МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий
кафедра прикладной математики

КУРСОВАЯ РАБОТА

по дисциплине Прикладные задачи математической статистики
(наименование дисциплины)

Тема курсовой работы Исследование статистических методов анализа данных

Студент группы ИМБО-01-20, Кильдишев А. С.
(учебная группа Ф.И.О. студента) *(подпись студента)*

Руководитель курсовой работы Буданцев А.В.
(должность, звание, учёная степень Ф.И.О.) *(подпись руководителя)*

Рецензент (при наличии)
(должность, звание, учёная степень Ф.И.О.) *(подпись рецензента)*

Работа представлена к защите «02» мая 2023 г.

Допущен к защите «02» мая 2023 г.

Москва 2023 г.

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий
кафедра прикладной математики

Утверждаю

Заведующий кафедрой **Дзержинский Роман Игоревич** ____________

Ф.И.О. *подпись*

«02» мая 2023 г.

ЗАДАНИЕ
на выполнение курсовой работы по дисциплине

Прикладные задачи математической статистики

Студент Кильдишев Александр Степанович Группа ИМБО-01-20

Тема: Исследование статистических методов анализа данных

Исходные данные: Исходные данные выданы в соответствии с вариантом

Перечень вопросов, подлежащих разработке, и обязательного графического материала: 1. Расчет статистических параметров. 2. Интервальные оценки параметров. Доверительные интервалы точечных оценок. 3. Идентификация распределений. 4. Генерация распределений. Проверка определений известных распределений. 5. Линейная регрессия. Оценка адекватности модели, оценка доверительных интервалов параметров. 6. Сглаживание временных рядов.

Срок представления к защите курсовой работы до «20» декабря 2023 г.

Задание на курсовую работу выдал ____________ Бузонцев Д.В.

подпись руководителя *Ф.И.О. руководителя*

Задание на курсовую работу получил ____________ «15» 09 2022 г.
Кильдишев А.С.

подпись обучающегося *Ф.И.О. обучающегося*

ОТЗЫВ

на курсовую работу по дисциплине

Прикладные задачи математической статистики

Студент группы ИМБО-01-20 Кильдишев Александр Степанович

(ФИО студента)

Характеристика курсовой работы

Критерий	Да	Нет	Не полностью
1. Соответствие содержания курсовой работы указанной теме			
2. Соответствие курсовой работы заданию			
3. Соответствие рекомендациям по оформлению текста, таблиц, рисунков и пр.			
4. Полнота выполнения всех пунктов задания			
5. Логичность и системность содержания курсовой работы			
6. Отсутствие фактических грубых ошибок			

Замечания:

Рекомендуемая оценка:

подпись руководителя *ФИО руководителя*

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современных условиях развития общества значительно вырос интерес к статистике как науке и ее широкому применению в практической деятельности. Сегодня уже никто не может отрицать значение и недооценивать роль статистики в общественной жизни.

Статистика, как любая другая наука, возникла из практических потребностей людей. Возникновение и развитие капитализма потребовало обширной и достоверной информации о состоянии производства, источниках сырья, рынках труда и сбыта продукции и т.п. Накапливается опыт в сборе, систематизации и обработке первичных статистических материалов. Появляется потребность в их анализе для выявления закономерностей общественного развития.

Без статистической информации невозможно познание закономерностей природных и социальных массовых явлений, их предвидение, а значит, и регулирование либо прямое управление, будь то на уровне отдельного предприятия, города или региона, на государственном или межгосударственном уровне.

Поэтому данная тема курсовой работы является актуальной для всех сфер, где требуется анализ данных.

Задачи, решаемые в данной курсовой работе:

- расчет статистических параметров;
- построение доверительного интервала точечных оценок;
- идентификация распределений;
- генерация распределений;
- оценка адекватности моделей;
- сглаживание временных рядов.

Цель курсовой работы — применение знаний, полученных на курсе «Прикладные задачи математической статистики».

1 РАСЧЕТ СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ

1.1 Постановка задачи

1. Необходимо собрать данные о месяцах рождения студентов группы.
2. Необходимо собрать данные о загаданных каждым студентом группы числах в диапазоне от 0 до 8.
3. Построить полигон относительной частоты для дискретного ряда распределения месяцев рождения студентов.
4. Построить гистограмму для дискретного ряда распределения длины рук студентов.
5. Найти выборочное среднее для обоих рядов распределений.
6. Найти выборочную дисперсию для обоих рядов распределений.
7. Найти выборочное СКО для обоих рядов распределений.
8. Найти коэффициент вариации для обоих рядов распределений.
9. Найти эмпирическую функцию распределения для обоих распределений.

Выборочное среднее считается по формуле (1.1):

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

Выборочная дисперсия считается по формуле (1.2):

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_i)^2 \tag{1.2}$$

Выборочное среднеквадратичное отклонение считается по формуле (1.3):

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} \tag{1.3}$$

Коэффициент вариации считается по формуле (1.4):

$$V_{\text{в}} = \frac{\sigma_{\text{в}}}{\bar{x}_{\text{в}}} \cdot 100\% \tag{1.4}$$

Выполним задачу в соответствии с заданием, в ходе выполнения воспользуемся собранным инструментарием.

1.2 Выполнение задачи

Начнём выполнение задачи с ряда распределения загаданных студентами чисел. Для полученных в результате опроса данных запишем относительную частоту и запишем их в удобном виде (Таблица 1.1).

Таблица 1.1 —Частота выбора случайного числа от 0 до 8

X_i	0	1	2	3	4	5	6	7
N_i	2	1	3	4	1	3	4	3
W_i	0.1	0.05	0.15	0.20	0.05	0.15	0.20	0.15

Построим полигон относительной частоты выбора студентами случайного числа от 0 до 8 (Рисунок 1.1). По оси X отложим загаданные числа, а по оси Y относительную частоту их выбора.

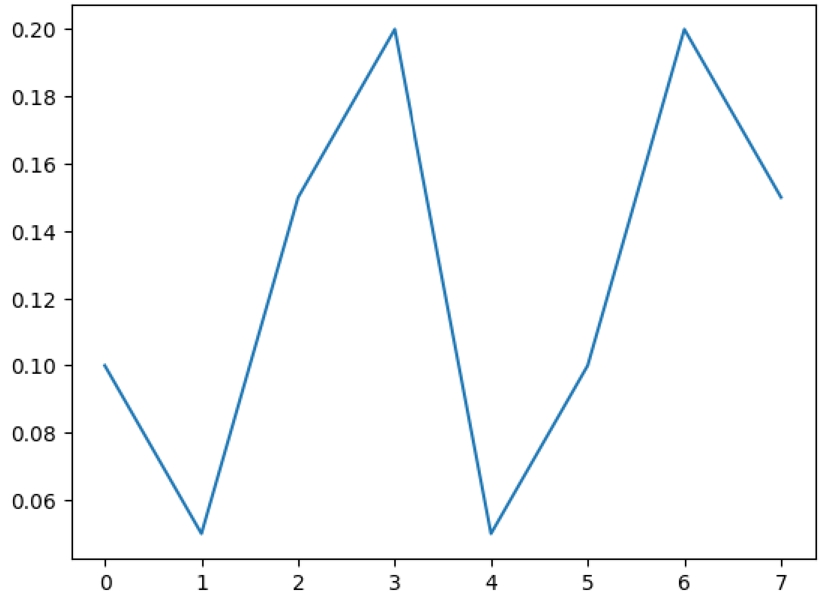

Рисунок 1.1 — Полигон относительной частоты выбора случайного числа от 0 до 8

Найдем выборочное среднее по формуле выборочного среднего для вариационного ряда(1.1).

Получаем: $\bar{x}_{\text{в}} = \frac{1}{20}(0 \cdot 2 + 1 \cdot 1 + .. + 6 \cdot 4 + 7 \cdot 3) = 3.9$

Найдем выборочную дисперсию по формуле выборочной дисперсии для вариационного ряда(1.2).

Получаем:$D_{\text{в}} = \frac{1}{20}(2(0 - \bar{x}_{\text{в}})^2 + 1(1 - \bar{x}_{\text{в}})^2 + .. + 4(6 - \bar{x}_{\text{в}})^2 +$
$3(7 - \bar{x}_{\text{в}})^2) = \frac{1}{20}(2(0 - 3.9)^2 + 1(1 - 3.9)^2 + .. + 4(6 - 3.9)^2 +$
$3(7 - 3.9)^2) = 5.09$

Найдем СКО по формуле среднеквадратичного отклонения для вариационного ряда(1.3).

Получаем: $\sigma_{\text{в}} = \sqrt{5.09} = 2.2561$

Найдем коэффициент вариации по формуле коэффициента вариации для вариационного ряда(1.4).

Получаем: $V_{\text{в}} = \frac{2.2561}{3.9} \cdot 100\% = 57.8488$

Найдём и построим эмпирическую функцию распределения (Рисунок 1.2)

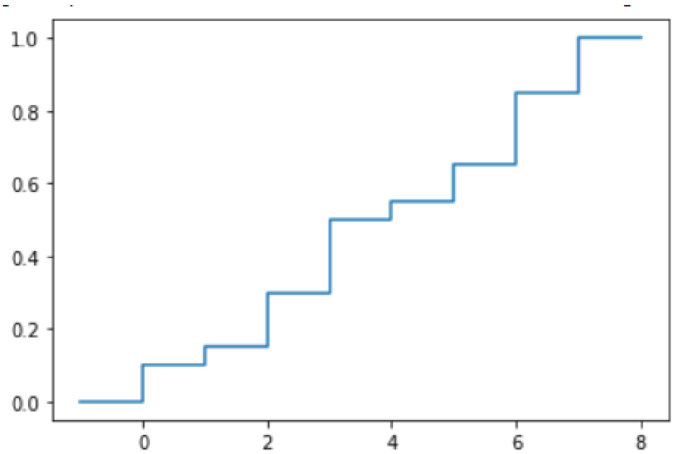

Рисунок 1.2 — Эмпирическая функция распределения частоты выбора случайного числа от 0 до 8

Приведем формулу (1.5) эмпирической функции распределения, по которой был построен график.

$$F^*(x) = \begin{cases} 0, x < 0 \\ 0.1, 0 \leq x < 1 \\ 0.15, 1 \leq x < 2 \\ 0.3, 2 \leq x < 3 \\ 0.5, 3 \leq x < 4 \\ 0.55, 4 \leq x < 5 \\ 0.65, 5 \leq x < 6 \\ 0.85, 6 \leq x < 7 \\ 1, x > 7 \end{cases} .(1.5)$$

Таким образом, задача для первой выборки полностью выполнена. Приступим к выполнению задания для распределения длины рук студентов.

Полученные в результате опроса данные разделим на классы и запишем в удобном виде (Таблица 1.2).

