Санкт-Петербургский политехнический университет

Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной

Физики

**Отчёт**

**по лабораторным работам №5-6**

**по дисциплине**

**«Математическая статистика»**

Выполнил студент:

Корьев М.А

группа: 5030102/10101

Проверил:

доцент

Баженов Александр Николаевич

Санкт-Петербург

2024 г

Оглавление

[**5 Коэффициент корреляции** 3](#_Toc167209498)

[**5.1 Двумерное нормальное распределение** 3](#_Toc167209499)

[**5.2 Корреляционный момент (ковариация) и коэффициент корреляции** 3](#_Toc167209500)

[**5.3 Выборочные коэффициенты корреляции** 4](#_Toc167209501)

[**5.3.1 Выборочный коэффициент корреляции Пирсона** 4](#_Toc167209502)

[**5.3.2 Выборочный квадрантный коэффициент корреляции** 4](#_Toc167209503)

[**5.3.3 Выборочный коэффициент ранговой корреляции Спирмена** 4](#_Toc167209504)

[**5.4 Эллипсы рассеивания** 5](#_Toc167209505)

[**5.5 Задание** 5](#_Toc167209506)

[**5.6. Результаты** 5](#_Toc167209507)

[**5.7 Выводы** 10](#_Toc167209508)

[**6 . Простая линейная регрессия** 11](#_Toc167209509)

[**6.0.1 Модель простой линейной регрессии** 11](#_Toc167209510)

[**6.0.2 Метод наименьших квадратов** 11](#_Toc167209511)

[**6.0.3 Расчётные формулы для МНК-оценок** 12](#_Toc167209512)

[**6.1 Робастные оценки коэффициентов линейной регрессии** 14](#_Toc167209513)

[**6.2 Задание** 16](#_Toc167209514)

[**6.3 Результаты** 16](#_Toc167209515)

[**6.4 Выводы** 18](#_Toc167209516)

[**7. Реализация** 19](#_Toc167209517)

# **5 Коэффициент корреляции**

## **5.1 Двумерное нормальное распределение**

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

Компоненты двумерной нормальной случайной величины также рас-  
пределены нормально с математическими ожиданиями 𝑥, 𝑦 и средними квадратическими отклонениями 𝜎𝑥, 𝜎𝑦 соответственно [1, с. 133-134].  
Параметр 𝜌 называется коэффициентом корреляции.

## **5.2 Корреляционный момент (ковариация) и коэффициент корреляции**

Ковариация между определяется как:

Коэффициент корреляции двух случайных величин и определяется как:

где — ковариация между и , а и — средние квадратические отклонения и соответственно.

## **5.3 Выборочные коэффициенты корреляции**

### **5.3.1 Выборочный коэффициент корреляции Пирсона**

Выборочный коэффициент корреляции Пирсона определяется как:

где , , — выборочные ковариация и дисперсии случайных величин и .

### **5.3.2 Выборочный квадрантный коэффициент корреляции**

Выборочный квадрантный коэффициент корреляции определяется как:

где , , и — количества точек с координатами , попавших соответственно в I, II, III и IV квадранты декартовой системы с осями , и с центром.

### **5.3.3 Выборочный коэффициент ранговой корреляции Спирмена**

Обозначим ранги, соответствующие значениям переменной , через , а ранги, соответствующие значениям переменной , через . Выборочный коэффициент ранговой корреляции Спирмена определяется как:

где — среднее значение рангов.

## **5.4 Эллипсы рассеивания**

Уравнение проекции эллипса рассеивания на плоскость :

Центр эллипса находится в точке с координатами ; оси симметрии эллипса составляют с осью углы, определяемые уравнением:

## **5.5 Задание**

Сгенерировать двумерные выборки размерами 20, 60, 100 для нормального двумерного распределения . Коэффициент корреляции взять равным 0, 0.5, 0.9. Каждая выборка генерируется 1000 раз и для неё вычисляются:

* среднее значение,
* среднее значение квадрата,
* дисперсия коэффициентов корреляции Пирсона, Спирмена и квадрантного коэффициента корреляции.

