Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Кафедра «Телематика (при ЦНИИ РТК)»

Отчет по лабораторной работе

Расчет коэффициентов корреляции

По дисциплине «Теория вероятностей и математическая статистика»

Выполнил
Студент гр.3630201/80101 _____ В.Н. Сеннов

Руководитель
доцент к.ф.-м.н. _____ А.Н. Баженов

«___»_____ 202__г.

Санкт-Петербург
2020

Содержание

Список таблиц

Список иллюстраций

1 Постановка задачи

Необходимо сгенерировать выборки размерами 20, 60, 100 элементов для двумерного нормального распределения $N(x, y, 0, 0, 1, 1, \rho)$, где $\rho = 0.0,\ 0.5,\ 0.9$, а также для смеси нормальных распределений:

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

Для каждого типа выборки необходимо вычислить коэффициенты корреляции Пирсона, Спирмена и квадрантный коэффициент корреляции 1000 раз, посчитать среднее значение и дисперсию. Также для каждого типа выборки нужно построить эллипс рассеивания.

2 Математическое описание

2.1 Двумерное нормальное распределение

Двумерное нормальное распределение в данной лабораторной имеет следующую плотность вероятности:

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

где ρ — коэффициент корреляции.

2.2 Коэффициенты корреляции

Коэффициентом корреляции двух величин называется величина ρ:

$$\rho = \frac{K}{\sigma_x \sigma_y},$$

где $K = M[(X - \bar{x})(Y - \bar{y})]$.

Выборочный коэффициент корреляции Пирсона r:

$$r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2 \frac{1}{n}\sum(y_i - \bar{y})^2}} \tag{3}$$

Выборочный квадрантный коэффициент корреляции r_Q:

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \tag{4}$$

где n — количество элементов выборки, а n_1, n_2, n_3, n_4 — количества точек, попавших соответственно в I, II, III, и IV квадранты координатной плоскости, параллельной OXY и имеющей центр в точке $(\operatorname{med} x; \operatorname{med} y)$.

Выборочный коэффициент ранговой корреляции Спирмена r_S:

$$r_S = \frac{\frac{1}{n}\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum(u_i - \bar{u})^2 \frac{1}{n}\sum(v_i - \bar{v})^2}}, \tag{5}$$

где u_i — ранг величины x_i, v_i — ранг величины y_i, n — размер выборки, $\bar{u} = \bar{v} = \frac{n+1}{2}$ — средний ранг.

2.3 Эллипсы рассеивания

Эллипсом рассеивания называется эллипс, задаваемый уравнением:

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - 2\rho\frac{(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = \lambda^2, \tag{6}$$

где λ — количество стандартных отклонений.

Оси эллипса пересекают ось OX под углом α:

$$\operatorname{tg} 2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$

3 Особенности реализации

Программа для выполнения лабораторной была написана на языке Python 3.8.2. Для генерации двумерных выборок был использован модуль **stats** библиотеки scipy. Для построения графиков использовалась библиотека Matplotlib.

Отображение эллипса рассеяния, определяемого формулой (6), было реализовано с использованием примера к библиотеке Matplotlib [1]. Эллипсы были построены с коэффициентом $\lambda = 3$. Выборка для смеси нормальных распределений (1) была сгенерирована так: была сгенерирована выборка для первого распределения размером 90% от необходимого размера, а потом была сгенерирована выборка для второго распределения размером 10% от необходимого размера. Итоговая выборка является их объединением.

В приложении А приведена ссылка на репозиторий с исходным кодом.

4 Результаты работы программы

В этом разделе представлены результаты работы программы. На рис. 1-3 представлены эллипсы рассеяния, построенные по формуле 6.

На рис. 1 изображены эллипсы рассеяния для выборок из 20 элементов.

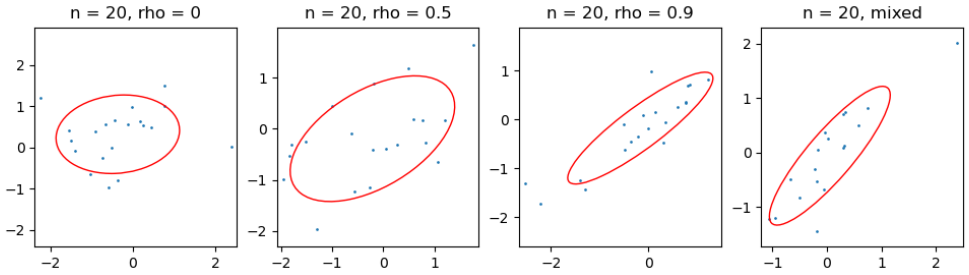

Рис. 1: Эллипсы рассеяния для выборок из 20 элементов

На рис. 2 изображены эллипсы рассеяния для выборок из 60 элементов.

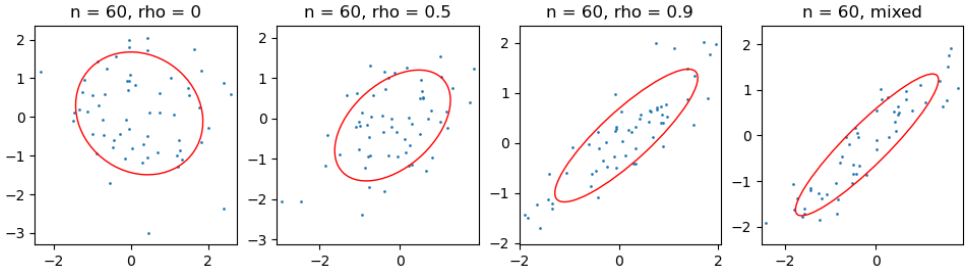

Рис. 2: Эллипсы рассеяния для выборок из 60 элементов

На рис. 3 изображены эллипсы рассеяния для выборок из 100 элементов.

