Санкт-Петербургский политехнический университет Петра Великого
Физико Механический институт
Высшая школа прикладной математики и вычислительной физики

Отчёт по лабораторной работе №5

по дисциплине
«Математическая статистика»

Выполнила студент
группы 5030102/90101 Кузин Иван Никитович

Проверил
Доцент, к.ф.-м.н. Баженов Александр Николаевич

Санкт-Петербург
2022

СОДЕРЖАНИЕ

СПИСОК ИЛЛЮСТРАЦИЙ **3**

1 Постановка задачи **4**

2 Теория **4**

2.1 Двумерное нормальное распределение 4

2.2 Корреляционный момент (ковариация) и коэффициент корреляции 4

2.3 Выборочные коэффициенты корреляции 5

2.3.1 Выборочный коэффициент корреляции Пирсона 5

2.3.2 Выборочный квадрантный коэффициент корреляции 5

2.3.3 Выборочный коэффициент ранговой корреляции Спирмена 5

2.4 Эллипсы рассеивания 6

3 Программная реализация **7**

4 Результаты **7**

4.1 Выборочные коэффициенты корреляции 7

4.2 Эллипсы рассеивания 9

5 Обсуждение **10**

5.1 Ядерные оценки плотности распределения 10

6 Приложение **11**

СПИСОК ИЛЛЮСТРАЦИЙ

1 Двумерное нормальное распределение, $n = 20$ 9
2 Двумерное нормальное распределение, $n = 60$ 10
3 Двумерное нормальное распределение, $n = 100$ 10

1 Постановка задачи

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:

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

Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

2 Теория

2.1 Двумерное нормальное распределение

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

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

Компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $\bar{x}$,$\bar{y}$ и средними квадратическими отклонениями σ_x, σ_y соответственно [1, с. 133-134]. Параметр ρ называется коэффициентом корреляции.

2.2 Корреляционный момент (ковариация) и коэффициент корреляции

Корреляционным моментом, иначе ковариацией, двух случайных величин X и Y называется математическое ожидание произведения отклонений этих случайных величин от их математических ожиданий [1, с. 141].

$$K = cov(X, Y) = M[(X - \bar{x})(Y - \bar{y})] \tag{3}$$

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

$$\rho = \frac{K}{\sigma_x\sigma_y} \tag{4}$$

Коэффициент корреляции — это нормированная числовая характеристика, являющаяся мерой близости зависимости между случайными величинами к линейной [1, с. 150].

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, —

$$r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2 \frac{1}{n}\sum(y_i - \bar{y})^2}} = \frac{K}{s_X s_Y}, \tag{5}$$

где K, s_X^2, s_Y^2 — выборочные ковариация и дисперсии с.в. X и Y [1, с. 535].

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

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \tag{6}$$

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 $(medx,\ medy)$

2.3.3 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 :

$$r_S = \frac{\frac{1}{n}\sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum (u_i - \bar{u})^2 \frac{1}{n}\sum (v_i - \bar{v})^2}}, \tag{7}$$

где $\bar{u} = \bar{v} = \frac{1+2+...+n}{n} = \frac{n+1}{2}$ — среднее значение рангов [1, с. 540-541].

2.4 Эллипсы рассеивания

Рассмотрим поверхность распределения, изображающую функцию (1). 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:

$$\frac{(x-\bar{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} = const \tag{8}$$

Уравнение эллипса 8 можно проанализировать обычными методами аналитической геометрии. Применяя их, убеждаемся, что центр эллипса 8 находится в точке с координатами $(\bar{x}, \bar{y})$; что касается направления осей симметрии эллипса, то они составляют с осью Ox углы, определяемые уравнением

$$tg(2\alpha) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \tag{9}$$

Это уравнение дает два значения углов: α и α_1, различающиеся на $\frac{\pi}{2}$.

Таким образом, ориентация эллипса 8 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 $(\bar{x}, \bar{y})$. Во всех точках каждого из таких эллипсов плотность распределения $N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho)$ постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче эллипсами рассеивания. Общие оси всех эллипсов рассеивания называются главными осями рассеивания [2, с. 193-194].

