Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики

Кафедра «Прикладная математика»

Отчет по лабораторной работе №5

по дисциплине
"Математическая статистика"

Выполнила студентка
группы 3630102/80201 Деркаченко Анна Олеговна

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

Санкт-Петербург
2021 г.

Содержание

Список иллюстраций

1. Постановка задачи

Дано нормальное двумерное распределение $N(x, y, 0, 0, 1, 1, \rho)$.

Необходимо:

1) Сгенерировать выборки размером 20, 60 и 100 элементов с коэффициентом корреляции ρ, равным 0, 0.5, 0.9
2) Сгенерировать выборки 1000 раз и вычислить среднее значение, среднее значение квадрата и дисперсию коэффициентов корреляции Пирсона, Спирмена и квадрантного коэффициента корреляции
3) Повторить все вычисления для смеси нормальных распределений:

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

4) 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

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

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. Параметр ρ называется коэффициентом корреляции.

2.2. Корреляционный момент (ковариация) и коэффициент корреляции

Корреляционный момент (ковариация) двух случайных величин X и Y - математическое ожидание произведения отклонений этих случайных величин от их математических ожиданий:

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

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

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

Коэффициент корреляции — 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:

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

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

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

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

2.4. Эллипсы рассеивания

При сечении поверхности функции нормального двумерного распределения плоскостями, паралелльными плоскости xOy, получаются эллипсы:

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

Центр эллипса находится в точке с координатами $(\bar{x}, \bar{y})$, а направления осей симметрии эллипса составляют с осью Ox углы, определяемые уравнением:

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

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

Таким образом, ориентация эллипса 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 данного распределения постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче эллипсами рассеивания. Общие оси всех эллипсов рассеивания называются главными осями рассеивания.

3. Реализация

Реализация лабораторной работы проводилась на языке Python в среде разработки PyCharm с использованием дополнительных библиотек:

- scipy
- numpy
- matplotlib
- statistics
- statsmodels

Исходный код лабораторной работы размещен в GitHub-репозитории.

URL: https://github.com/derkanw/Mathstat/tree/main/lab5

4. Результаты

4.1. Вычислительные характеристики распределения

	r	r_S	r_Q
$\rho = 0$			
$E(z)$	-0.0166	-0.0188	0.0
$E(z^2)$	0.0232	0.0245	0.04
$D(z)$	0.0514	0.053	0.0515
$\rho = 0.5$			
$E(z)$	0.5061	0.4677	0.4
$E(z^2)$	0.2562	0.2187	0.16
$D(z)$	0.0334	0.0373	0.0501
$\rho = 0.9$			
$E(z)$	0.9037	0.8797	0.8
$E(z^2)$	0.8167	0.7739	0.64
$D(z)$	0.0027	0.005	0.0289

Таблица 1. Характеристики распределения размерностью n = 20

	r	r_S	r_Q
$\rho = 0$			
$E(z)$	0.7933	0.7579	0.6
$E(z^2)$	0.6294	0.5744	0.36
$D(z)$	0.0091	0.0136	0.0386
$\rho = 0.5$			
$E(z)$	0.0016	0.0014	0.0
$E(z^2)$	0.0085	0.0082	0.0044
$D(z)$	0.0183	0.0181	0.0167
$\rho = 0.9$			
$E(z)$	0.5091	0.4894	0.3333
$E(z^2)$	0.2592	0.2395	0.1111
$D(z)$	0.0101	0.0113	0.015

Таблица 2. Характеристики распределения размерностью n = 60

	r	r_S	r_Q
$\rho = 0$			
$E(z)$	0.9029	0.8886	0.7333
$E(z^2)$	0.8153	0.7897	0.5378
$D(z)$	0.0006	0.001	0.0087
$\rho = 0.5$			
$E(z)$	0.7953	0.7736	0.6
$E(z^2)$	0.6324	0.5985	0.36
$D(z)$	0.0026	0.0037	0.0119
$\rho = 0.9$			
$E(z)$	-0.0089	-0.0078	0.0
$E(z^2)$	0.005	0.005	0.0064
$D(z)$	0.0107	0.0105	0.0096

