Санкт-Петербургский Политехнический Университет Петра Великого
Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Лабораторная работа №3

по дисциплине “Математическая статистика”

Обучающаяся: А.Д. Балакшина (группа 5030102/20101)

Преподаватель: А.Н. Баженов

Санкт-Петербург

2025

Содержание

1 Формулировка задачи

1. Сгенерировать двумерные выборки размерами 20, 60, 100 для нормального двумерного распределения $N(x, y, 0, 0, 1, 1, \rho)$.

2. Коэффициент корреляции ρ взять равным 0, 0.5, 0.9.

3. Каждая выборка генерируется 1000 раз, и для неё вычисляются среднее значение, среднее значение квадрата и дисперсия коэффициентов корреляции Пирсона, Спирмена и квадратного коэффициента корреляции.

4. Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

Повторить все вычисления для смеси нормальных распределений:

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

2 Формализация

1. **Коэффициент корреляции Пирсона**:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

2. **Коэффициент корреляции Спирмена**:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

   где d_i — разность рангов x_i и y_i.

3. **Квадратный коэффициент корреляции**:

$$r^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \right)^2.$$

3 Выполнение работы

Лабораторная работа выполнена на языке программирования Python 3.12 с использованием библиотек numpy, scipy, mathplotlib.
Были сгенерированны выборки, выполнены вычисления (выводились в консоль в формате таблиц LATEX), построены эллипсы равновероятности (сохранялись в виде картинок png). Программа отработала корректно.

4 Результаты

4.1 Нормальное распределение

ρ	Method	Mean	Mean of squares	Variance
0.0	$Pearson^{(1)}$	0.0078	0.0552	0.0552
0.0	$Spearman^{(2)}$	0.0053	0.0535	0.0535
0.0	$Quadratic^{(3)}$	0.0031	0.0538	0.0538
0.5	$Pearson^{(1)}$	0.4786	0.2615	0.0325
0.5	$Spearman^{(2)}$	0.4481	0.2368	0.0361
0.5	$Quadratic^{(3)}$	0.2188	0.1292	0.0813
0.9	$Pearson^{(1)}$	0.8949	0.8035	0.0026
0.9	$Spearman^{(2)}$	0.8653	0.7536	0.0048
0.9	$Quadratic^{(3)}$	0.7868	0.6396	0.0205

Таблица 1: Statistics for correlation coefficients (n = 20)

ρ	Method	Mean	Mean of squares	Variance
0.0	$Pearson^{(1)}$	0.0040	0.0167	0.0167
0.0	$Spearman^{(2)}$	0.0029	0.0164	0.0164
0.0	$Quadratic^{(3)}$	-0.0033	0.0153	0.0153
0.5	$Pearson^{(1)}$	0.4961	0.2566	0.0105
0.5	$Spearman^{(2)}$	0.4745	0.2368	0.0117
0.5	$Quadratic^{(3)}$	0.2426	0.0887	0.0299
0.9	$Pearson^{(1)}$	0.8989	0.8087	0.0006
0.9	$Spearman^{(2)}$	0.8831	0.7810	0.0011
0.9	$Quadratic^{(3)}$	0.8000	0.6464	0.0064

Таблица 2: Statistics for correlation coefficients (n = 60)

ρ	Method	Mean	Mean of squares	Variance
0.0	$Pearson^{(1)}$	-0.0059	0.0097	0.0097
0.0	$Spearman^{(2)}$	-0.0057	0.0095	0.0094
0.0	$Quadratic^{(3)}$	0.0012	0.0105	0.0105
0.5	$Pearson^{(1)}$	0.5027	0.2588	0.0061
0.5	$Spearman^{(2)}$	0.4824	0.2393	0.0066
0.5	$Quadratic^{(3)}$	0.2505	0.0810	0.0182
0.9	$Pearson^{(1)}$	0.9001	0.8106	0.0004
0.9	$Spearman^{(2)}$	0.8873	0.7880	0.0006
0.9	$Quadratic^{(3)}$	0.8043	0.6507	0.0038

Таблица 3: Statistics for correlation coefficients (n = 100)

4.2 Смесь распределений

Таблица 4: Statistics for sample size n = 20

Statistic	$Pearson^{(1)}$	$Spearman^{(2)}$	$Quadratic^{(3)}$
Mean	0.4141	0.6596	0.5276
Mean squared	0.3198	0.4624	0.3842
Variance	0.1483	0.0273	0.1058

Таблица 5: Statistics for sample size n = 60

Statistic	$Pearson^{(1)}$	$Spearman^{(2)}$	$Quadratic^{(3)}$
Mean	0.3959	0.6860	0.4167
Mean squared	0.2108	0.4789	0.2351
Variance	0.0541	0.0083	0.0615

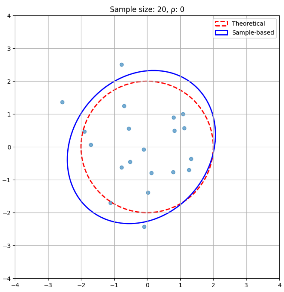
Таблица 6: Statistics for sample size n = 100