Таблица 1.2 —Длина рук студентов и её разделение по классам

$\{X_i\}$	{67, 70, 70}	{71, 71, 74, 74}	{75, 75, 75, 75, 77, 77}	{79, 79, 79, 79, 80, 81}	{86}
N_i	3	4	6	6	1
$[a_i, b_i]$	[67, 70.8)	[70.8, 74.6)	[74.6, 78.4)	[78.4, 82.2)	[82.2, 86]

Построим гистограмму классов распределения длины рук студентов (Рисунок 1.3). По оси X отложим длину рук, а по оси Y количество элементов в получившихся классах.

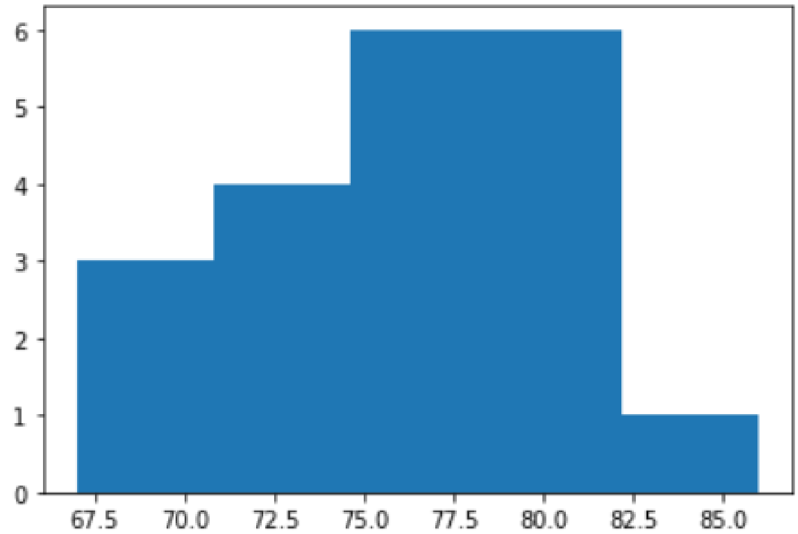

Рисунок 1.3 — Гистограмма классов распределения длины рук студентов

Найдем выборочное среднее по формуле выборочного среднего для вариационного ряда(1.1).

Получаем: $\bar{x}_{\text{в}} = \frac{1}{20}(67 + 2 \cdot 70 +..+ 81 + 86) = 75.7$

Найдем выборочную дисперсию по формуле выборочной дисперсии для вариационного ряда(1.2).

Получаем:$D_{\text{в}} = \frac{1}{20}((67 - \bar{x}_{\text{в}})^2 + 2(70 - \bar{x}_{\text{в}})^2 +..+ (81 - \bar{x}_{\text{в}})^2 + (86 - \bar{x}_{\text{в}})^2) = \frac{1}{20}((67 - 75.7)^2 + 2(70 - 75.7)^2 +..+ (81 - 75.7)^2 + (86 - 75.7)^2) = 18.804$

Найдем СКО по формуле среднеквадратичного отклонения для вариационного ряда(1.3).

Получаем: $\sigma_{\text{в}} = \sqrt{18.804} = 4.3364$

Найдем коэффициент вариации по формуле коэффициента вариации для вариационного ряда(1.4).

Получаем: $V_{\text{в}} = \frac{4.3364}{75.7} \cdot 100\% = 5.7283$

Найдём и построим эмпирическую функцию распределения (Рисунок 1.4)

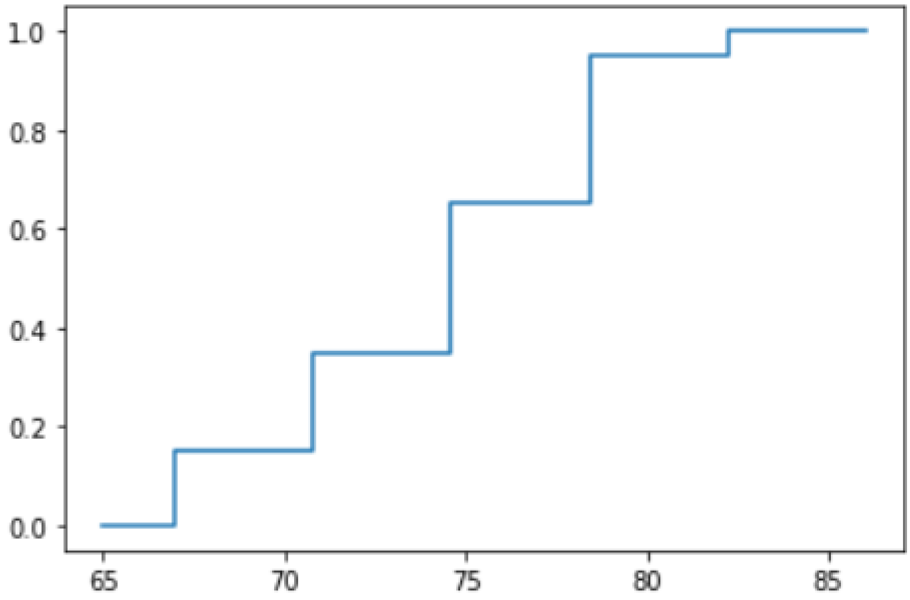

Рисунок 1.4 — Эмпирическая функция распределения классов длины рук студентов

Приведем формулу (1.6) эмпирической функции распределения, по которой был построен график.

$$F^*(x) = \begin{cases} 0, x < 67 \\ 0.15, 67 \leq x < 70.8 \\ 0.35, 70.8 \leq x < 74.6 \\ 0.65, 74.6 \leq x < 78.4 \\ 0.95, 78.4 \leq x < 82.2 \\ 1, x \geq 82.2 \end{cases} \quad .(1.6)$$

Таким образом, задача для второй выборки полностью выполнена.

По итогу мы провели работу по сбору данных и по расчёту статистических параметров для них. Рассчитанные параметры включают выборочное среднее, выборочную дисперсию, выборочное среднеквадратичное отклонение, а также коэффициент вариации. По полученным вариационным рядам нами были построены полигоны частот и эмпирические функции распределений. Для наглядности нами были построены графики найденных эмпирических функций.

2 ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ТОЧЕЧНЫХ ОЦЕНОК

2.1 Постановка задачи

1. Скачать папку с исходными данными по ссылке.
2. Открыть папку, соответствующую номеру своей группы.
3. Открыть папку, соответствующую номеру своего варианта.
4. В папке data можете найти 4 ряда данных реализации случайной величины.
5. Для каждого из четырех рядов данных необходимо провести следующие расчёты:

- подсчитать выборочные статистики для среднего и стандартного отклонения;
- для выборочного среднего $\bar{x}_в$ подсчитать границы доверительного интервала по правилу нормального распределения, используя таблицу критических значений функции Лапласа $\Phi(x)$и по правилу t-распределения Стьюдента используя таблицу критических значений $t_{\gamma,n}$t-распределения при значении уверенности γ = 0.95;
- для выборочного среднеквадратического отклонения $\sigma_в$ подсчитать границы доверительного интервала по оценке χ^2-распределения при значении уверенности γ = 0.95.

Выборочное среднее вычисляется по формуле (2.1):

$$\bar{x}_в = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2.1}$$

где N — число значений реализации случайной величины;

$xi \in \mathbb{R}$, $i \in \overline{1, N}$ —реализации нашей случайной величины.

Выборочное среднеквадратическое отклонение вычисляется по формуле (2.2):

$$\sigma_{в} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x}_{в})^2}{N-1}}. \tag{2.2}$$

Границы доверительного интервала по правилу нормального распределения находятся по формуле (2.3):

$$\hat{x}_{в} \in \left[\bar{x}_{в} - x_{\gamma}\frac{\sigma_{в}}{\sqrt{N}}, \bar{x}_{в} + x_{\gamma}\frac{\sigma_{в}}{\sqrt{N}}\right], \Phi(x_{\gamma}) = \frac{\gamma}{2}. \tag{2.3}$$

Границы доверительного интервала по правилу t-распределения находятся по формуле (2.4):

$$\hat{x}_{в} \in \left[\bar{x}_{в} - t_{(1-\gamma,N-1)} \cdot \frac{\sigma_{в}}{\sqrt{N}}, \bar{x}_{в} + t_{(1-\gamma,N-1)} \cdot \frac{\sigma_{в}}{\sqrt{N}}\right]. \tag{2.4}$$

Границы доверительного интервала по оценке χ^2-распределения находится по формуле (2.5):

$$\sigma_{в} \in \left[\frac{\sigma_{в} \cdot \sqrt{N-I}}{\sqrt{\chi^2_{\left(\frac{1+\gamma}{2},N-1\right)}}}, \frac{\sigma_{в} \cdot \sqrt{N-I}}{\sqrt{\chi^2_{\left(\frac{1-\gamma}{2},N-1\right)}}}\right] \tag{2.5}$$

Выполним задачу в соответствии с заданием, в ходе выполнения воспользуемся собранным инструментарием.

2.2 Выполнение задачи

Рассмотрим ход решения на примере второго ряда данных (Таблица 2.1). Решения для остальных рядов будут идентичны.

Таблица 2.1 — Второй ряд данных

i	1	2	3	4	5	6	7	8	9	10
X_i	42.06	47.41	41.98	26.06	30.3	40.61	35.98	30.7	49.62	37.44

Найдём выборочное среднее по формуле выборочного среднего для ряда(2.1).

Получим: $\bar{x}_{\text{в}} = \frac{1}{10}(42.06 + 47.41 + .. + 49.62 + 37.44) = 38.22$

Найдем выборочное СКО по формуле выборочного среднеквадратичного отклонения(2.2).

Получим: $\sigma_{\text{в}} = \sqrt{\frac{(42.06-38.22)^2+(47.41-38.22)^2+..+(37.44-38.22)^2}{10-1}} = \sqrt{\frac{(3.84)^2+(9.19)^2+..+(-1.22)^2}{9}} = 7.615$

Найдём границы доверительного интервала для выборочного среднего по правилу нормального распределения, воспользовавшись формулой (2.3).

$$\bar{x}_{\text{в}} \in \left[38.22 - 1.96\frac{7.615}{\sqrt{10}}, 38.22 + 1.96\frac{7.615}{\sqrt{10}}\right] => \bar{x}_{\text{в}} \in [33.496, 42.935]$$

Найдём границы доверительного интервала для выборочного среднего по правилу Т-распределения Стьюдента, воспользовавшись формулой (2.4).

$$\bar{x}_{\text{в}} \in \left[38.22 - 2.23\frac{7.615}{\sqrt{10}}, 38.22 + 2.23\frac{7.615}{\sqrt{10}}\right] => \bar{x}_{\text{в}} \in [32.846, 43.586]$$

Найдём границы доверительного интервала для выборочного среднеквадратического отклонения по оценке χ^2-распределения, воспользовавшись формулой (2.5).

$$\sigma_{\text{в}} \in \left[\frac{7.615 \cdot \sqrt{9}}{\sqrt{\chi^2_{\left(\frac{1+0.95}{2},9\right)}}}, \frac{7.615 \cdot \sqrt{9}}{\sqrt{\chi^2_{\left(\frac{1-0.95}{2},9\right)}}}\right] => \sigma_{\text{в}} \in [5.238,13.902]$$

Таким образом, задача для второго набора данных выполнена. Далее проделаем аналогичные вычисления для остальных наборов.