Повторить все вычисления для смеси нормальных распределений:

Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

## **5.6. Результаты**

Table 1 Двумерное нормальное распределение, n = 20

|  |  |  |  |
| --- | --- | --- | --- |
| ρ = 0 | r |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| ρ = 0.5 | r |  |  |
|  | 0.488 | 0.458 | 0.329 |
|  | 0.269 | 0.244 | 0.156 |
|  | 0.031 | 0.034 | 0.048 |
|  |  |  |  |
| ρ = 0.9 | r |  |  |
|  | 0.894 | 0.863 | 0.681 |
|  | 0.802 | 0.749 | 0.494 |
|  | 0.002 | 0.005 | 0.03 |



Table 2 Смесь нормальных распределений, n=20

|  |  |  |  |
| --- | --- | --- | --- |
| ρ = 0 | r |  |  |
|  | 0.787 | 0.763 | 0.573 |
|  | 0.623 | 0.587 | 0.345 |
|  | 0.004 | 0.005 | 0.016 |

Изображение выглядит как диаграмма, снимок экрана, линия, График

Автоматически созданное описание

Table 3 Двумерное нормальное распределение, n = 60

|  |  |  |  |
| --- | --- | --- | --- |
| ρ = 0 | r |  |  |
|  | 0.001 | 0.001 | -0.002 |
|  | 0.018 | 0.018 | 0.017 |
|  | 0.018 | 0.018 | 0.017 |
|  |  |  |  |
| ρ = 0.5 | r |  |  |
|  | 0.497 | 0.476 | 0.331 |
|  | 0.256 | 0.237 | 0.124 |
|  | 0.01 | 0.011 | 0.015 |
|  |  |  |  |
| ρ = 0.9 | r |  |  |
|  | 0.898 | 0.882 | 0.704 |
|  | 0.806 | 0.779 | 0.503 |
|  | 0.001 | 0.001 | 0.008 |

Изображение выглядит как диаграмма, линия, График

Автоматически созданное описание

Table 4 Смесь нормальных распределений, n=60

|  |  |  |  |
| --- | --- | --- | --- |
| ρ = 0 | r |  |  |
|  | 0.792 | 0.775 | 0.581 |
|  | 0.628 | 0.602 | 0.344 |
|  | 0.001 | 0.002 | 0.006 |

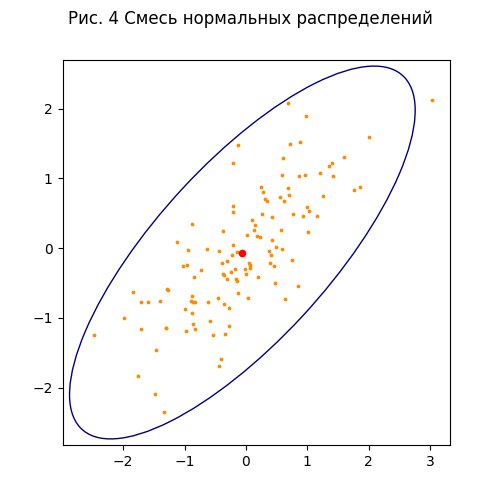


Table 5 Двумерное нормальное распределение, n = 100

|  |  |  |  |
| --- | --- | --- | --- |
| ρ = 0 | r |  |  |
|  | -0.001 | -0.001 | 0.0 |
|  | 0.01 | 0.01 | 0.01 |
|  | 0.01 | 0.01 | 0.01 |
|  |  |  |  |
| ρ = 0.5 | r |  |  |
|  | 0.327 | 0.475 | 0.327 |
|  | 0.252 | 0.233 | 0.116 |
|  | 0.006 | 0.007 | 0.009 |
|  |  |  |  |
| ρ = 0.9 | r |  |  |
|  | 0.898 | 0.884 | 0.703 |
|  | 0.807 | 0.782 | 0.5 |
|  | 0.0 | 0.001 | 0.005 |



Table 6 Смесь нормальных распределений, n=100

|  |  |  |  |
| --- | --- | --- | --- |
| ρ = 0 | r |  |  |
|  | 0.791 | 0.774 | 0.581 |
|  | 0.626 | 0.601 | 0.341 |
|  | 0.001 | 0.001 | 0.003 |

Изображение выглядит как снимок экрана, текст, диаграмма

Автоматически созданное описание

## **5.7 Выводы**

Из представленных таблиц можно сделать несколько выводов:

1. **Зависимость от размера выборки:** В целом, с увеличением размера выборки (от 20 до 100) оценки выборочных коэффициентов корреляции становятся более точными и средние значения приближаются к теоретическим значениям коэффициента корреляции. Это подтверждает закон больших чисел и позволяет делать более точные выводы о структуре данных.
2. **Влияние коэффициента корреляции:** При увеличении коэффициента корреляции ρ от 0 до 0.9 выборочные коэффициенты корреляции также увеличиваются. Это свидетельствует о том, что с увеличением линейной зависимости между переменными выборочные коэффициенты корреляции становятся более близкими к теоретическим значениям.
3. **Эффект смеси нормальных распределений:** В смеси нормальных распределений наблюдается увеличение дисперсии оценок выборочных коэффициентов корреляции по сравнению с чистым нормальным распределением. Это связано с тем, что смесь добавляет дополнительную неопределенность в данные, что отражается на точности оценок.
4. **Сравнение методов оценки:** В целом, выборочные коэффициенты корреляции Пирсона , Спирмена и квадрантный коэффициент корреляции демонстрируют схожие результаты, однако их поведение может немного различаться в зависимости от конкретной выборки и параметров распределения.

