Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

# ОТЧЕТ

**Тема:** ***Двумерное нормальное распределение***

Направление: 01.03.02 Прикладная математика и информатика

Выполнил студент гр. 33631/4 Камалетдинова Ю.

Преподаватель Баженов А.

Санкт-Петербург
2019

# Содержание

# Постановка задачи

В данной лабораторной работе рассматриваются двумерные распределения. Необходимо сгенерировать двумерные выборки $x_n = (x_1, \dots, x_n),\ y_n = (y_1, \dots, y_n)$ размерами $n = 20,\ 60,\ 100$ из нормального двумерного распределения. Коэффициенты корреляции взять равными $\rho = 0,\ 0.5,\ 0.9$. Формула для плотности распределения приведена ниже

$$N(x, y, 0, 0, 1, 1, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\} \tag{1}$$

Каждая выборка генерируется $N = 1000$ раз, и для каждоый выборки вычисляются среднее значение, среднее значение квадрата и дисперсия следующих коэффициентов

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{cov(x, y)}{\sqrt{\sigma_x^2 \sigma_y^2}} \text{ — коэффициент корреляции Пирсона,} \tag{2}$$

где $\overline{x},\ \overline{y}$ — выборочные средние $x_n,\ y_n,\ \sigma_x^2,\ \sigma_y^2$ — выборочные дисперсии

$$r_s = \rho_{rg_x, rg_y} = \frac{cov(rg_x, rg_y)}{\sqrt{\sigma_x^2 \sigma_y^2}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \text{ — коэффициент корреляции Спирмена,} \tag{3}$$

где $rg_x,\ rg_y$ — ранговые переменные, $d_i = rg(x_i) - rg(y_i)$ — разность двух рангов наблюдений. Формула для рассчета из источника [2]

$$r_q = \frac{(n_I + n_{III}) - (n_{II} + n_{IV})}{n} \text{ — квадрантный коэффициент корреляции,} \tag{4}$$

где $n_i,\ \ i = I,\ II,\ III,\ IV$ — число наблюдений, попавших в $i$ - ый квадрант на плоскости

Приведем формулы для вычисления выборочного среднего, квадрата выборочного среднего и выборочной дисперсии в двумерном случае

$$\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{i_k},\ \ k = 1,\ 2 \tag{5}$$

$$\overline{x}_k^2 = \frac{1}{n}\sum_{i=1}^{n} x_{i_k}^2,\ \ k = 1,\ 2 \tag{6}$$

$$\overline{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{i_k} - \overline{x}_k)^2,\ \ k = 1,\ 2 \tag{7}$$

Требуется повторить вычисления характеристик (5), (6), (7) корреляционных коэффициентов (2), (3), (4) для смеси нормальных распределений

$$f(x,y) = 0.9N(x,y,0,0,1,1,0.9) + 0.1N(x,y,0,0,10,10,-0.9) \tag{8}$$

Полученные выборки необходимо изобразить на плоскости и изобразить эллипс равновероятности

# Реализация

Для выполнения поставленной задачи будем пользоваться библиотеками для языка Python: *numpy, scipy* – расчеты, законы распределения вероятностей; *matplotlib, seaborn* – визуализация результатов. Ход работы:

- Задаем распределение с заданными параметрами
- Формируем двойной цикл: внешний — по объемам выборок $n$, внутренний — по корреляционным коэффициентам $\rho$

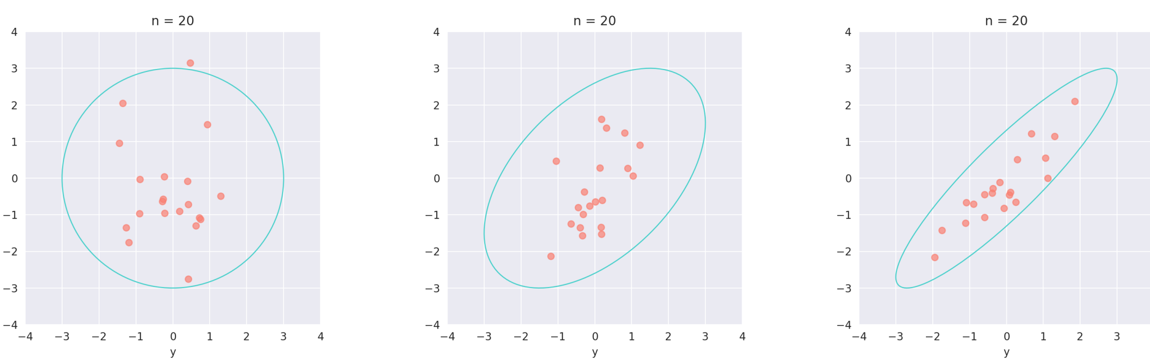
- На каждой итерации цикла для выборки строим 99% доверительный эллипс, теоретическое описание которого находится по ссылкам [1], [3]; изображаем выборки и эллипс в одних осях

