Санкт-Петербургский политехнический университет имени Петра Великого
Физико-Механический институт

Лабораторная работа №2

по дисциплине

Математическая статистика

Выполнил студент группы 5030102/00101 Маковеев Лев
Преподаватель Баженов Александр Николаевич

Санкт-Петербург, 2023

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:

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

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

   Исследовать точность (чувствительность) критерия χ^2 - сгенерировать выборки равномерного распределения и распределения Лапласа малого объема (например, 20 элементов). Проверить их на нормальность.

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 Выборочные коэффициенты корреляции

2.1.1 Выборочный коэффициент корреляции Пирсона

Пусть по выборке значений $\{x_i, y_i\}$, двумерной с.в. (X, Y) требуется оценить коэффициент корреляции ρ. Естественной оценкой для ρ служит его статистический аналог в виде выборочного коэффициента корреляции, предложенного К.Пирсоном:

$$r_P = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{K}{s_X s_Y}$$

где K, s_X, s_Y - выборочные ковариация и дисперсии с.в. X и Y.

2.1.2 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, соответствующие значениям переменной Y, через v.

Выборочный коэффициент ранговой корреляции Спирмена определяется как выборочный коэффициент корреляции Пирсона между рангами u, v переменных X, Y:

$$r_S = \frac{\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum(u_i - \bar{u})^2 \sum(v_i - \bar{v})^2}} = \frac{K}{s_X s_Y}$$

2.1.3 Выборочный квадрантный коэффициент корреляции

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

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

2.2 Эллипсы рассеивания

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

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

2.3 Методы наименьших квадратов и модулей

Пусть имеется выборка $\{x_i, y_i\}$, функция $Q(x, \theta)$, где θ - неизвестный параметр (или вектор). Требуется найти такой θ^*, чтобы $Q(x, \theta)$ максимально приближала выборку $\{x_i, y_i\}$.

В методе наименьших квадратов θ^* находится из соображения, что θ^* даёт минимум функционалу

$$S_{sq}(\theta) = \sum_{i=1}^{n}(y_i - Q(x_i, \theta))^2$$

В методе наименьших модулей θ^* находится из соображения, что θ^* даёт минимум функционалу

$$S_{sq}(\theta) = \sum_{i=1}^{n}|y_i - Q(x_i, \theta)|$$

2.4 Метод максимального правдоподобия

Пусть $\{x_i\}_{i=1}^n$ - выборка из нормального распределения $N(a, \sigma^2)$. Оба параметра a, σ неизвестны.

Функция правдоподобия нормального распределения:

$$L(a, \sigma^2) = \prod_{i=1}^{n} f_{N(a,\sigma^2)}(x_i) = \frac{1}{(2\pi\sigma)^{n/2}}\exp\left(-\frac{\sum\limits_{i=1}^{n}(x_i - a)^2}{2\sigma^2}\right)$$

Необходимо найти (a_*, σ_*^2) такие, чтобы $L(a_*, \sigma_*^2) = \underset{(a,\sigma^2)}{max} L(a, \sigma^2)$.

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

Пусть $\{x_i\}_{i=1}^n$ - выборка. H_0 - выдвинутая гипотеза о законе распределения с функцией распределения $F(x)$.

Разобьём генеральную совокупность, т.е. множество значений изучаемой случайной величины X на k непересекающихся подмножеств $\Delta_1, \Delta_2, ..., \Delta_k$, где $\Delta_i = (a_{i-1}, a_i]$. Построим $p_i = F(a_i) - F(a_{i-1})$.

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

Статистика критерия согласия χ^2 Пирсона определяется соотношением

$$\chi^2 = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i}$$

Если получившееся значения статистики меньше, чем табличное значение распределения χ^2 с $k-1$ степенями свободы степени $1-\alpha$, где α - уровень значимости выставляемой гипотезы, то гипотеза H_0 принимается. Если больше - гипотеза H_0 отвергается.