Результаты вычислений среднего выборочного и выборочного СКО для первого набора данных:

$$\bar{x}_{\text{в}} = 25.55$$

$$\sigma_{\text{в}} = 7.81$$

Результат вычисления границ доверительного интервала среднего выборочного по правилу нормального распределения для первого набора данных:

$$\bar{x}_{\text{в}} \in [24.023, 27.085]$$

Результат вычисления границ доверительного интервала среднего выборочного по правилу Т-распределения для первого набора данных::

$$\bar{x}_{\text{в}} \in [24.0,27.109]$$

Результат вычисления границ доверительного интервала выборочного среднеквадратичного отклонения по оценке χ^2-распределения для первого набора данных:

$$\sigma_{\text{в}} \in [6.858,9.073]$$

Результаты вычислений среднего выборочного и выборочного СКО для третьего набора данных:

$$\bar{x}_{\text{в}} = 45.72$$

$$\sigma_{\text{в}} = 3.155$$

Результат вычисления границ доверительного интервала среднего выборочного по правилу нормального распределения для третьего набора данных:

$$\bar{x}_{\text{в}} \in [45.104,46.341]$$

Результат вычисления границ доверительного интервала среднего выборочного по правилу Т-распределения для третьего набора данных:

$$\bar{x}_{в} \in [45.095, 46.351]$$

Результат вычисления границ доверительного интервала выборочного среднеквадратичного отклонения по оценке χ^2-распределения для третьего набора данных:

$$\sigma_{в} \in [2.770, 3.665]$$

Результаты вычислений среднего выборочного и выборочного СКО для четвёртого набора данных:

$$\bar{x}_{в} = 71.592$$

$$\sigma_{в} = 5.016$$

Результат вычисления границ доверительного интервала среднего выборочного по правилу нормального распределения для четвёртого набора данных:

$$\bar{x}_{в} \in [68.484,74.701]$$

Результат вычисления границ доверительного интервала среднего выборочного по правилу Т-распределения для четвёртого набора данных:

$$\bar{x}_{в} \in [68.055,75.129]$$

Результат вычисления границ доверительного интервала выборочного среднеквадратичного отклонения по оценке χ^2-распределения для четвёртого набора данных:

$$\sigma_{в} \in [3.450, 9.157]$$

По итогу мы провели работу по постройке доверительных интервалов для вычислений среднего выборочного и выборочного СКО четырёх наборов данных.

3 ИДЕНТИФИКАЦИЯ РАСПРЕДЕЛЕНИЙ

3.1 Постановка задачи

Для выполнения практического задания необходимо:

1. Скачать папку с исходными данными по ссылке. Открыть папку, соответствующую своей группе. Далее открыть папку с вариантом, совпадающим с вашим номером в списке.
2. В папке 6 файлов с данными. Данные имеют различные распределения:

- 1 и 4 файлы - нормальное распределение;
- 2 и 5 файлы - равномерное распределение;
- 3 и 6 файлы - показательное распределение.

3. Методом Пирсона для каждого файла нужно проверить истинное распределение и одно ложное. При расчете теоретических частот в качестве параметров распределений брать их точечные несмещенные оценки.
4. В отчет вставить гистограммы по каждому файлу, расчетные формулы и результаты проверки распределения.
5. При проверке распределения методом анаморфоз нужно построить 3 графика для каждого из 6 файлов. Каждый график представлен в координатах соответствующей анаморфозы. Тот график, на котором достигнуто спрямление, соответствует истинному распределению. Для проверки качества спрямления необходимо построить линейный тренд (провести линейную регрессию) и показать значения коэффициента детерминации. Он должен быть близок к 1.
6. По параметрам прямой найти параметры распределения.

7. В выводах сравнить результаты, полученные двумя способами.

Математическое ожидание дискретной величины рассчитывается по формуле (3.1):

$$M = \sum_{i=1}^{N} x_i p_i \tag{3.1}$$

Среднее квадратичное отклонение дискретной величины рассчитывается по формуле (3.2):

$$\sigma = \sqrt{\sum_{i=1}^{N} p_i (x_i - M)^2} \tag{3.2}$$

Оптимальное количество интервалов, на которое разбивается случайная величина рассчитывается по формуле Стерджеса (3.3):

$$n = 1 + \lfloor log\, N \rfloor \tag{3.3}$$

Критерий Пирсона рассчитывается по формуле (3.4):

$$\chi^2 = \sum_{i=1}^{g} \frac{(n_i - n_i')^2}{n_i'}, \tag{3.4}$$

$$n_i' = N \cdot P\,(X(i)|\Theta) = N \cdot (P(Z_i \mid \Theta) - P(Z_i - 1 \mid \Theta))$$

где $P(x|\Theta) = \int_{-\infty}^{x} p(t|\Theta)\, dt$ – известная функция распределения, вычисленная с параметрами $\Theta = (\theta 1,\ \theta 2, \ldots)$, которые необходимы для расчета значения данной функции.

$P(x|\Theta)$ для нормального распределения рассчитывается по формуле (3.5):

$$p(x|\Theta) = \frac{1}{\sqrt{2n\sigma^2}} \cdot \int_{-\infty}^{x} exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \tag{3.5}$$

$P(x|\Theta)$ для равномерного распределения рассчитывается по формуле (3.6):

$$P(x|\Theta) = \begin{cases} 0, x \in (-\infty, a) \\ \frac{1}{b-a} \cdot (x-a), x \in [a, b) \\ 1, \in [b, \infty) \end{cases} \tag{3.6}$$

$P(x|\Theta)$ для показательного распределения рассчитывается по формуле (3.7):

$$P(x|\Theta) = \begin{cases} 1 - e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases} \tag{3.7}$$

Коэффициент корреляции

Выполним задачу в соответствии с заданием, в ходе выполнения воспользуемся собранным инструментарием.

3.2 Выполнение задачи

Начнём выполнение задачи с определения распределений для вариационного ряда из первого файла. Выборка состоит из 32 значений.

Разделим её на классы, количество которых найдем по формуле Стерджеса(3.1):

$$n = 1 + \lfloor log\,32 \rfloor = 6$$

Запишем полученное разделение в удобном виде, вычислим относительные частоты и середины интервалов для каждого класса (Таблица 3.1):

Таблица 3.1 — Разделение выборки из первого файла по классам

i	1	2	3	4	5	6
[a_i, b_i]	[-5.616,-4.217)	[-4.217,-2.817)	[-2.817,-1.418)	[-1.418,-0.018)	[-0.018,1.381)	[1.381,2.781]
n_i	1	4	12	8	6	1
p_i	0.03125	0.125	0.375	0.25	0.1875	0.03125
X_i	-4.916	-3.517	-2.118	-0.718	0.681	2.081

Рассчитаем математическое ожидание и СКО для вариационного ряда из первого файла по формуле математического ожидания дискретной величины(3.1) и формуле среднего квадратичного отклонения дискретной величины(3.2):

$$M = -1.374, \qquad \sigma = 1.564$$

Построим гистограмму классов, по оси OX отложим среднее значение по классам, а по оси OY отложим относительную частоту вхождения значений выборки в соответствующий класс. По вычисленным значениям математического ожидания и СКО также построим график нормального распределения (Рисунок 3.1).

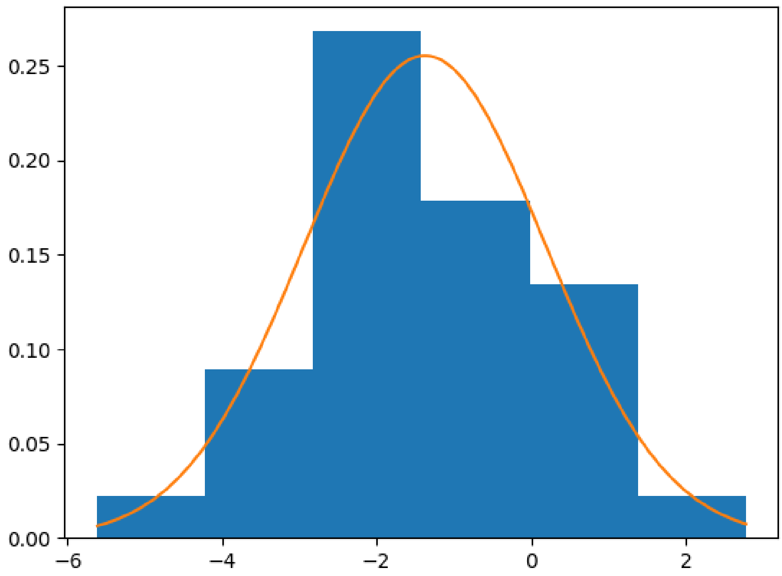

Рисунок 3.1 — Гистограмма вариационного ряда и нормальное распределение, построенное по его параметрам

Распределение из файла похоже на нормальное. Воспользуемся критерием согласия Пирсона по формуле (3.4) и проверим, является ли оно таковым.

Рассчитаем критерий Пирсона с использованием вычислительных средств. Получим значение, равное 1.2156 при критической границе 9.3484, из чего следует вывод о нормальности распределения данной выборки по критерию Пирсона.

Так же проверим, является ли распределение равномерным. По среднему выборочному построим график равномерного распределения для ряда (Рисунок 3.2).

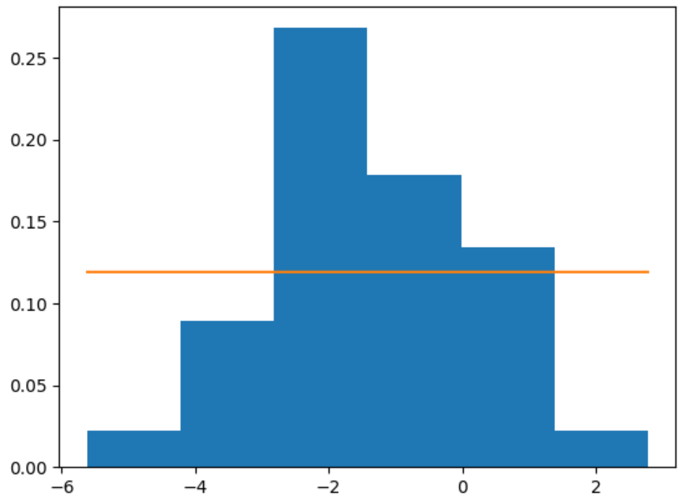

Рисунок 3.2 — Гистограмма вариационного ряда и равномерное распределение, построенное по его параметрам

При проверке на равномерное распределение по критерию Пирсона, получим значение, равное 17.125, при критической границе 9.3484, из чего следует вывод о не равномерности распределения данной выборки по критерию Пирсона.