3 Программная реализация

Лабораторная работа выполнена на языке Python вресии 3.7 в среде разработки JupyterLab. Использовались дополнительные библиотеки:

1. scipy (генерация выборок)
2. statsmodels, statistics (построение эмпирических функций распределения)
3. matplotlib (визуализация)
4. numpy (вычисление ряда числовых характеристик)

В приложении находится ссылка на GitHub репозиторий с исходныи кодом.

4 Результаты

4.1 Выборочные коэффициенты корреляции

$\rho = 0$	r	r_S	r_Q
E(z)	-0.018	-0.012	0.0
$E(z^2)$	0.026	0.027	0.04
D(z)	0.052	0.052	0.052
$\rho = 0.5$	r	r_S	r_Q
E(z)	0.5	0.473	0.4
$E(z^2)$	0.25	0.224	0.16
D(z)	0.032	0.036	0.047
$\rho = 0.9$	r	r_S	r_Q
E(z)	0.905	0.882	0.6
$E(z^2)$	0.819	0.778	0.36
D(z)	0.002	0.005	0.029

Таблица 1: Двумерное нормальное распределение, n = 20

$\rho = 0$	r	r_S	r_Q
E(z)	-0.002	0.002	0.0
$E(z^2)$	0.009	0.009	0.004
D(z)	0.018	0.018	0.018
$\rho = 0.5$	r	r_S	r_Q
E(z)	0.503	0.481	0.333
$E(z^2)$	0.253	0.231	0.111
D(z)	0.01	0.011	0.014
$\rho = 0.9$	r	r_S	r_Q
E(z)	0.902	0.887	0.733
$E(z^2)$	0.813	0.787	0.538
D(z)	0.001	0.001	0.01

Таблица 2: Двумерное нормальное распределение, n = 60

$\rho = 0$	r	r_S	r_Q
E(z)	0.003	0.003	0.0
$E(z^2)$	0.005	0.005	0.006
D(z)	0.011	0.011	0.01
$\rho = 0.5$	r	r_S	r_Q
E(z)	0.499	0.477	0.32
$E(z^2)$	0.249	0.228	0.102
D(z)	0.006	0.006	0.009
$\rho = 0.9$	r	r_S	r_Q
E(z)	0.901	0.888	0.72
$E(z^2)$	0.812	0.788	0.518
D(z)	0.0	0.001	0.005

Таблица 3: Двумерное нормальное распределение, n = 100

$size = 20$	r	r_S	r_Q
E(z)	0.795	0.882	0.6
$E(z^2)$	0.632	0.778	0.36
D(z)	0.01	0.005	0.038
$size = 60$	r	r_S	r_Q
E(z)	0.792	0.887	0.6
$E(z^2)$	0.628	0.787	0.36
D(z)	0.003	0.001	0.011
$size = 100$	r	r_S	r_Q
E(z)	0.79	0.888	0.56
$E(z^2)$	0.624	0.788	0.314
D(z)	0.002	0.001	0.007

Таблица 4: Смесь нормальных распределений

4.2 Эллипсы рассеивания

Для уравнения эллипса выбиралась константа равная $const = 2 \cdot (2 \cdot \sigma)$