2.6 Доверительные интервалы

2.6.1 Доверительный интервал для математического ожидания Стьюдента

Пусть $\{x_i\}_{i=1}^n$ - выборка из нормальной генеральной совокупности. На её основе строим выборочное среднее $\bar{x}$ и выборочное среднее квадратичное отклонение s. Пусть $t_{1-\alpha/2}(n-1)$ - квантиль распределения Стьюдента с $n-1$ степенями свободы и порядка $1-\alpha/2$. Тогда:

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

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

2.6.2 Доверительный интервал для среднего квадратического отклонения

Пусть $\{x_i\}_{i=1}^n$ - выборка из нормальной генеральной совокупности. На её основе строим выборочное среднее $\bar{x}$ и выборочное среднее квадратичное отклонение s. Тогда

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

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

3 Результаты

3.1 Выборочные коэффициенты корреляции

ρ	$E(r_P)$	$E(r_P^2)$	$D(r_P)$	$E(r_S)$	$E(r_S^2)$	$D(r_S)$	$E(r_Q)$	$E(r_Q^2)$	$D(r_Q)$
0	0.008	0.054	0.054	0.007	0.057	0.057	0.01	0.052	0.052
0.5	0.49	0.27	0.032	0.46	0.25	0.035	0.32	0.15	0.045
0.9	0.90	0.80	0.003	0.87	0.76	0.005	0.70	0.51	0.028
MIX	0.87	0.76	0.003	0.84	0.72	0.006	0.67	0.48	0.032

Таблица 1: Результаты вычислений при n=20

ρ	$E(r_P)$	$E(r_P^2)$	$D(r_P)$	$E(r_S)$	$E(r_S^2)$	$D(r_S)$	$E(r_Q)$	$E(r_Q^2)$	$D(r_Q)$
0	-0.003	0.016	0.016	-0.003	0.016	0.016	-0.003	0.016	0.016
0.5	0.49	0.25	0.01	0.47	0.23	0.011	0.33	0.12	0.015
0.9	0.90	0.81	0.001	0.88	0.78	0.001	0.71	0.51	0.009
MIX	0.88	0.77	0.001	0.86	0.74	0.002	0.68	0.47	0.009

Таблица 2: Результаты вычислений при n=60

ρ	$E(r_P)$	$E(r_P^2)$	$D(r_P)$	$E(r_S)$	$E(r_S^2)$	$D(r_S)$	$E(r_Q)$	$E(r_Q^2)$	$D(r_Q)$
0	0.0	0.01	0.01	-0.001	0.01	0.01	0.001	0.01	0.01
0.5	0.50	0.25	0.006	0.48	0.23	0.006	0.33	0.12	0.009
0.9	0.90	0.81	0.0	0.89	0.79	0.001	0.71	0.51	0.005
MIX	0.88	0.77	0.001	0.86	0.75	0.001	0.67	0.46	0.005