Также для достоверности идентифицируем распределение по методу анаморфоз. Для этого построим графики распределения выборки в спрямлённых координатах и найдём квадраты коэффициентов корреляции спрямлённых графиков с эталонными графиками нормального, равномерного и показательного распределения.

Построим график выборки, полученной из первого файла, в спрямлённых для нормального распределения координатах (Рисунок 3.2).

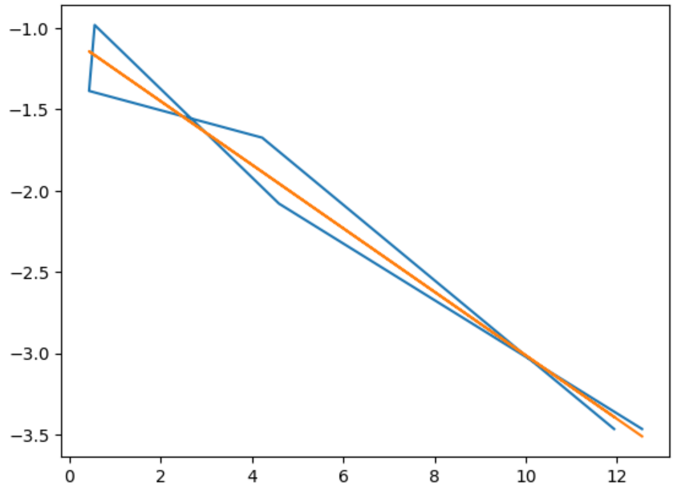

Рисунок 3.2 — График первой выборки в спрямлённых для нормального распределения координатах

Значение метрики R^2 для данного графика равно 0.97147.

Построим график выборки, полученной из первого файла, в спрямлённых для показательного распределения координатах (Рисунок 3.3).

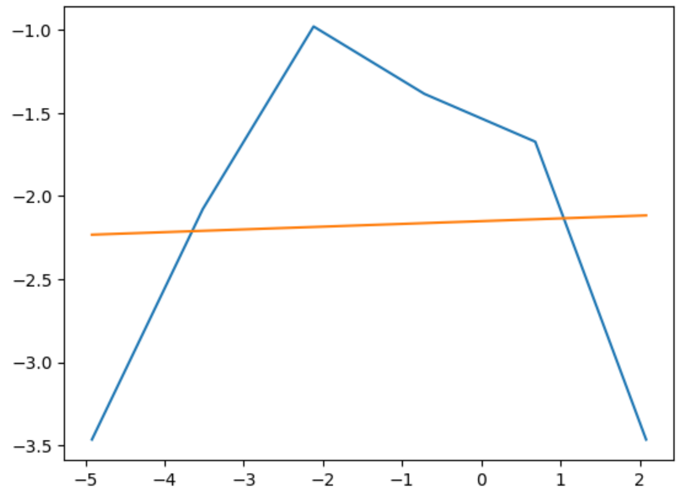

Рисунок 3.3 — График первой выборки в спрямлённых для показательного распределения координатах

Значение метрики R^2 для данного графика равно 0.00167.

Построим график выборки, полученной из первого файла, в спрямлённых для равномерного распределения координатах (Рисунок 3.4).

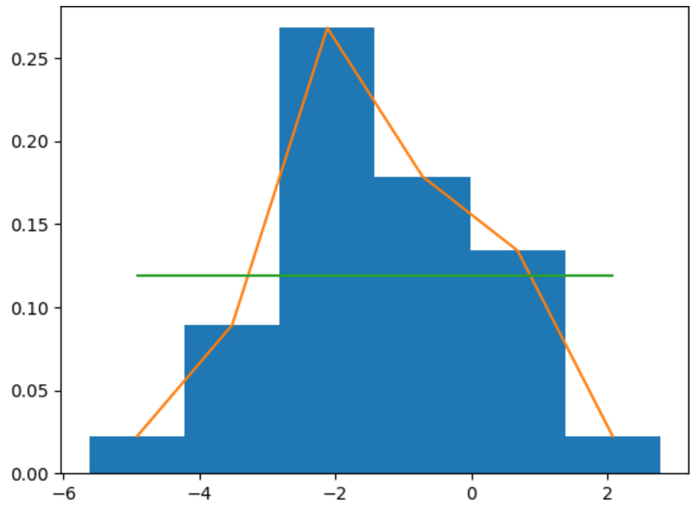

Рисунок 3.4 — График первой выборки в спрямлённых для равномерного распределения координатах

Значение метрики R^2 для данного графика равно 0.00063.

Наибольшее значение коэффициента корреляции R^2 было получено при спрямлении координат для проверки нормального распределения. Это значение равно 0.97147 и оно больше критического значения 0.85, из чего следует, что данное распределение является нормальным.

Вычислим параметры линейной регрессии, характеризующие параметры распределения:$a = -0.1954 \approx \frac{-1}{2\sigma^2}$, $b = -1.0686 \approx -\ln(\sqrt{2\pi\sigma^2})$

Проведём аналогичные исследования распределений для остальных файлов.

Построим гистограмму классов для выборки из второго файла и сравним её с нормальным распределением (Рисунок 3.5).

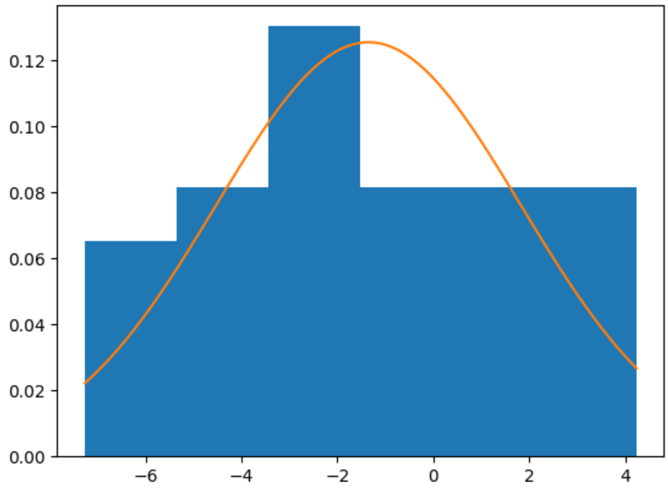

Рисунок 3.5 — Гистограмма вариационного ряда и нормальное распределение, построенное по его параметрам

Рассчитаем критерий Пирсона для проверки нормальности распределения с использованием вычислительных средств. Получим значение, равное 6.0789 при критической границе 9.3484, из чего, следует вывод о нормальности распределения по критерию Пирсона.

Так же проверим, является ли распределение равномерным. Построим график равномерного распределения для ряда (Рисунок 3.6).

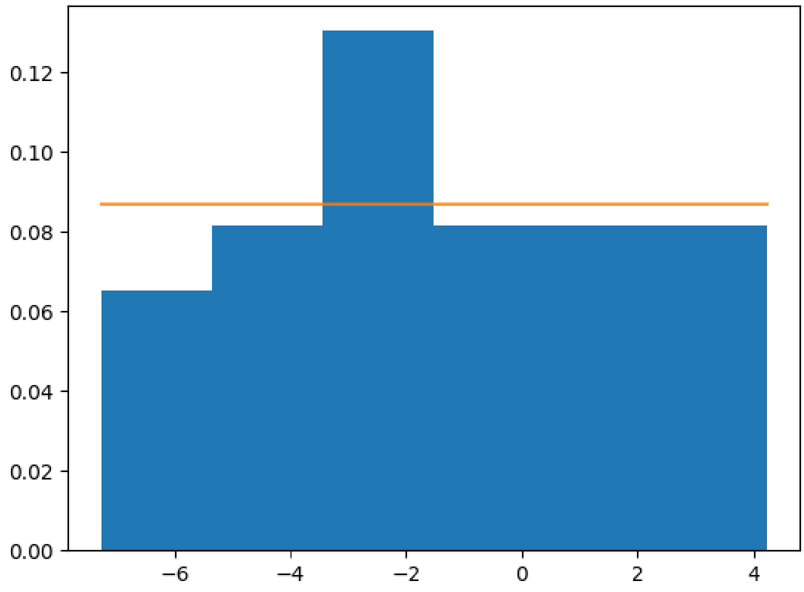

Рисунок 3.6 — Гистограмма вариационного ряда и равномерное распределение, построенное по его параметрам

При проверке на равномерное распределение по критерию Пирсона, получим значение, равное 1.8125, при критической границе 9.3484, из чего следует вывод равномерности распределения данной выборки по критерию Пирсона.

Идентифицируем распределение по методу анаморфоз.

Построим график выборки, полученной из второго файла, в спрямлённых для нормального распределения координатах (Рисунок 3.7).

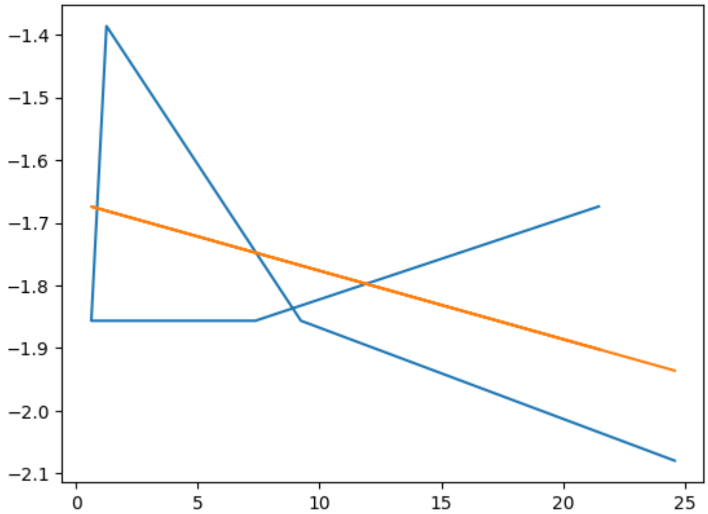

Рисунок 3.7 — График второй выборки в спрямлённых для нормального распределения координатах

Значение метрики R^2 для данного графика равно 0.22386.

Построим график выборки, полученной из второго файла, в спрямлённых для показательного распределения координатах (Рисунок 3.8).