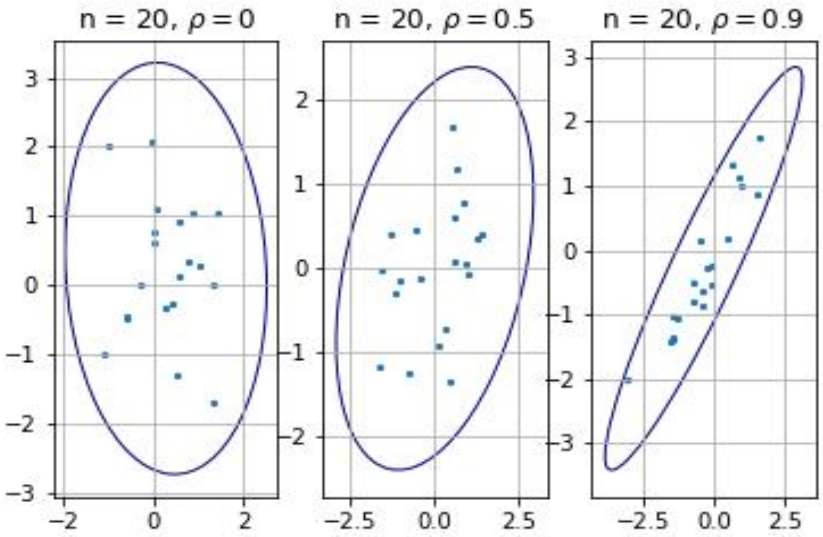

Рис. 1: Двумерное нормальное распределение, $n = 20$

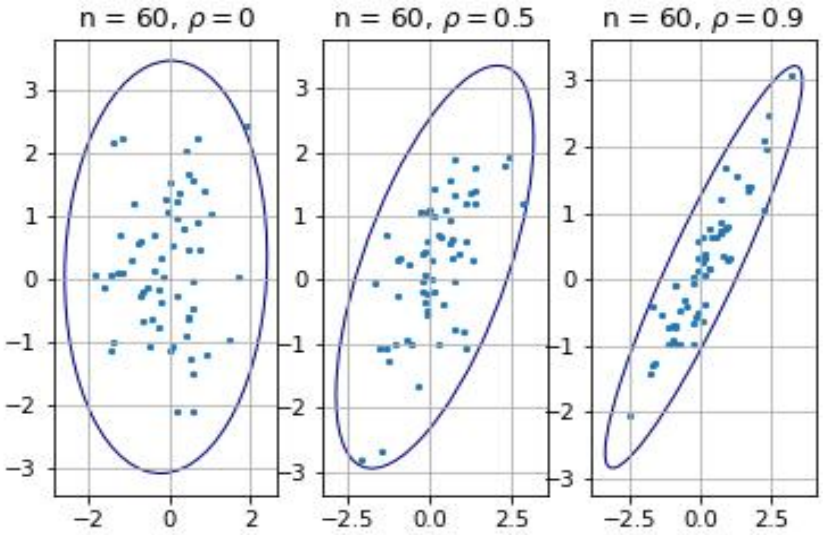

Рис. 2: Двумерное нормальное распределение, $n = 60$

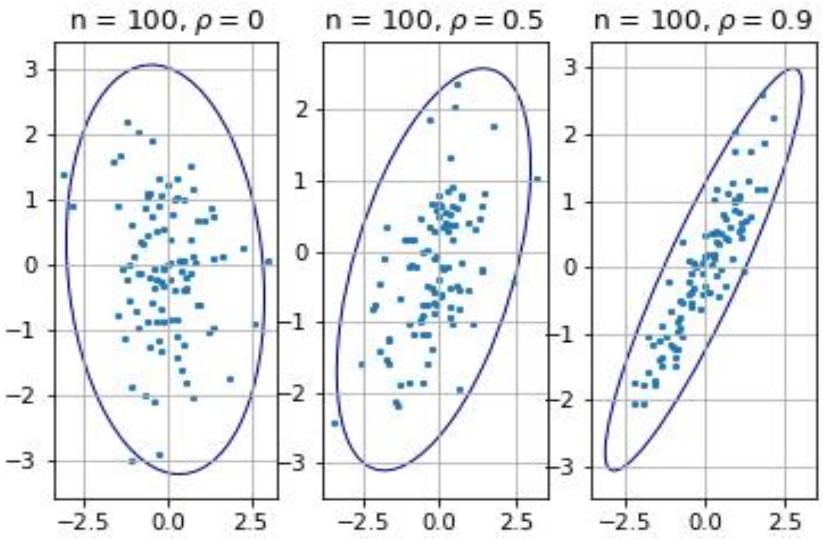

Рис. 3: Двумерное нормальное распределение, $n = 100$

5 Обсуждение

5.1 Ядерные оценки плотности распределения

Для двумерного нормального распределения дисперсии выборочных коэффициентов корреляции упорядочены следующим образом: $r < r_S < r_Q$; для смеси распределений получили обратную картину: $r_Q < r_S < r$.

Процент попавших элементов выборки в эллипс рассеивания (95%-ная доверительная область) примерно равен его теоретическому значению (95%).

6 Приложение

Код программы GitHub URL:

https://github.com/workivan/mat-ver-stat