Таблица 3: Результаты вычислений при n=100

3.2 Эллипсы рассеивания

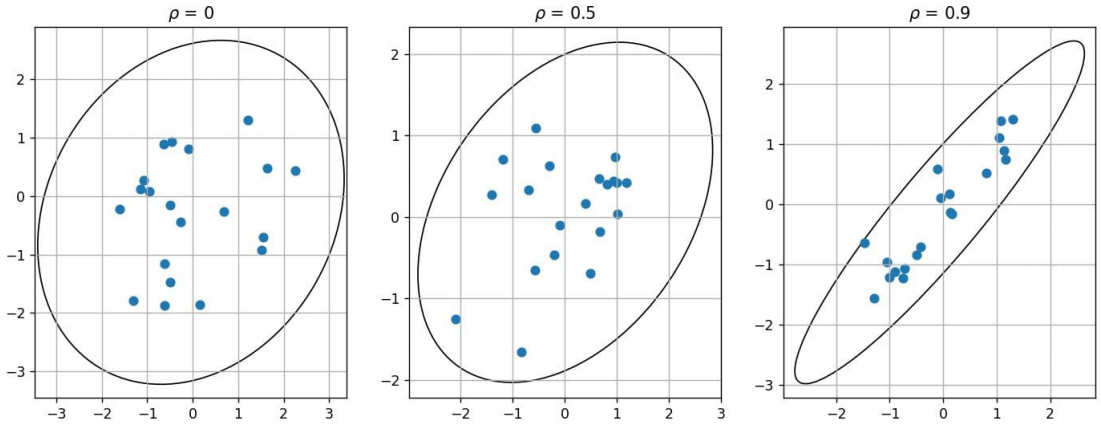

Рис. 1: Двумерное нормальное распределение, n = 20

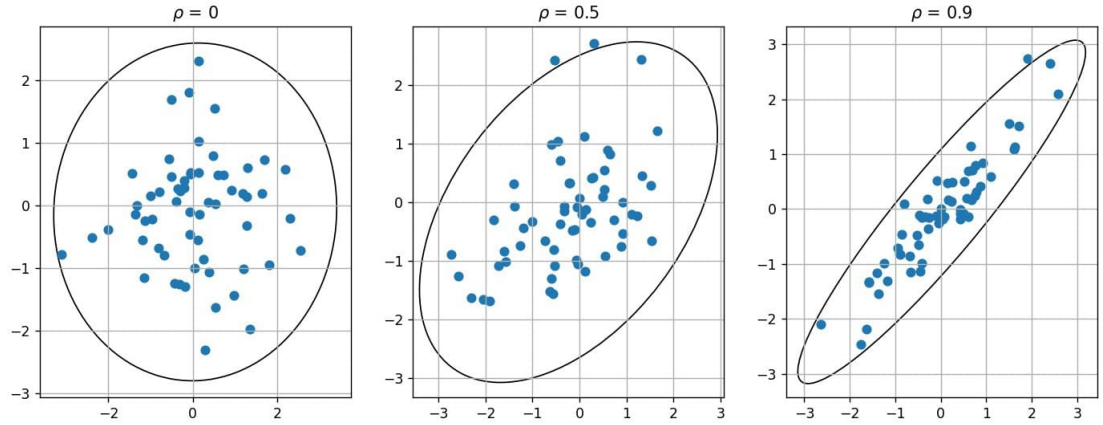

Рис. 2: Двумерное нормальное распределение, n = 60

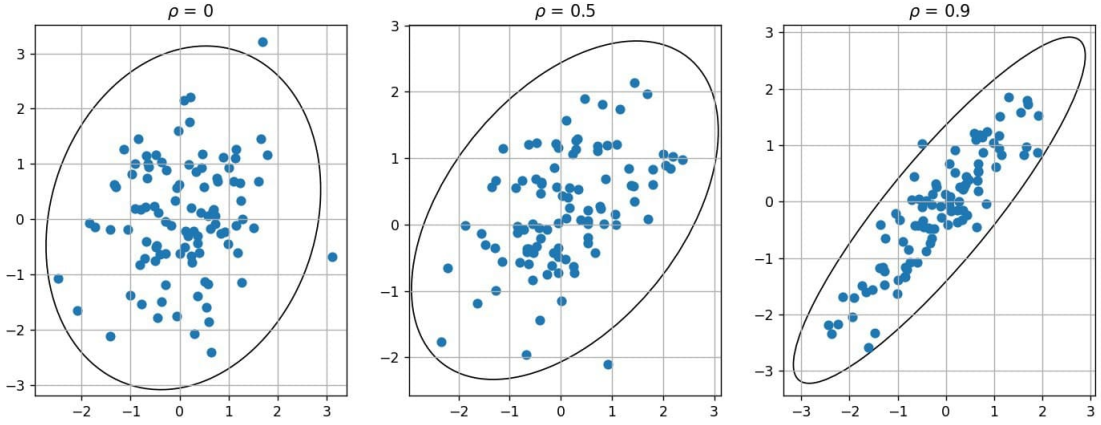

Рис. 3: Двумерное нормальное распределение, n = 100

3.3 Оценки коэффициентов линейной регрессии

3.3.1 Выборка без возмущений

1. Метод наименьших квадратов: $a = 1.84, b = 1.80$
2. Метод наименьших модулей: $a = 1.87, b = 1.96$

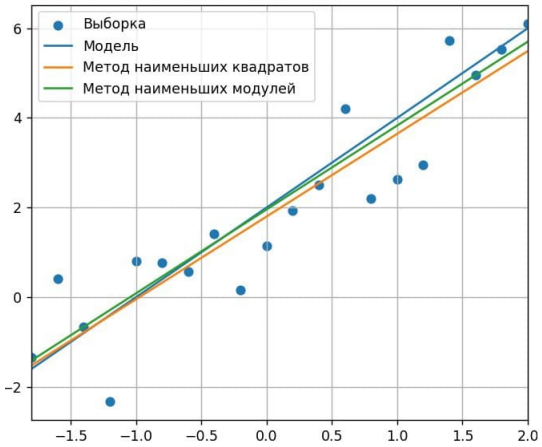

Рис. 4: Выборка без возмущений

3.3.2 Выборка с возмущениями

1. Метод наименьших квадратов: $a = 0.56, b = 1.80$
2. Метод наименьших модулей: $a = 2.01, b = 1.69$