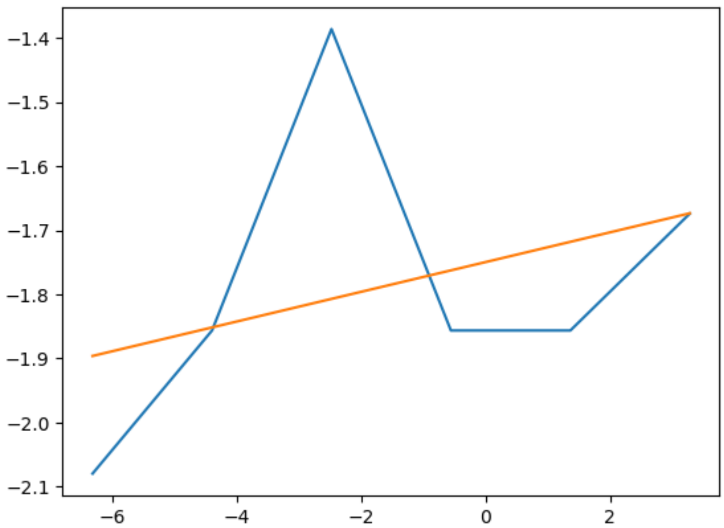

Рисунок 3.8 — График второй выборки в спрямлённых для показательного распределения координатах

Значение метрики R^2 для данного графика равно 0.12680.

Построим график выборки, полученной из второго файла, в спрямлённых для равномерного распределения координатах (Рисунок 3.9).

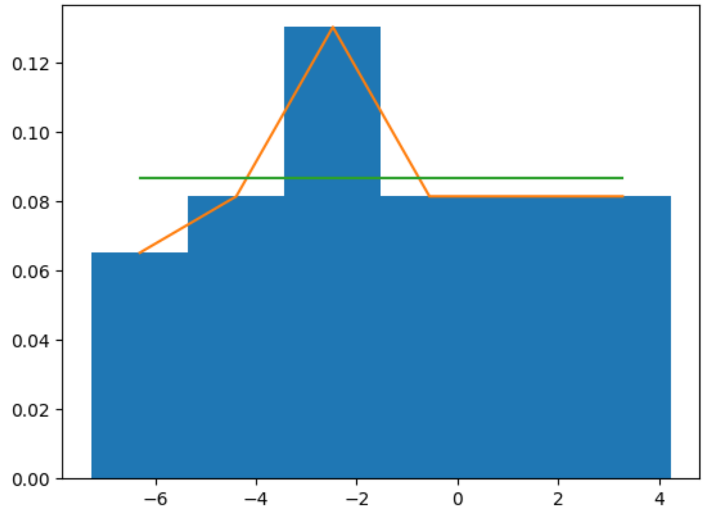

Рисунок 3.9 — График второй выборки в спрямлённых для равномерного распределения координатах

Значение метрики R^2 для данного графика равно 0.07368.

Ни одно из значений коэффициента корреляции R^2 не больше критического значения 0.85, из чего следует, что данное распределение не является ни нормальным, ни равномерным, ни показательным по методу анаморфоз.

Построим гистограмму классов для выборки из третьего файла и сравним её с показательным распределением (Рисунок 3.10).

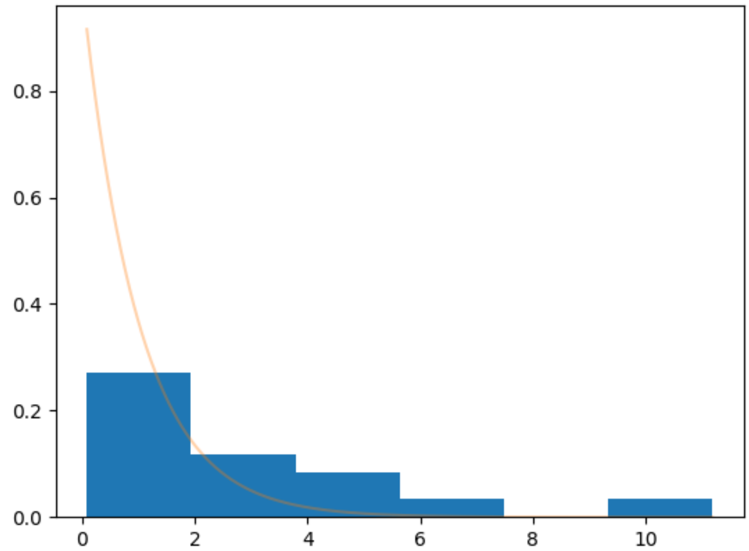

Рисунок 3.10 — Гистограмма вариационного ряда и показательное распределение, построенное по его параметрам

Рассчитаем критерий Пирсона для проверки показательности распределения с использованием вычислительных средств. Получим значение, равное 6.0789 при критической границе 9.3484, из чего следует, что распределение выборки является показательным по критерию Пирсона.

Так же проверим, является ли распределение равномерным. Построим график равномерного распределения для ряда (Рисунок 3.11).

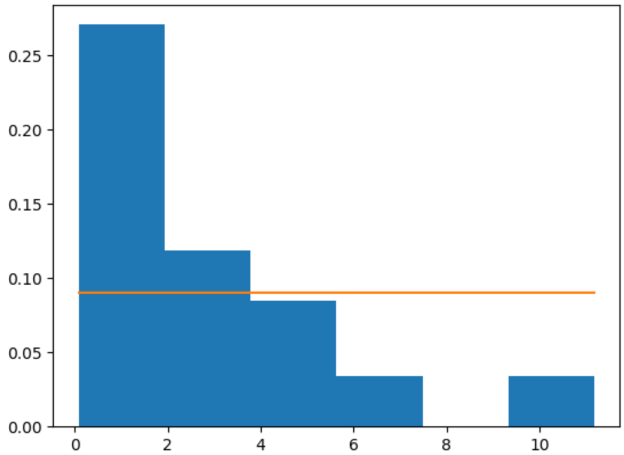

Рисунок 3.11 — Гистограмма вариационного ряда и равномерное распределение, построенное по его параметрам

При проверке на равномерное распределение по критерию Пирсона, получим значение, равное 31.375, при критической границе 9.3484, из чего следует вывод о не равномерности распределения данной выборки по критерию Пирсона.

Идентифицируем распределение по методу анаморфоз.

Построим график выборки, полученной из третьего файла, в спрямлённых для нормального распределения координатах (Рисунок 3.12).

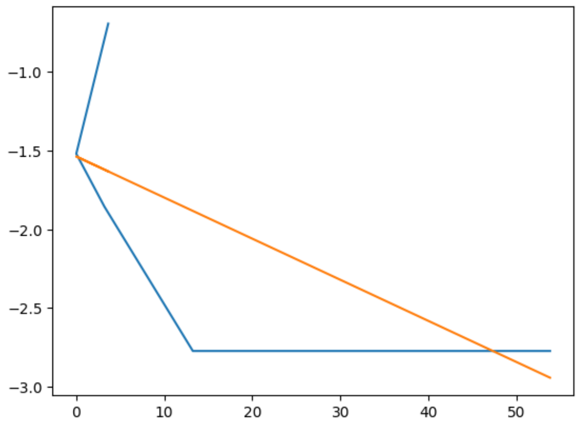

Рисунок 3.12 — График третьей выборки в спрямлённых для нормального распределения координатах

Значение метрики R^2 для данного графика равно 0.43726.

Построим график выборки, полученной из третьего файла, в спрямлённых для показательного распределения координатах (Рисунок 3.13).

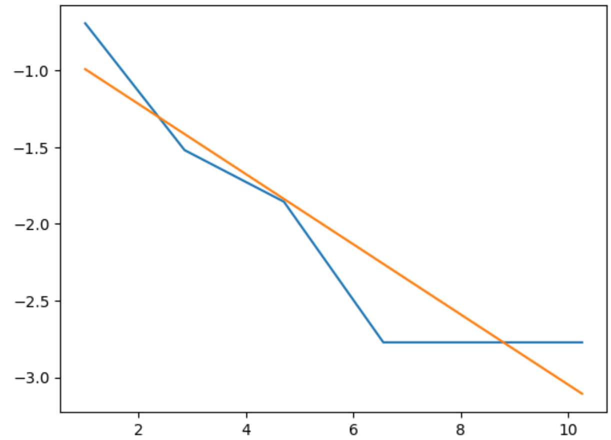

Рисунок 3.13 — График третьей выборки в спрямлённых для показательного распределения координатах

Значение метрики R^2 для данного графика равно 0.84817.

Построим график выборки, полученной из третьего файла, в спрямлённых для равномерного распределения координатах (Рисунок 3.14).

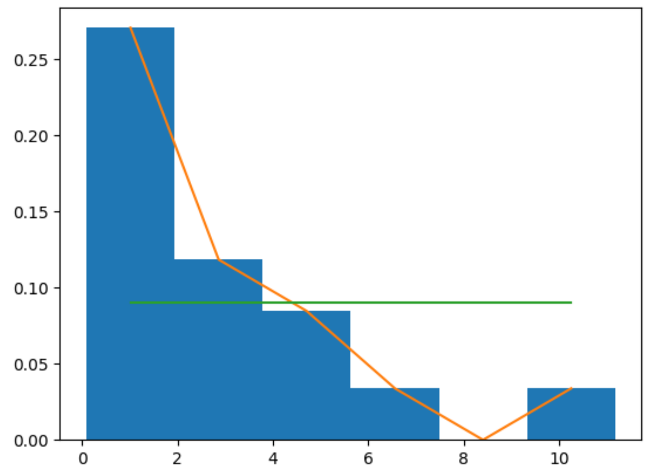

Рисунок 3.14 — График третьего выборки в спрямлённых для равномерного распределения координатах

Значение метрики R^2 для данного графика равно 0.75435.

Наибольшее значение коэффициента корреляции R^2 достигается при проверке на показательное распределение и оно равно 0.84817. Однако это значение меньше критического значения 0.85, из чего следует, что данное распределение не является ни нормальным, ни равномерным, ни показательным по методу анаморфоз.

Построим гистограмму классов для выборки из четвёртого файла и сравним её с нормальным распределением (Рисунок 3.15).

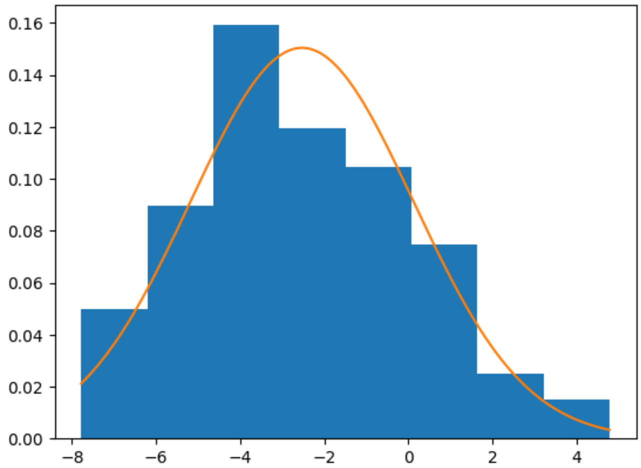

Рисунок 3.15 — Гистограмма вариационного ряда и нормальное распределение, построенное по его параметрам

Рассчитаем критерий Пирсона для проверки нормальности распределения с использованием вычислительных средств. Получим значение, равное 5.0204 при критической границе 12.8325, из чего следует, что распределение выборки является нормальным по критерию Пирсона.

Так же проверим, является ли распределение равномерным. Построим график равномерного распределения для ряда (Рисунок 3.16).