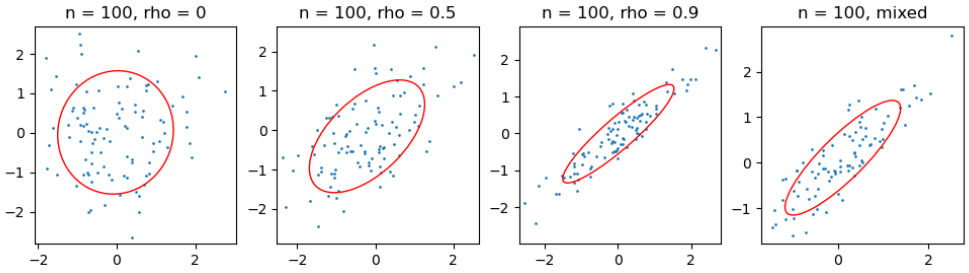

Рис. 3: Эллипсы рассеяния для выборок из 100 элементов

В таблицах 1-3 представлены коэффициенты корреляции r, r_Q, r_S, рассчитанные по формулам (3), (4) и (5) соответственно.

Средние значения $E(z)$ коэффициентов корреляции округлены до первого знака $\sqrt{D(z)}$, дисперсия $D(z)$ округлена до первого значащей цифры.

$\rho = 0$	r	r_S	r_Q
$E(z)$	0.0	0.0	0.0
$D(z)$	0.05	0.05	0.05
$\rho = 0.5$	r	r_S	r_Q
$E(z)$	0.5	0.5	0.3
$D(z)$	0.03	0.03	0.04
$\rho = 0.9$	r	r_S	r_Q
$E(z)$	0.89	0.87	0.7
$D(z)$	0.003	0.004	0.03
Смесь (1)	r	r_S	r_Q
$E(z)$	0.89	0.86	0.7
$D(z)$	0.003	0.005	0.03

Таблица 1: Двумерное нормальное распределение, $n = 20$

$\rho = 0$	r	r_S	r_Q
$E(z)$	0.0	0.0	0.0
$D(z)$	0.02	0.02	0.02
$\rho = 0.5$	r	r_S	r_Q
$E(z)$	0.5	0.5	0.3
$D(z)$	0.01	0.01	0.02
$\rho = 0.9$	r	r_S	r_Q
$E(z)$	0.90	0.88	0.7
$D(z)$	0.0007	0.001	0.009
Смесь (1)	r	r_S	r_Q
$E(z)$	0.90	0.88	0.7
$D(z)$	0.0008	0.001	0.01

Таблица 2: Двумерное нормальное распределение, $n = 60$

$\rho = 0$	r	r_S	r_Q
$E(z)$	0.0	0.0	0.0
$D(z)$	0.01	0.01	0.01
$\rho = 0.5$	r	r_S	r_Q
$E(z)$	0.50	0.48	0.3
$D(z)$	0.006	0.006	0.008
$\rho = 0.9$	r	r_S	r_Q
$E(z)$	0.90	0.89	0.71
$D(z)$	0.0004	0.0006	0.005
Смесь (1)	r	r_S	r_Q
$E(z)$	0.90	0.89	0.71
$D(z)$	0.0004	0.0006	0.005

Таблица 3: Двумерное нормальное распределение, $n = 100$

Заключение

В рамках лабораторной работы были сгенерированы выборки разного размера для двумерных нормальных распределений с различными коэффициентами корреляции и для смеси двумерных нормальных распределений. Для полученных выборок были рассчитаны выборочные коэффициенты корреляции Пирсона и Спирмена, квадрантный коэффициент корреляции. Также для выборок были построены эллипсы рассеивания.

По построенным эллипсам видно, что чем больше коэффициент корреляции, тем более выражена разница осей. Для выборки размером 100 элементов с нулевой корреляцией эллипс очень близок к окружности.

Для таких выборок видно, что коэффициент корреляции Пирсона очень близок к коэффициенту корреляции выборки. Коэффициент корреляции Спирмена менее точен, но тоже достаточно близок к коэффициенту корреляции выборки. Квадратный коэффициент корреляции для выборок с ненулевой корреляцией заметно отличается от коэффициента корреляции выборки.

Смесь двух двумерных нормальных распределений нельзя отличить от чистого двумерного нормального распределения по выборочным коэффициентам корреляции, но при большом объеме выборки эллипсы рассеивания заметно отличаются.

Программа для лабораторной была написана языке Python 3.8.2, для генерации выборок была использована библиотека scipy, для построения графиков использовалась библиотека Matplotlib.

Список литературы

[1] Plot a confidence ellipse of a two-dimensional dataset. // Matplotlib documentation. — URL: https://matplotlib.org/3.1.1/gallery/statistics/confidence_ellipse.html#sphx-glr-gallery-statistics-confidence-ellipse-py . — (дата обращения: 24.11.2020)

A Репозиторий с исходным кодом

Исходный код программы для данной лабораторной размещен на сервисе GitHub.

Ссылка на репозиторий: https://github.com/Vovan-S/TV-Lab1.