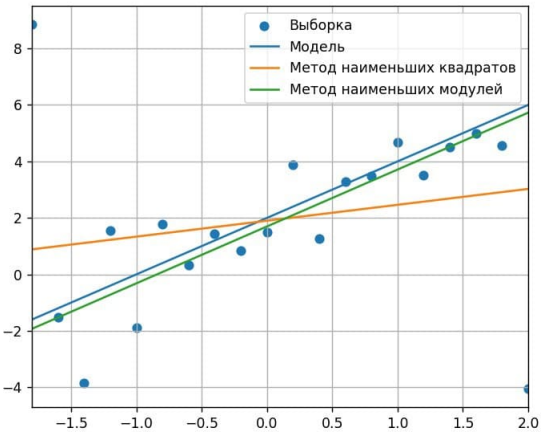

Рис. 5: Выборка с возмущениями

3.4 Проверка гипотезы о законе распределения генеральной совокупности

3.4.1 Метод максимального правдоподобия

$\bar{\mu} = -0.12 \quad \bar{\sigma} = 1.05$

3.4.2 Критерий хи-квадрат

1. *Нормальное распределение n = 100*

   $k = 7$

   $p = \{0.13, 0.20, 0.27, 0.22, 0.12, 0.04, 0.01\}$

   $n = \{11, 20, 33, 21, 11, 2, 2\}$

   $\chi^2 = 4.13$

   $\chi^2_{0.95}(6) = 12.6$

   $\chi^2 < \chi^2_{0.95}(6) =>$ гипотеза $N(x, 0, 1)$ принимается.

2. *Распределение Лапласа n = 20*

   $k = 4$

   $p = \{0.11, 0.60, 0.28, 0.01\}$

   $n = \{4, 12, 2, 1\}$

   $\chi^2 = 7.13$

   $\chi^2_{0.95}(3) = 7.8$

   $\chi^2 < \chi^2_{0.95}(6) =>$ гипотеза $L(x, 0, 1)$ принимается.

3. *Равномерное распределение n = 20*

   $k = 4$

   $p = \{0.16, 0.34, 0.34, 0.17\}$

   $n = \{6, 4, 6, 3\}$

   $\chi^2 = 3.82$

   $\chi^2_{0.95}(3) = 7.8$

   $\chi^2 < \chi^2_{0.95}(6) =>$ гипотеза $U(x, -\sqrt{3}, \sqrt{3})$ принимается.

3.5 Доверительные интервалы для параметров нормального распределения

1. *Нормальное распределение n = 20*

   - Интервальная оценка макс. правдоподобия: $-2.71 < x < 2.5$
   - Доверительный интервал мат. ожидания Стьюдента: $-0.37 < m < 0.76$
   - Доверительный интервал стд отклонения χ^2: $0.92 < \sigma < 1.76$

2. *Нормальное распределение n = 100*

   - Интервальная оценка макс. правдоподобия: $-1.99 < x < 2.07$
   - Доверительный интервал мат. ожидания Стьюдента: $-0.30 < m < 0.14$
   - Доверительный интервал стд отклонения χ^2: $0.97 < \sigma < 1.28$

4 Обсуждение

4.1 Выборочные коэффициенты корреляции

Для двумерного нормального распределения дисперсии выборочные коэффициенты корреляции Пирсона и Спирмена примерно равны для любого размера выборки и приближают выбранный коэффициент корреляции. Выборочный квадрантный корреляционный коэффициент меньше чем коэффициенты Пирсона и Спирмена, и прямо пропорционально зависит от выбранного коэффициента корреляции.

4.2 Оценки коэффициентов линейной регрессии

По полученным результатам можно сказать, что критерии наименьших квадратов и наименьших модулей примерно одинакого оценивают коэффициенты линейной регрессии на выборке без возмущений. На выборке с сильными крайними возмущенями критерий наименьших модулей показывет себя намного более устойчиво, чем критерий наименьших квадратов.

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

Гипотеза H_0 о нормальном законе распределения $N(x, \bar{\mu}, \bar{\sigma})$ на уровне значимости $\alpha = 0.05$ согласуется с выборкой для нормального распределения $N(x, 0, 1)$.
Также видно, что из-за очень маленького размера выборок, сгенерированных по равномерному закону и закону Лапласа, гипотеза H_0 оказалась принята.

4.4 Доверительные интервалы

Для любого размера выборки доверительные интервалы для параметров μ и σ накрывают их. Для большей выборки доверительные интервалы являются соответственно более точными, т.е. меньшими по длине.

5 Дополнительно

Код лабораторной работы:
https://github.com/S0krat/MakoveevLev_mathstat2023/tree/main/Lab2/src