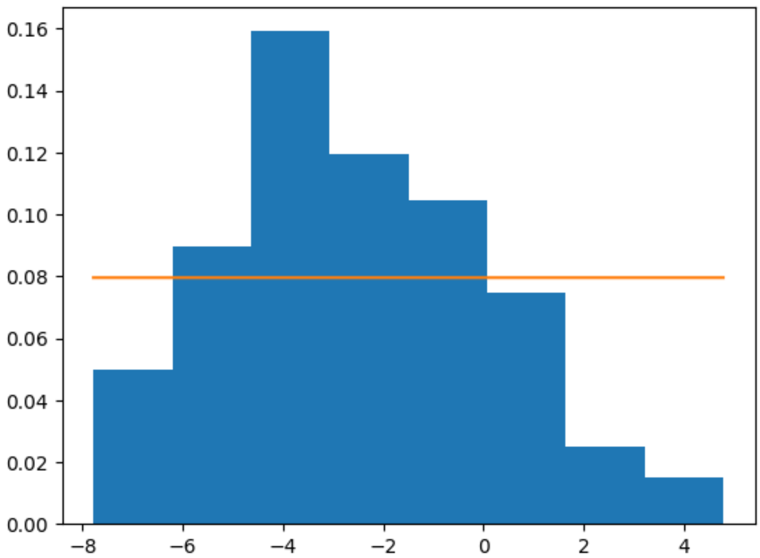

Рисунок 3.16 — Гистограмма вариационного ряда и равномерное распределение, построенное по его параметрам

При проверке на равномерное распределение по критерию Пирсона, получим значение, равное 42.25, при критической границе 12.8325, из чего следует вывод о не равномерности распределения данной выборки по критерию Пирсона.

Идентифицируем распределение по методу анаморфоз.

Построим график выборки, полученной из четвёртого файла, в спрямлённых для нормального распределения координатах (Рисунок 3.17).

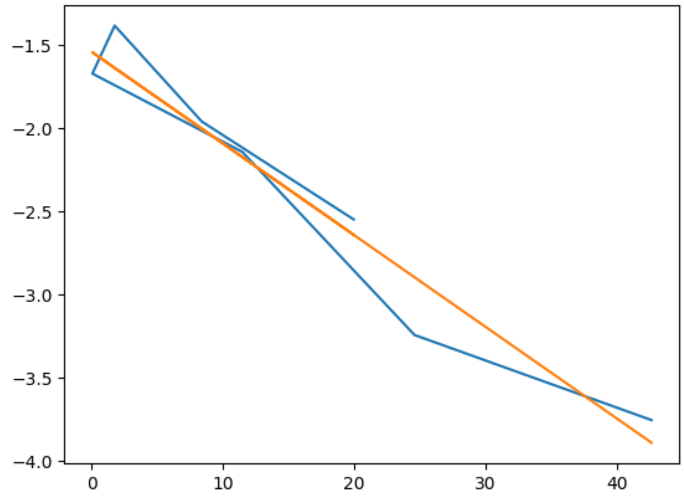

Рисунок 3.17 — График четвёртой выборки в спрямлённых для нормального распределения координатах

Значение метрики R^2 для данного графика равно 0.94921.

Построим график выборки, полученной из четвёртого файла, в спрямлённых для показательного распределения координатах (Рисунок 3.18).

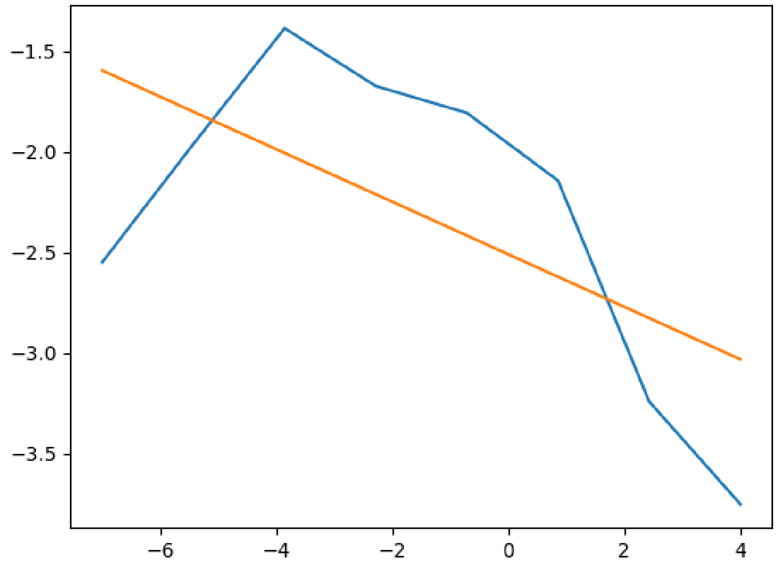

Рисунок 3.18— График четвёртой выборки в спрямлённых для показательного распределения координатах.

Значение метрики R^2 для данного графика равно 0.37883.

Построим график выборки, полученной из четвёртого файла, в спрямлённых для равномерного распределения координатах (Рисунок 3.19).

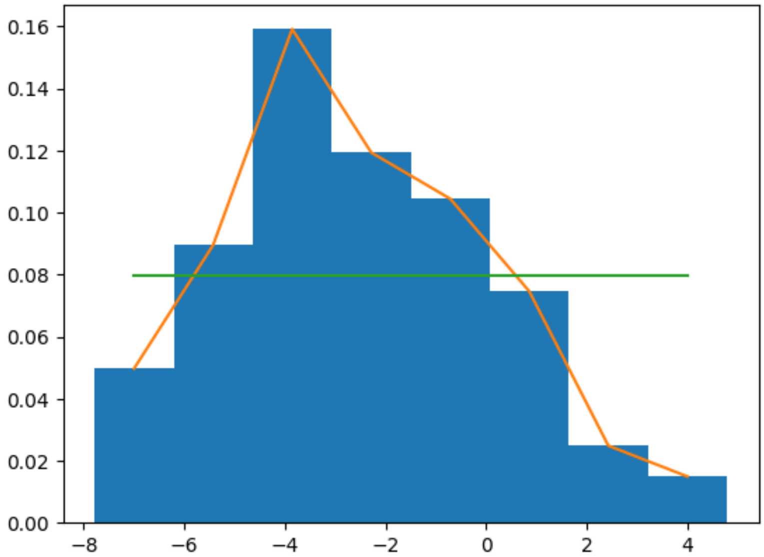

Рисунок 3.19 — График четвёртой выборки в спрямлённых для равномерного распределения координатах

Значение метрики R^2 для данного графика равно 0.24852.

Наибольшее значение коэффициента корреляции R^2 достигается при проверке на нормальное распределение и оно равно 0.94921. Это значение больше критического значения 0.85, из чего следует, что данное распределение является нормальным по методу анаморфоз.

Вычислим параметры линейной регрессии, характеризующие параметры распределения: $a = -0.05502 \approx \frac{-1}{2\sigma^2}$, $b = -1.54344 \approx -\ln(\sqrt{2\pi\sigma^2})$

Построим гистограмму классов для выборки из пятого файла и сравним её с нормальным распределением (Рисунок 3.20).

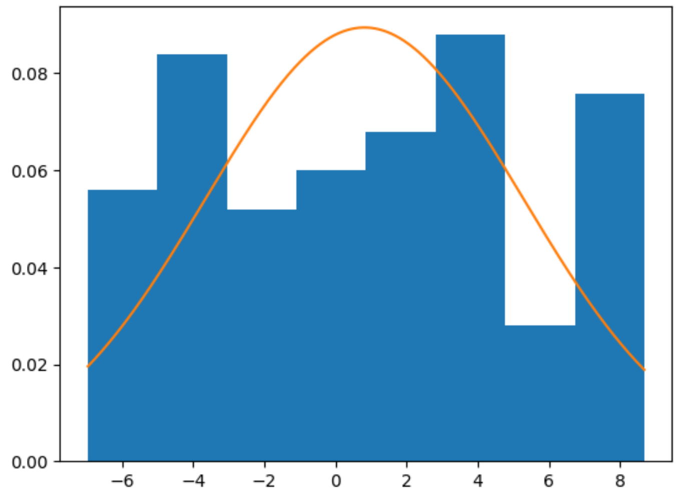

Рисунок 3.20 — Гистограмма вариационного ряда и нормальное распределение, построенное по его параметрам

Рассчитаем критерий Пирсона для проверки нормальности распределения с использованием вычислительных средств. Получим значение, равное 41.79 при критической границе 12.8325, из чего следует, что распределение выборки не является нормальным по критерию Пирсона.

Так же проверим, является ли распределение равномерным. Построим график равномерного распределения для ряда (Рисунок 3.21).

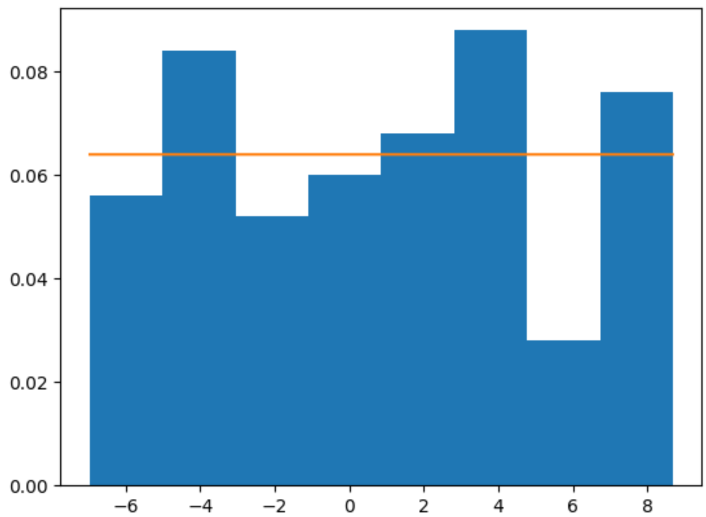

Рисунок 3.21 — Гистограмма вариационного ряда и равномерное распределение, построенное по его параметрам

При проверке на равномерное распределение по критерию Пирсона, получим значение, равное 10.375, при критической границе 12.8325, из чего следует вывод равномерности распределения данной выборки по критерию Пирсона.

Идентифицируем распределение по методу анаморфоз.

Построим график выборки, полученной из пятого файла, в спрямлённых для нормального распределения координатах (Рисунок 3.22).

