

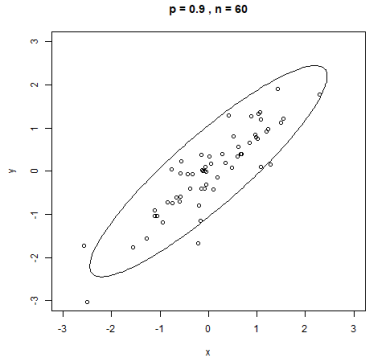

Рисунок 4.6. $\rho = 0.9, n = 60$

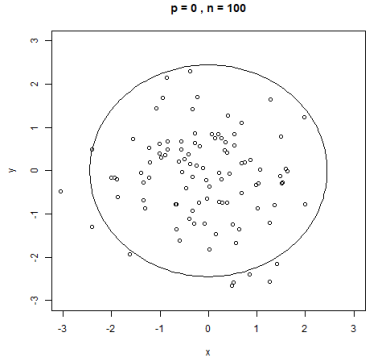

Рисунок 4.7. $\rho = 0, n = 100$

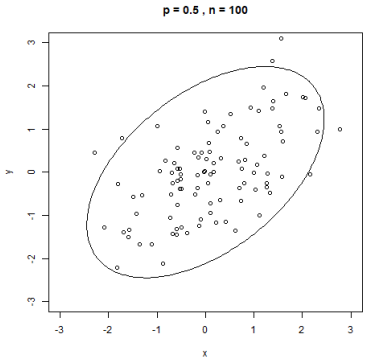

Рисунок 4.8. $\rho = 0.5, n = 100$

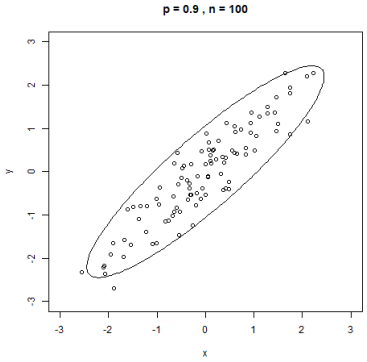

Рисунок 4.9. $\rho = 0.9, n = 100$

4.3. Эллипсы рассеивания для выборки n=3

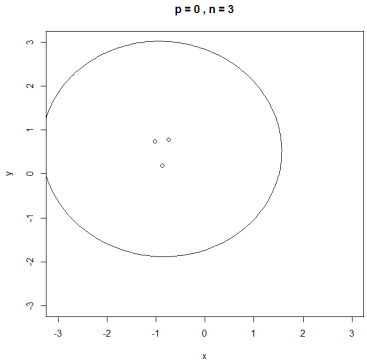

Рисунок 4.10. $\rho = 0, n = 3$

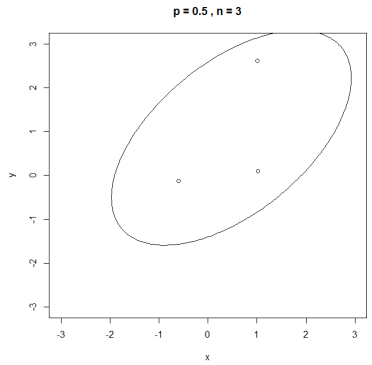

Рисунок 4.11. $\rho = 0.5, n = 3$

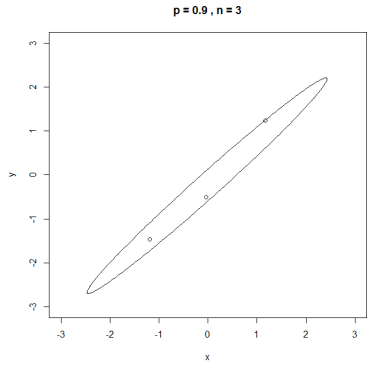

Рисунок 4.12. $\rho = 0.9, n = 3$

4.4. Оценки коэффициентов линейной регрессии

4.4.1. Выборка без возмущений

Критерий наименьших квадратов:

$$\widehat{a} \approx 1.92 \tag{42}$$

$$\widehat{b} \approx 1.89 \tag{43}$$

Критерий наименьших модулей:

$$\widehat{a} \approx 1.71 \tag{44}$$

$$\widehat{b} \approx 1.92 \tag{45}$$

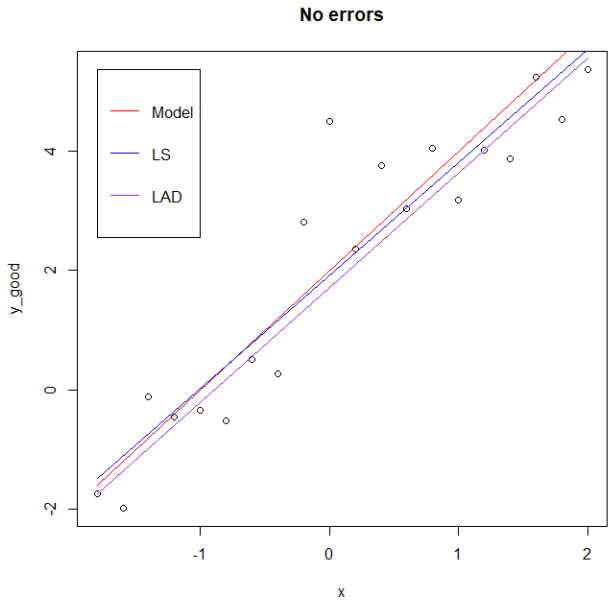

Рисунок 4.13. Без возмущений

4.4.2. Выборка с возмущением

Критерий наименьших квадратов:

$$\widehat{a} \approx 2.07 \tag{46}$$

$$\widehat{b} \approx 0.46 \tag{47}$$

Критерий наименьших модулей:

$$\widehat{a} \approx 1.78 \tag{48}$$

$$\widehat{b} \approx 1.86 \tag{49}$$

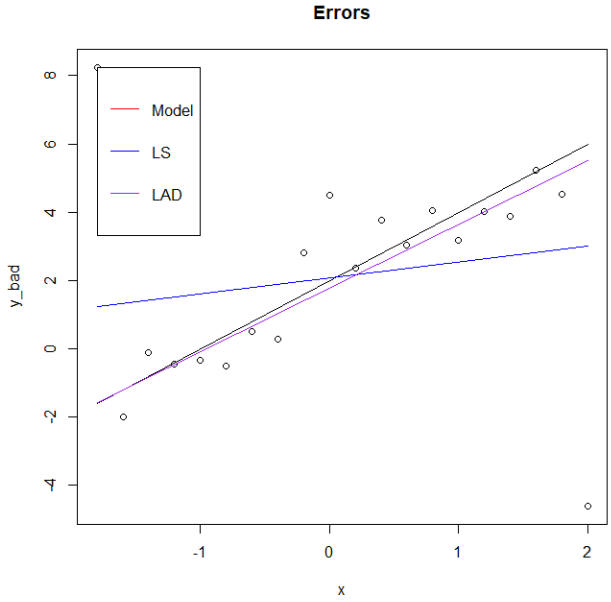

Рисунок 4.14. С возмущениями

4.5. Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

4.5.1. Нормальное распределение

Метод максимального правдоподобия:

$$\widehat{\mu} \approx -0.16 \tag{50}$$

$$\widehat{\sigma} \approx 0.86 \tag{51}$$

Критерий согласия χ^2:

i	a_{i-1}, a_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
1	$-\infty$, -2	2	0.015	1.599	0.411	0.106
2	-2, -1.5	8	0.043	4.319	3.680	3.136
3	-1.5, -1	11	0.105	10.499	0.500	0.023
4	-1, -0.5	12	0.183	18.303	-6.303	2.171
5	-0.5, 0	24	0.229	22.885	1.114	0.054
6	0, 0.5	20	0.205	20.524	-0.524	0.013
7	0.5, 1	12	0.132	13.203	-1.203	0.109
8	1, 1.5	7	0.061	6.091	0.908	0.135
9	1.5, 2	3	0.020	2.014	0.985	0.481
10	2, $+\infty$	1	0.006	0.569	0.430	0.325
$\sum$	-	100	1	100	0	χ^2_B=6.558

Таблица 4.5: Вычисление χ^2 при проверке гипотезы H_0 о нормальном законе распределения $N(x, \widehat{\mu}, \widehat{\sigma})$

Количество промежутков $k = 10$.

Уровень значимости $\alpha = 0.05$.

Тогда квантиль $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(9) \approx 16.918$.

Сравнивая $\chi^2_B = 6.558$ и $\chi^2_{0.95}(9) \approx 16.918$, видим, что $\chi^2_B < \chi^2_{0.95}(9)$.