Таблица 3. Характеристики распределения размерностью n = 100

	r	r_S	r_Q
$n = 20$			
$E(z)$	0.4987	0.479	0.32
$E(z^2)$	0.2487	0.2294	0.1024
$D(z)$	0.0061	0.0069	0.0088
$n = 60$			
$E(z)$	0.9006	0.889	0.72
$E(z^2)$	0.811	0.7904	0.5184
$D(z)$	0.0004	0.0006	0.0054
$n = 100$			
$E(z)$	0.7917	0.7734	0.56
$E(z^2)$	0.6268	0.5981	0.3136
$D(z)$	0.0015	0.0021	0.0071

Таблица 4. Смесь нормальных распределений

4.2. Эллипсы рассеивания

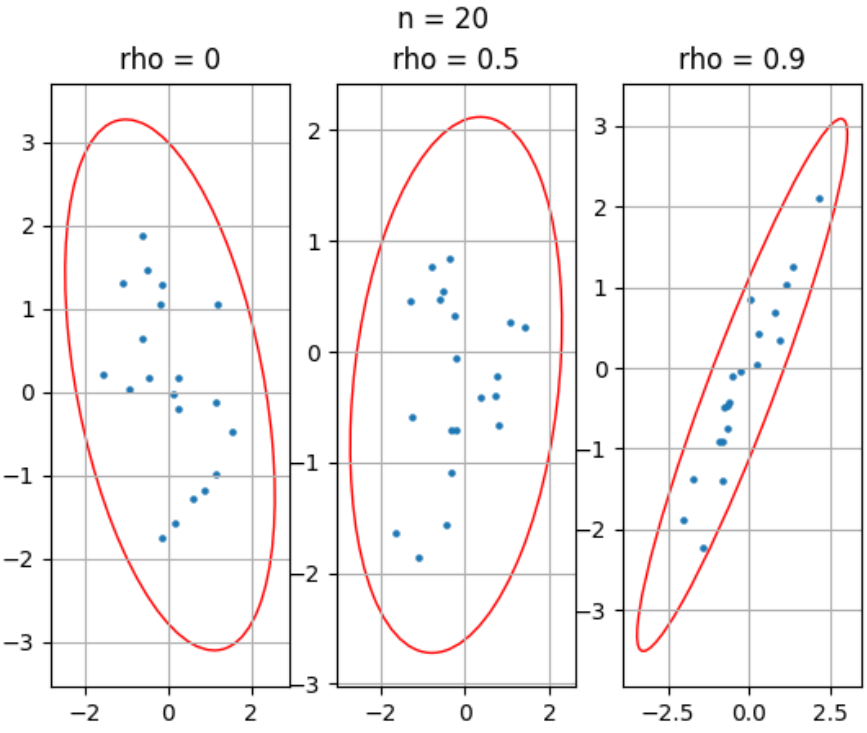

Рис. 1. Эллипс рассеивания для 20 элементов

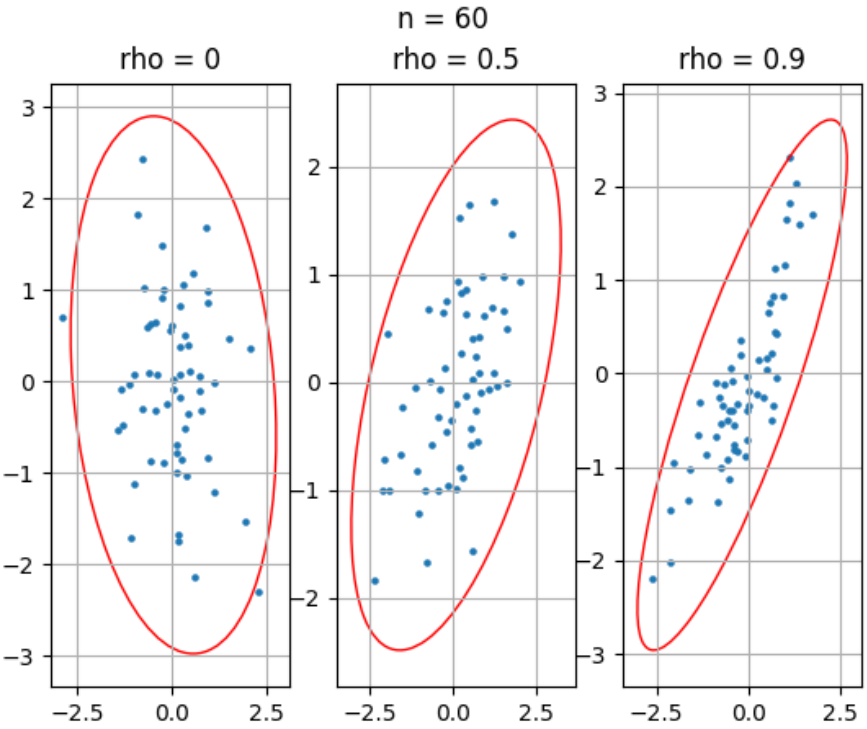

Рис. 2. Эллипс рассеивания для 60 элементов

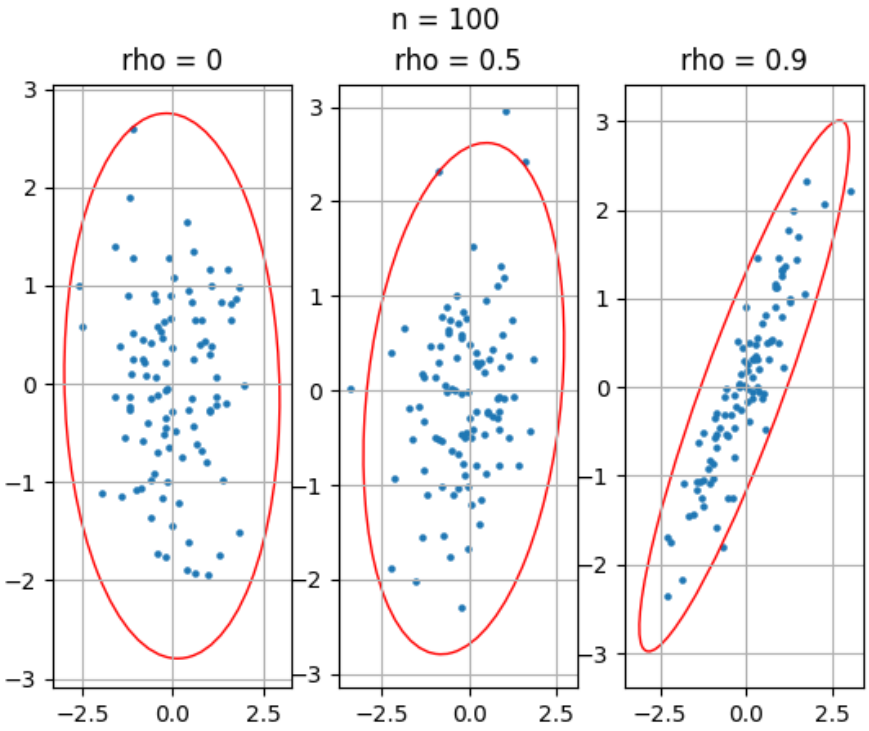

Рис. 3. Эллипс рассеивания для 100 элементов

5. Обсуждение

Исходя из таблиц характеристик распределений разных размерностей и смеси распределений, можно можно сделать вывод, что значения $E(z), E(z^2)$ для величин r, r_S, r_Q в большинстве случаев подчиняются соотношению $r > r_S > r_Q$. А для дисперсии наблюдается обратное: $r < r_S < r_Q$. При этом с ростом размерности выборки дисперсия уменьшается для более приближающегося к нулю коэффициента корреляции Стоит также сказать, что дисперсии для смеси распределений имеют меньшие значения по сравнению с соответствующими показателями для обычного распределения.

На графиках эллипсов рассеивания можно наблюдать, что почти все элементы выборки попадают в границы данных эллипсов, исходя из чего можно сделать вывод, что полученный результат можно назвать соответствующим его теоретической оценке. При этом достаточно большая часть значений концентрируется в центре эллипсов, что подтверждает вывод о теоретическом значении центра эллипса, координаты которого представляются в виде среднего для исследуемых величин X, Y.