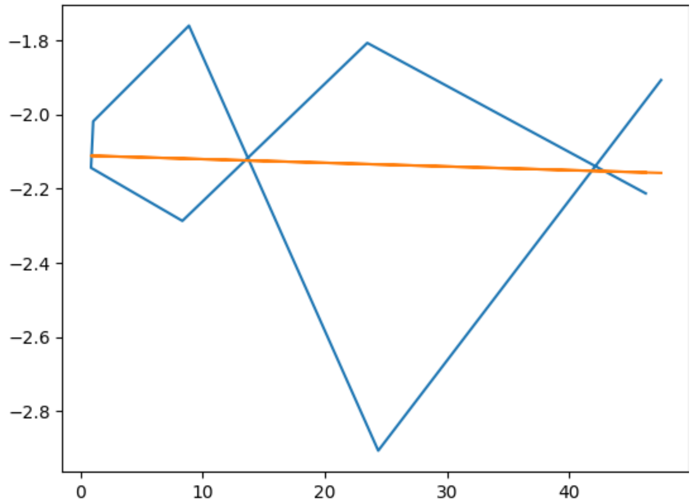

Рисунок 3.22 — График пятой выборки в спрямлённых для нормального распределения координатах

Значение метрики R^2 для данного графика равно 0.00259.

Построим график выборки, полученной из пятого файла, в спрямлённых для показательного распределения координатах (Рисунок 3.23).

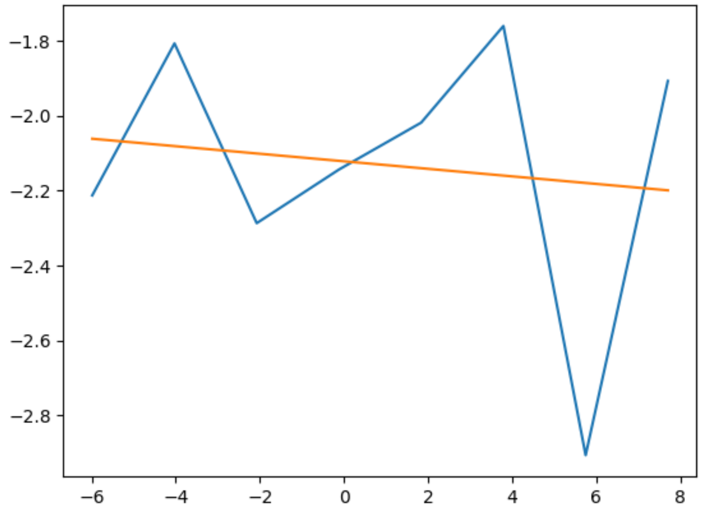

Рисунок 3.23 — График пятой выборки в спрямлённых для показательного распределения координатах

Значение метрики R^2 для данного графика равно 0.01735.

Построим график выборки, полученной из пятого файла, в спрямлённых для равномерного распределения координатах (Рисунок 3.24).

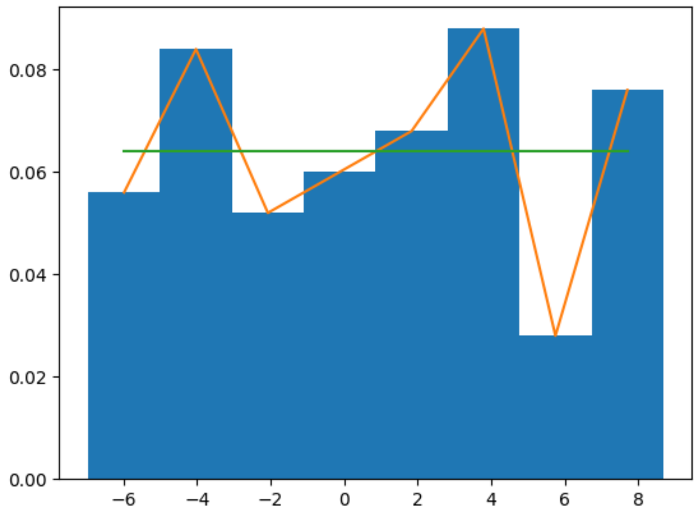

Рисунок 3.24 — График пятой выборки в спрямлённых для равномерного распределения координатах

Значение метрики R^2 для данного графика равно 0.00129.

Ни одно из значений коэффициента корреляции R^2 не больше критического значения 0.85, из чего следует, что данное распределение не является ни нормальным, ни равномерным, ни показательным по методу анаморфоз.

Построим гистограмму классов для выборки из шестого файла и сравним её с нормальным распределением (Рисунок 3.25).

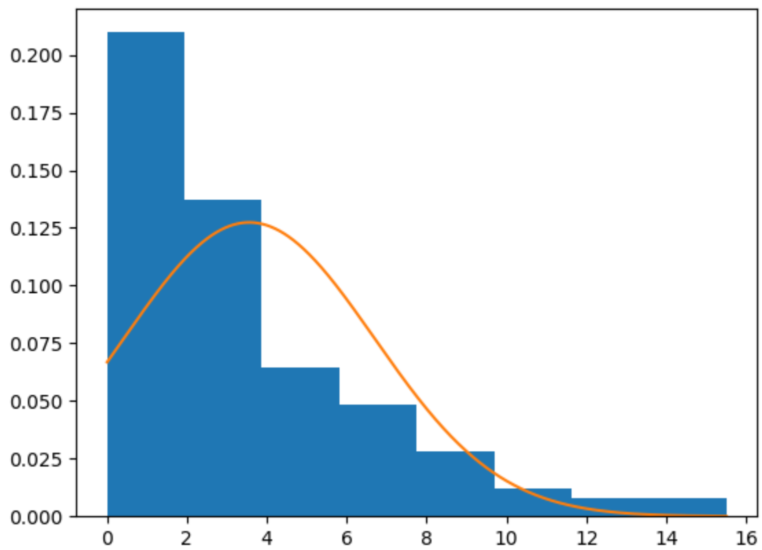

Рисунок 3.25 — Гистограмма вариационного ряда и нормальное распределение, построенное по его параметрам

Рассчитаем критерий Пирсона для проверки нормальности распределения с использованием вычислительных средств. Получим значение, равное 97.3464 при критической границе 12.8325, из чего следует, что распределение выборки не является нормальным по критерию Пирсона.

Так же проверим, является ли распределение показательным. Построим график показательного распределения для ряда (Рисунок 3.26).

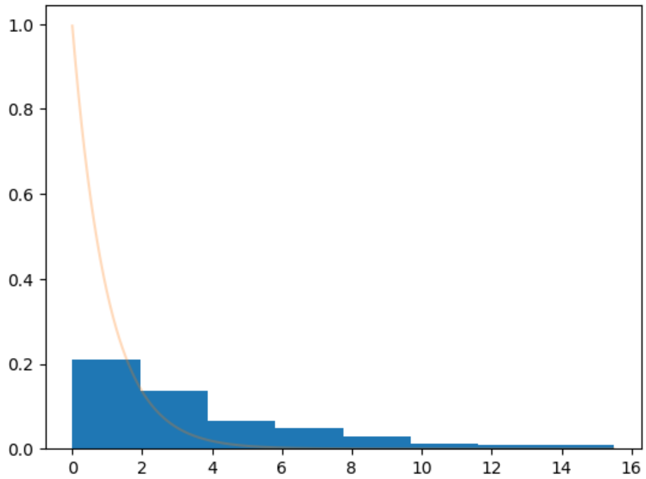

Рисунок 3.26 — Гистограмма вариационного ряда и показательное распределение, построенное по его параметрам

При проверке на равномерное распределение по критерию Пирсона, получим значение, равное 1.5385, при критической границе 14.4494, из чего следует вывод показательности распределения данной выборки по критерию Пирсона.

Идентифицируем распределение по методу анаморфоз.

Построим график выборки, полученной из шестого файла, в спрямлённых для нормального распределения координатах (Рисунок 3.27).

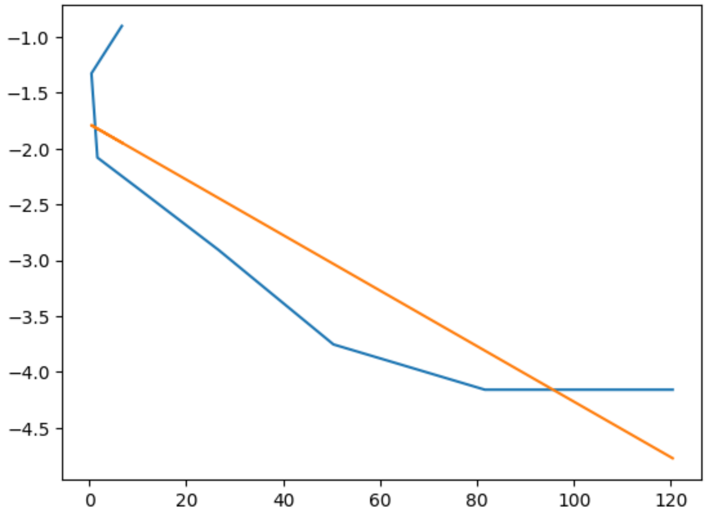

Рисунок 3.27 — График шестой выборки в спрямлённых для нормального распределения координатах

Значение метрики R^2 для данного графика равно 0.75353.

Построим график выборки, полученной из шестого файла, в спрямлённых для показательного распределения координатах (Рисунок 3.28).

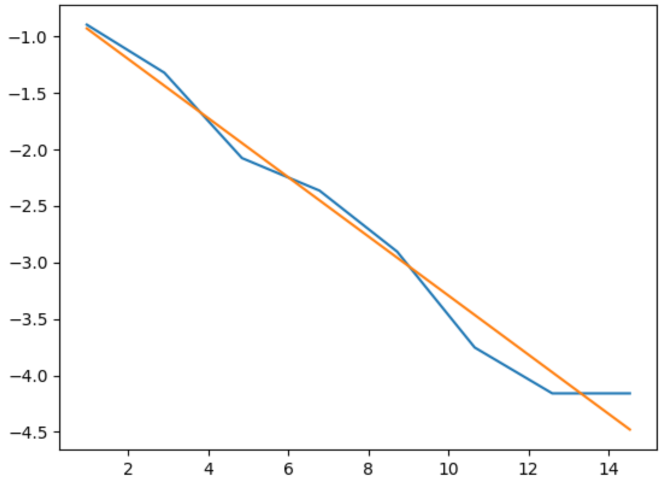

Рисунок 3.28 — График шестой выборки в спрямлённых для показательного распределения координатах

Значение метрики R^2 для данного графика равно 0.97626.

Построим график выборки, полученной из шестого файла, в спрямлённых для равномерного распределения координатах (Рисунок 3.29).

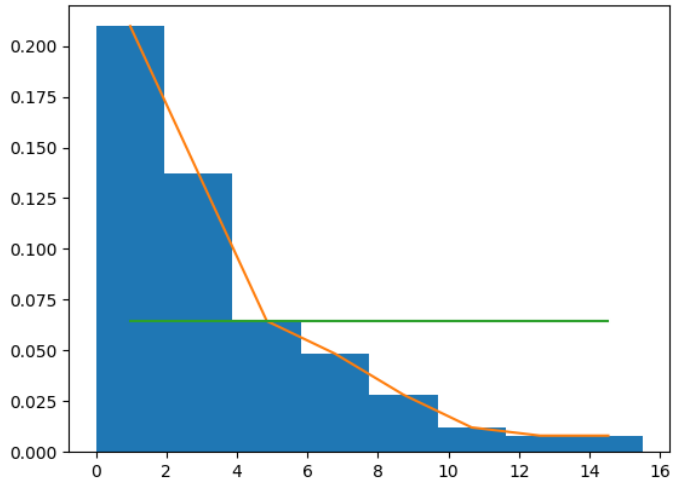

Рисунок 3.29 — График шестой выборки в спрямлённых для равномерного распределения координатах

Значение метрики R^2 для данного графика равно 0.80197.