**6 . Простая линейная регрессия**

**6.0.1 Модель простой линейной регрессии**

Регрессионную модель описания данных называют простой линейной регрессией, если

где — заданные числа (значения фактора); — наблюдаемые значения отклика; — независимые, нормально распределённые с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); — неизвестные параметры, подлежащие оцениванию.

В модели зависит от одного фактора , и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика . Погрешности результатов измерений в этой модели полагают существенно меньшими погрешностей результатов измерений , так что ими можно пренебречь.

**6.0.2 Метод наименьших квадратов**

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распространённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

Задача минимизации квадратичного критерия называется задачей метода наименьших квадратов (МНК), а оценки параметров , реализующие минимум критерия, называют МНК-оценками.

**6.0.3 Расчётные формулы для МНК-оценок**

МНК-оценки параметров и находятся из условия обращения функции в минимум.

Для нахождения МНК-оценок выпишем необходимые условия экстремума:

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы получим:

Разделим оба уравнения на :

и, используя известные статистические обозначения для выборочных первых и вторых начальных моментов:

получим:

откуда МНК-оценку наклона прямой регрессии находим по формуле Крамера:

а МНК-оценку определяем непосредственно из первого уравнения системы:

Заметим, что определитель системы:

если среди значений есть различные, что и будем предполагать.

Доказательство минимальности функции в стационарной точке проведём с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

Этот результат вместе с условием означает, что в стационарной точке функция имеет минимум.

**6.1 Робастные оценки коэффициентов линейной регрессии**

Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию в данных редких, но больших по величине выбросов) может быть обеспечена различными способами. Одним из них является использование метода наименьших модулей вместо метода наименьших квадратов:

Напомним, что использование метода наименьших модулей в задаче оценивания параметра сдвига распределений приводит к оценке в виде выборочной медианы, обладающей робастными свойствами. В отличие от этого случая и от задач метода наименьших квадратов, на практике задача решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу.

Здесь мы рассмотрим простейшую в вычислительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок в другом виде:

В формулах заменим выборочные средние и соответственно на робастные выборочные медианы , среднеквадратические отклонения и на робастные нормированные интерквартильные широты и , выборочный коэффициент корреляции — на знаковый коэффициент корреляции :

где

\begin{equation}

Уравнение регрессии здесь имеет вид:

Статистики выборочной медианы и интерквартильной широты обладают робастными свойствами в силу того, что основаны на центральных порядковых статистиках, малочувствительных к большим по величине выбросам в данных. Статистика выборочного знакового коэффициента корреляции робастна, так как знаковая функция чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии обладает очевидными робастными свойствами устойчивости к выбросам по координате , но она довольно груба.

**6.2 Задание**

Найти оценки коэффициентов линейной регрессии , используя 20 точек на отрезке с равномерным шагом равным 0.2. Ошибку считать нормально распределённой с параметрами . В качестве эталонной зависимости взять . При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей. Проделать то же самое для выборки, у которой в значения и вносятся возмущения 10 и -10.

**6.3 Результаты**

В результате анализа методом наименьших квадратов (МНК) и методом наименьших модулей (МНМ) для данных с выбросами и без выбросов были получены следующие коэффициенты регрессии:

Без выбросов:

МНК:

МНМ:

С выбросами:

МНК:

МНМ:

Изображение выглядит как текст, линия, диаграмма, карта

Автоматически созданное описание

Figure 7 оценки коэффициентов линейной регрессии

**6.4 Выводы**

Для данных без выбросов оба метода (МНК и МНМ) дают схожие результаты, близкие к истинным значениям коэффициентов , . Это подтверждает, что оба метода корректно оценивают параметры модели при отсутствии выбросов.

Для данных с выбросами метод наименьших квадратов (МНК) показывает значительное смещение, особенно для коэффициента (с 1.502 до 0.074), что указывает на сильную чувствительность МНК к выбросам.

Метод наименьших модулей (МНМ), напротив, показывает более стабильные результаты в присутствии выбросов. Коэффициенты и изменяются незначительно (с 1.706 и 1.561 до 1.817 и 1.407 соответственно), что указывает на устойчивость МНМ к выбросам.

Метод наименьших модулей (МНМ) является более предпочтительным для оценки параметров линейной регрессии в присутствии выбросов, так как он демонстрирует большую устойчивость и меньшее смещение по сравнению с методом наименьших квадратов (МНК).

## **7. Реализация**

Ссылка на GitHub-репозиторий: https://github.com/1Qwix1/MathStatistics