- Генерируем выборку и вычисляем корреляционные коэффициенты по формулам (2), (3), (4) 1000 раз

- Находим среднее, среднее квадрата и дисперсию корреляций по формулам (5), (6), (7)

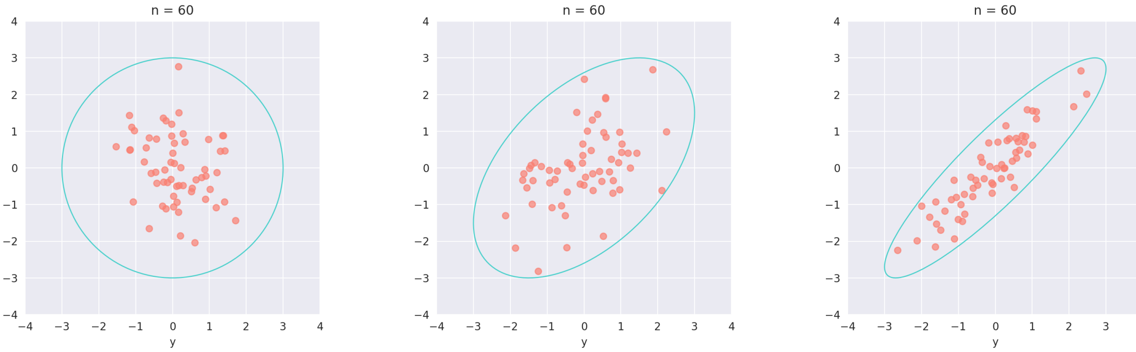
# Результат

### Выборки из двумерного нормального распределения и 99%-эллипс



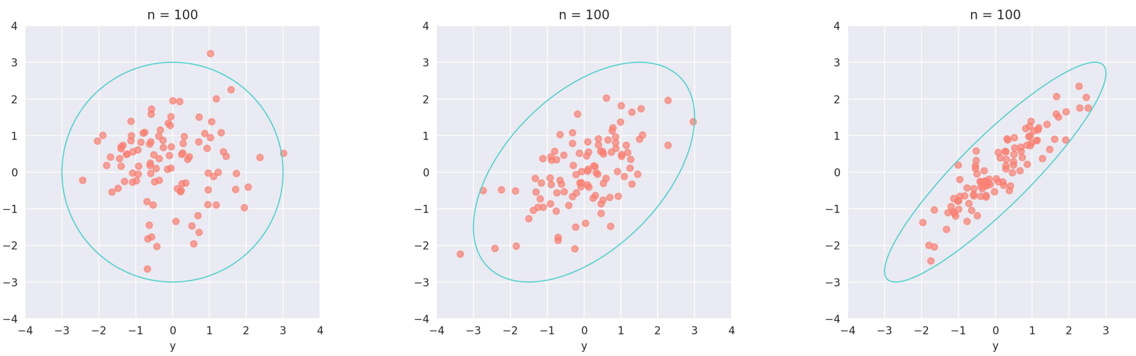

(a) (b) (c)

Рис. 1: $\rho$: (a) 0.0; (b) 0.5; (c) 0.9




(a) (b) (c)

Рис. 2: $\rho$: (a) 0.0; (b) 0.5; (c) 0.9




(a) (b) (c)

Рис. 3: $\rho$: (a) 0.0; (b) 0.5; (c) 0.9

### Характеристики для двумерного нормального распределения

| $n = 20, \rho = 0.0$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.0132 | 0.0086 | 0.0073 |
| $E(z^2)$ | 0.0505 | 0.0517 | 0.0489 |
| $D(z)$ | 0.0503 | 0.0517 | 0.0488 |

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_p^2) < E(r_s^2)$$
$$D(r_q) < D(r_p) < D(r_s)$$

| $n=20, \rho=0.5$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.4893 | 0.4628 | 0.3263 |
| $E(z^2)$ | 0.2712 | 0.2484 | 0.1530 |
| $D(z)$ | 0.0318 | 0.0343 | 0.0465 |

$E(r_q) < E(r_s) < E(r_p)$
$E(r_q^2) < E(r_s^2) < E(r_p^2)$
$D(r_p) < D(r_s) < D(r_q)$

| $n=20, \rho=0.9$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.8942 | 0.8645 | 0.7170 |
| $E(z^2)$ | 0.8019 | 0.7518 | 0.5384 |
| $D(z)$ | 0.0024 | 0.0044 | 0.0243 |

$E(r_q) < E(r_s) < E(r_p)$
$E(r_q^2) < E(r_s^2) < E(r_p^2)$
$D(r_p) < D(r_s) < D(r_q)$

| $n=60, \rho=0.0$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.0024 | 0.0012 | -0.0028 |
| $E(z^2)$ | 0.0172 | 0.0175 | 0.0167 |
| $D(z)$ | 0.0172 | 0.0175 | 0.0167 |