4.5.2. Распределение Лапласа

Метод максимального правдоподобия:

$$\widehat{\mu} \approx -0.35 \tag{52}$$

$$\widehat{\sigma} \approx 2.92 \tag{53}$$

Критерий согласия χ^2:

i	a_{i-1}, a_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
1	$-\infty$, -4	1	0.10	3.20	-2.20	1.51
2	-4, -2	2	0.18	5.42	-3.42	2.15
3	-2, 0	14	0.26	7.83	6.16	4.85
4	0, 2	11	0.24	7.22	3.77	1.96
5	2, $+\infty$	2	0.21	6.31	-4.31	2.94
$\sum$	-	30	1	30	0	χ^2_B=13.44

Таблица 4.6: Вычисление χ^2 при проверке гипотезы H_1 о нормальности закона распределения Лапласа

Количество промежутков $k = 5$.

Уровень значимости $\alpha = 0.05$.

Тогда квантиль $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(4) \approx 9.49$.

Сравнивая $\chi^2_B = 13.44$ и $\chi^2_{0.95}(4) \approx 9.49$, видим, что $\chi^2_B > \chi^2_{0.95}(4)$.

4.6. Доверительные интервалы для параметров нормального распределения

Значения выборочного среднего и выборочного среднего квадратического равны:

- $m = 0.03$, $\sigma = 0.99$ при $n = 20$
- $m = 0.12$, $\sigma = 1.03$ при $n = 100$

4.7. Доверительные интервалы для параметров произвольного распределения. Асимптотический подход

Значения выборочного среднего и выборочного среднего квадратического равны:

- $m = 0.03$, $\sigma = 0.99$ при $n = 20$
- $m = 0.12$, $\sigma = 1.03$ при $n = 100$

$n = 20$	m	σ
	$-0.44 < m < 0.47$	$0.77 < \sigma < 1.48$
$n = 100$	m	σ
	$-0.08 < m < 0.32$	$0.91 < \sigma < 1.20$

Таблица 4.7: Доверительные интервалы для параметров нормального распределения

$n = 20$	m	σ
	$-0.40 < m < 0.47$	$0.44 < \sigma < 1.53$
$n = 100$	m	σ
	$-0.08 < m < 0.32$	$0.79 < \sigma < 1.27$

Таблица 4.8: Доверительные интервалы для параметров произвольного распределения. Асимптотический подход

5. Обсуждение

5.1. Выборочные коэффициенты корреляции и эллипсы рассеивания

Сравним дисперсии выборочных коэффициентов корреляции.
Для двумерного нормального распределения дисперсии выборочных коэффициентов корреляции упорядочены следующим образом: $r < r_S < r_Q$.
Для смеси нормальных распределений дисперсии выборочных коэффициентов корреляции упорядочены следующим образом: $r < r_S < r_Q$.
Процент попавших элементов выборки в эллипс рассеивания (95%-ная доверительная область) примерно равен его теоретическому значению (95%).

5.2. Оценки коэффициентов линейной регрессии

Рассмотрим пары чисел (a, b) и $(\widehat{a}, \widehat{b})$ как двумерные векторы. Оценим отклонение полученных результатов от эталонных по норме разности для выборки без возмущений. Для критерия наименьших квадратов по бесконечной норме получим 0.11, для критерия наименьших модулей - 0.29. Значит, критерий наименьших квадратов точнее оценивает коэффициенты линейной регрессии на выборке без возмущений.
Критерий наименьших модулей точнее оценивает коэффициенты линейной регрессии на выборке с возмущениями.
Критерий наименьших модулей устойчив к редким крупным выбросам.

5.3. Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Заключаем, что гипотеза H_0 о нормальном законе распределения $N(x, \widehat{\mu}, \widehat{\sigma})$, на уровне значимости $\alpha = 0.05$, согласуется с выборкой.
Гипотеза о нормальности закона распределения Лапласа не согласуется с полученной выборкой.

5.4. Доверительные интервалы для параметров распределения

- Генеральные характеристики ($m = 0$ и $\sigma = 1$) накрываются построенными доверительными интервалами
- Доверительные интервалы, полученные по большей выборке, являются соответственно более точными, т.е. меньшими по длине
- Доверительные интервалы для параметров нормального распределения более надёжны (меньше по длине), так как основаны на точном, а не асимптотическом распределении

6. Приложения

Репозиторий на Github с кодом лабораторной работы:
`https://github.com/VsevolodMelnikov/Math_Stat/tree/master`