Statistic	$Pearson^{(1)}$	$Spearman^{(2)}$	$Quadratic^{(3)}$
Mean	0.3942	0.6914	0.3886
Mean squared	0.1910	0.4832	0.1955
Variance	0.0356	0.0053	0.0445

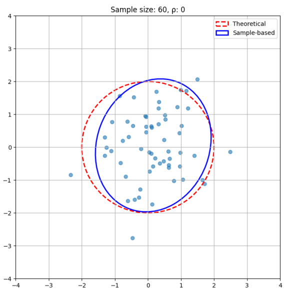
В скобках указан номер формулы, по которой проводился расчёт

5 Эллипсы равновероятностей

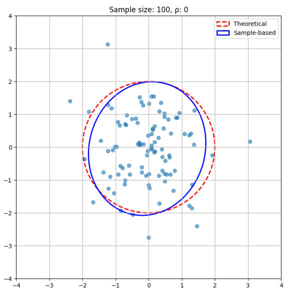
5.1 Нормальное распределение



(a) Эллипс для n=10

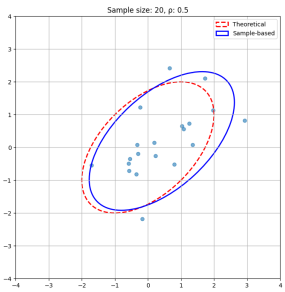


(b) Эллипс длян=60

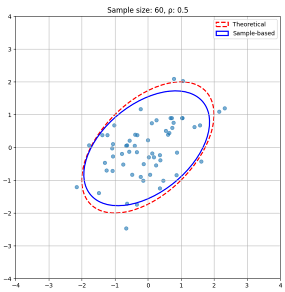


(c) Эллипс для n=1000

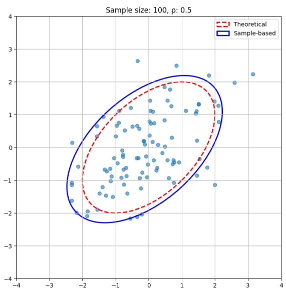
Рис. 1: Эллипсы равновероятности для разных размеров выборки при $\rho = 0$



(a) Эллипс для n=10

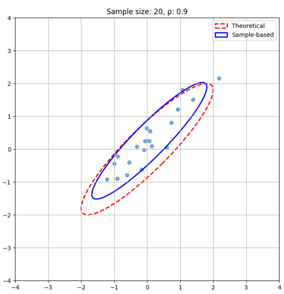


(b) Эллипс для n=60

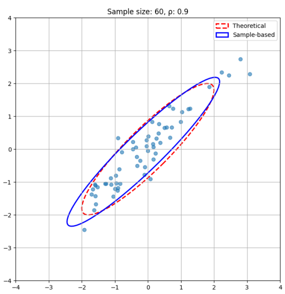


(c) Эллипс для n=1000

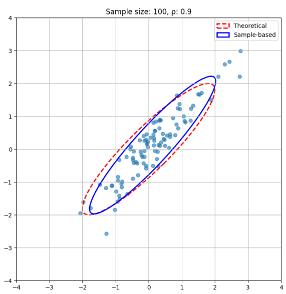
Рис. 2: Эллипсы равновероятности для разных размеров выборки при $\rho = 0.5$



(a) Эллипс для n=10



(b) Эллипс для n=60



(c) Эллипс для n=1000

Рис. 3: Эллипсы равновероятности для разных размеров выборки при $\rho = 0.9$

6 Выводы

Оценка коэффициентов корреляции

В ходе работы были исследованы три коэффициента корреляции: Пирсона, Спирмена и квадратный коэффициент корреляции. Для нормального распределения с разными значениями ρ (0, 0.5, 0.9) и размерами выборок (20, 60, 100) были получены следующие результаты:

- В большой части результатов среднее значение коэффициентов близко к теоретическому значению ρ, что подтверждает их несмещённость при больших размерах выборок.
- Дисперсия коэффициентов уменьшается с увеличением размера выборки, что согласуется с законом больших чисел.

Анализ смеси нормальных распределений

Для смеси $0.9N(0, 0, 1, 1, 0.9) + 0.1N(0, 0, 10, 10, -0.9)$:

- Коэффициенты корреляции показали смещение оценок из-за наличия выбросов и второй компоненты с отрицательной корреляцией.
- Дисперсия оценок увеличилась, что объясняется высокой дисперсией второй компоненты смеси.
- Коэффициент Спирмена оказался более устойчивым к аномалиям по сравнению с коэффициентом Пирсона.

Визуализация данных

Эллипсы равновероятности для нормального распределения отражают степень корреляции: при $\rho = 0.9$ эллипс вытянут, а при $\rho = 0$ принимает форму окружности. То есть на графиках точек для $\rho = 0.9$ наблюдается явная линейная зависимость, а для $\rho = 0$ точки распределены хаотично.

Заключение

Лабораторная работа подтвердила важность выбора коэффициента корреляции в зависимости от характера данных:

- Для нормальных данных без аномалий подходит коэффициент Пирсона.
- Для данных с выбросами или смесями распределений рекомендуется использовать коэффициент Спирмена.
- Квадратный коэффициент корреляции может быть дополнительным инструментом для анализа сложных зависимостей.