$E(r_s) < E(r_p) < E(r_q)$
$E(r_q^2) < E(r_p^2) < E(r_s^2)$
$D(r_q) < D(r_p) < D(r_s)$

| $n=60, \rho=0.5$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.4942 | 0.4725 | 0.3372 |
| $E(z^2)$ | 0.2544 | 0.2344 | 0.1287 |
| $D(z)$ | 0.0101 | 0.0111 | 0.0150 |

$E(r_q) < E(r_s) < E(r_p)$
$E(r_q^2) < E(r_s^2) < E(r_p^2)$
$D(r_p) < D(r_s) < D(r_q)$

| $n=60, \rho=0.9$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.8991 | 0.8832 | 0.7103 |
| $E(z^2)$ | 0.8090 | 0.7811 | 0.5122 |
| $D(z)$ | 0.0007 | 0.0010 | 0.0077 |

$E(r_q) < E(r_s) < E(r_p)$
$E(r_q^2) < E(r_s^2) < E(r_p^2)$
$D(r_p) < D(r_s) < D(r_q)$

| $n = 100, \rho = 0.0$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | -0.0081 | -0.0064 | -0.0062 |
| $E(z^2)$ | 0.0107 | 0.0108 | 0.0101 |
| $D(z)$ | 0.0106 | 0.0107 | 0.0101 |

$E(r_q) < E(r_s) < E(r_p)$

$E(r_q^2) < E(r_p^2) < E(r_s^2)$

$D(r_q) < D(r_p) < D(r_s)$

| $n = 100, \rho = 0.5$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.4980 | 0.4775 | 0.3341 |
| $E(z^2)$ | 0.2538 | 0.2344 | 0.1206 |
| $D(z)$ | 0.0058 | 0.0064 | 0.0090 |

$E(r_q) < E(r_s) < E(r_p)$

$E(r_q^2) < E(r_s^2) < E(r_p^2)$

$D(r_p) < D(r_s) < D(r_q)$

| $n = 100, \rho = 0.9$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.8994 | 0.8868 | 0.7134 |
| $E(z^2)$ | 0.8093 | 0.7869 | 0.5137 |
| $D(z)$ | 0.0004 | 0.0006 | 0.0048 |

$E(r_q) < E(r_s) < E(r_p)$

$E(r_q^2) < E(r_s^2) < E(r_p^2)$

$D(r_p) < D(r_s) < D(r_q)$

### Характеристики для смеси двумерных нормальных распределений

| $n = 20$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.6886 | 0.6547 | 0.4866 |
| $E(z^2)$ | 0.4904 | 0.4485 | 0.2738 |
| $D(z)$ | 0.0162 | 0.0200 | 0.0370 |

$E(r_q) < E(r_s) < E(r_p)$

$E(r_q^2) < E(r_s^2) < E(r_p^2)$

$D(r_p) < D(r_s) < D(r_q)$

| $n = 60$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.6948 | 0.6702 | 0.4885 |
| $E(z^2)$ | 0.4874 | 0.4555 | 0.2506 |
| $D(z)$ | 0.0047 | 0.0063 | 0.0119 |

$E(r_q) < E(r_s) < E(r_p)$

$E(r_q^2) < E(r_s^2) < E(r_p^2)$

$D(r_p) < D(r_s) < D(r_q)$

| $n = 100$ | $r_p$ | $r_s$ | $r_q$ |
|---|---|---|---|
| $E(z)$ | 0.7003 | 0.6796 | 0.4952 |
| $E(z^2)$ | 0.4932 | 0.4653 | 0.2525 |
| $D(z)$ | 0.0028 | 0.0035 | 0.0073 |

$$E(r_q) < E(r_s) < E(r_p)$$
$$E(r_q^2) < E(r_s^2) < E(r_p^2)$$
$$D(r_p) < D(r_s) < D(r_q)$$

Рассмотрим полученные соотношения. На некоррелированных данных квадрантный коэффициент корреляции имеет наименьшую дисперсию. Также дисперсия коэффициента Пирсона всегда меньше дисперсии коэффициента Спирмена вне зависимости от объема выборки или корреляции двумерного нормального распределения (1). Можно полагать, что в случае такого распределения лучше рассчитывать коэффициент корреляции Пирсона.

Неравенства для смеси двух нормальных распределений (8) совпадают с неравенствами для двумерного нормального распределения с корреляциями $\rho = 0.5,\ 0.9$.

# Список литературы


[1] *Eisele, R.* (2018). How to plot a covariance error ellipse. URL: https://www.xarg.org/2018/04/how-to-plot-a-covariance-error-ellipse/

[2] *Zwillinger, D. and Kokoska, S.* (2000). CRC Standard Probability and Statistics Tables and Formulae. Chapman & Hall: New York. 2000.

[3] *Ллойд Э., Ледерман У.* Справочник по прикладной статистике. Том 1. М.: Финансы и статистика, 1989. - 510 с.