Наибольшее значение коэффициента корреляции R^2 достигается при проверке на показательное распределение, и оно равно 0.97626. Это значение больше критического значения 0.85, из чего следует, что данное распределение является показательным по методу анаморфоз.

Вычислим параметры линейной регрессии, характеризующие параметры распределения: $a = -0.26120 \approx -\lambda$, $E[x] = \frac{1}{\lambda} = 3.8285$

Сравним критерий Пирсона и метод анаморфоз. Результаты проверки по методу анаморфоз и по критерию Пирсона совпадают для 1, 3, 4, 6 рядов, то есть для вариационных рядов и нормальным и показательным распределениями. Для рядов с равномерным распределением, критерий Пирсона подтвердил гипотетическое распределение, однако метод анаморфоз выдал противоположный результат, из этого можно сделать вывод, что метод анаморфоз более чувствителен к выбросам в выборке, чем критерий Пирсона.

Используя метод Пирсона, проверены для каждого файла одно ложное распределение и истинное. Из полученных значений сделан вывод о распределения рядов. Также для проверки распределения изучен метод анаморфоз. Каждый из рядов построен в спрямлённых координатах для нормального распределения, экспоненциального распределения и равномерном распределении. После спрямления рассчитаны коэффициенты детерминации и сделан вывод о распределение рядов.

4 ГЕНЕРАЦИЯ РАСПРЕДЕЛЕНИЙ

4.1 Постановка задачи

Сгенерировать выборку нормального распределения $Y \sim N$ (μ, σ^2) используя определение центральной предельной теоремы. На основе $n \approx 10 \div 20$ равномерно распределенных случайных реализаций случайных величин образовать новую выборку по определению центральной предельной теоремы.

Если $Yi \sim U$ (a_i, b_i), i = 1, 2,…, n, где Y_i— равномерно распределенная реализация случайной величины со случайными параметрами $a_i \in \mathbb{R}$, $b_i \in \mathbb{R}$, то ожидаемая нормально распределенная величина Y будет найдена по формуле (4.1):

$$Y = \sum_{i=1}^{n} Y_i, i = 1,2, \dots, n \tag{4.1}$$

Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность нормального распределения по несмещенным точечным оценкам $\hat{\mu}$, $\hat{\sigma}$ и провести тест на нормальное распределение с помощью критерия χ^2-Пирсона и методом анаморфоз. Степени свободы рассчитывать, как $k = n - 1$.

Для генерации выборок рекомендуется пользоваться встроенными в компьютерные статистические пакеты функциями генерации равномерно распределённых случайных величин, которые задаются с помощью параметров границ интервала генерации чисел a и b.

Сгенерировать выборку χ^2-распределения $R \sim \chi_k^2$ используя определение распределения χ^2. На основе Z-оценок нормально распределенных случайных реализаций случайных величин $L_i \sim N(\mu_i,\ \sigma_i^2)$ образовать новую выборку по определению χ^2-распределения по формуле (4.2):

$$R = \sum_{i=1}^{n} Z\,[L_i]^2,\ Z[L_i] = \frac{L - E[L]}{\sigma[L]},\ L_i \sim N(\mu_i, \sigma_i^2), i\ =\ 1, 2, \dots , n \tag{4.2}$$

Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность χ^2 распределения с $k = n - 1$ степенями свободы и провести тест на χ^2 с помощью критерия χ^2-Пирсона.

Для генерации нормально распределенных реализаций случайных величин рекомендуется пользоваться встроенными в статистические пакеты функциями для генерации значений выборки из нормального распределения, которые задаются с помощью параметров математического ожидания μ и стандартного отклонения σ^2.

Сгенерировать выборку распределения Фишера на основе определения. На основе двух случайных реализаций Y_1, Y_2 случайных величин, распределенных по χ^2-распределению со степенями свободы $d1$, $d2$ соответственно, сгенерировать выборку, распределенную по распределению Фишера $S \sim F\ (d1,\ d2)$ в соответствии с определением по формуле (4.3):

$$s = \frac{Y_1/d_1}{Y_2/d_2}, S \sim F(d1, d2) \tag{4.3}$$

Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность $F(d1,\ d2)$ распределения и провести тест на распределение Фишера с помощью критерия χ^2-Пирсона.

Для генерации выборки фиксированного размера из распределения χ^2 рекомендуется пользоваться встроенными в статистические пакеты функциями для генерации случайных выборок из распределения χ^2 с df степенями свободы.

Реализаций Y_1, Y_2, …, Y_n случайных величин, распределенных по стандартному нормальному распределению $Y_i \sim N$ (0, 1), i = 1, 2, …, n, сгенерировать выборку $T \sim t(n)$, распределенную по t-распределению Стьюдента с df = n степенями свободы в соответствии с определением по формуле (4.4):

$$T = \frac{Y_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}}, Y0 \sim N\ (0, 1) \qquad (4.4)$$

Для получившейся выборки построить гистограмму, визуализировать на гистограмме теоретическую плотность $t(n)$, полученную через аналитическую формулу (4.5) с использованием бета-функции:

$$P_t(x|n) = \frac{1}{\sqrt{n}B\left(\frac{1}{2},\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{\frac{-n+1}{2}}, \qquad (4.5)$$

где $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$,определённая при Re x> 0, Re y> 0.

Для получившейся выборки провести тест на t-распределение Стьюдента с помощью критерия χ^2-Пирсона, используя в качестве функции вероятности распределения по формуле (4.6):

$$P_t(x|n) = \int_{-\infty}^{x} p_t(z|n)\,dz. \qquad (4.6)$$

Выполним задачу в соответствии с заданием, в ходе выполнения воспользуемся собранным инструментарием.

4.2 Выполнение задачи

Сгенерируем выборку нормального распределения $Y{\sim}N(\mu,\ \sigma^2)$ используя определение центральной предельной теоремы (4.1). На основе n = 20 равномерно распределенных случайных реализаций случайных величин.

Для получившейся выборки построим гистограмму, визуализируем на гистограмме теоретическую плотность нормального распределения по несмещенным точечным оценкам $\hat{\mu}$, $\hat{\sigma}$ и проведём тест на нормальное распределение с помощью критерия χ^2-Пирсона и метода анаморфоз. Степени свободы рассчитаем, как $k = n - 1$.

Сгенерируем 20 равномерно распределённых случайных величин, сложим полученные случайные величины поэлементно. ЦПТ распределение полученной случайной величины является нормальным.

Построим гистограмму по полученному нормальному распределению, и плотность распределения с полученными параметрами (Рисунок 4.1).

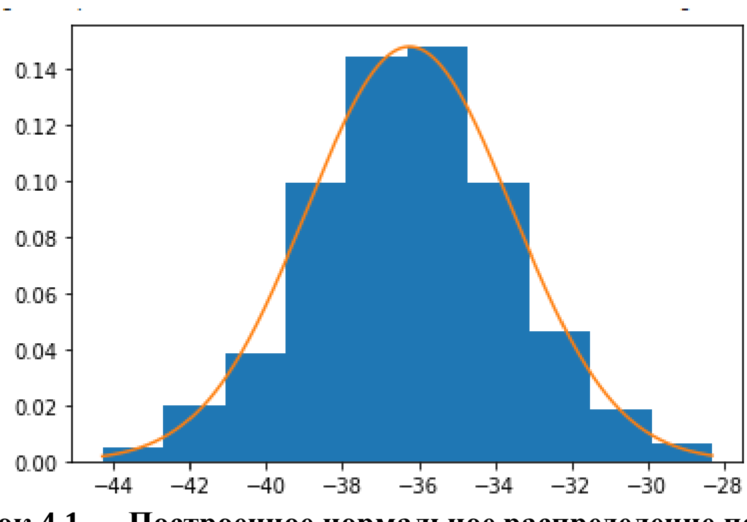

Рисунок 4.1 — Построенное нормальное распределение по ЦПТ

Проверим полученную случайную величину на вид распределения. Проверим с помощью метода анаморфоз (Рисунок 4.2).

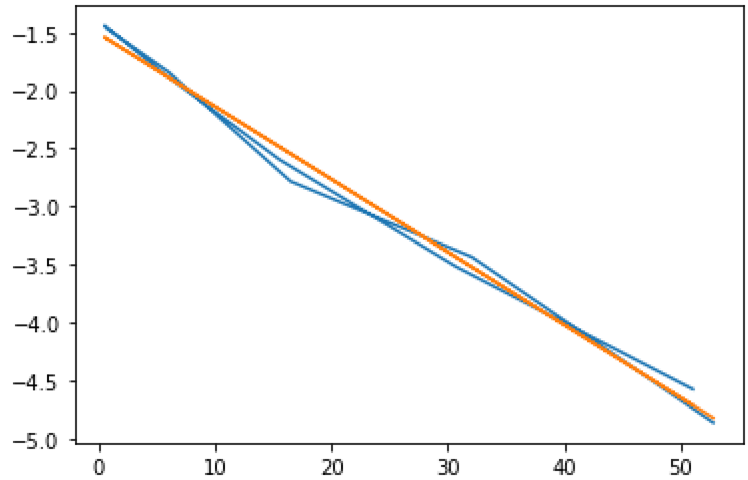

Рисунок 4.2 — График случайной величины в спрямленных для нормального распределения координатах

Полученный коэффициент корреляции равен 0.9910, что больше критического значения 0.85, что значит, что случайная величина распределена нормально.

Проверим также по критерию χ^2-Пирсона с n–1 степенями свободы. При вычислении было получено значение 5.219862, критическое значение равно 30.14353, можно с уверенность утверждать, что случайная величина распределена нормально.

Сгенерируем выборку χ^2-распределения $R \sim \chi_k^2$ используя определение распределения χ^2. На основе Z-оценок нормально распределенных случайных реализаций случайных величин $L_i \sim N(\mu_i,\ \sigma_i^2)$ образуем новую выборку по определению χ^2-распределения по формуле (4.2).

Для получившейся выборки построим гистограмму, визуализируем на гистограмме теоретическую плотность χ^2 распределения с $k = n - 1$ степенями свободы и проведем тест на χ^2 с помощью критерия χ^2-Пирсона.

Сгенерируем n=20 нормально распределённых величин. Для каждой посчитаем z-оценку. Поэлементно сложим возведённые во вторую степень z-оценки. Таким образом, получим случайную величину с χ^2 распределением по определению (Рисунок 4.3).

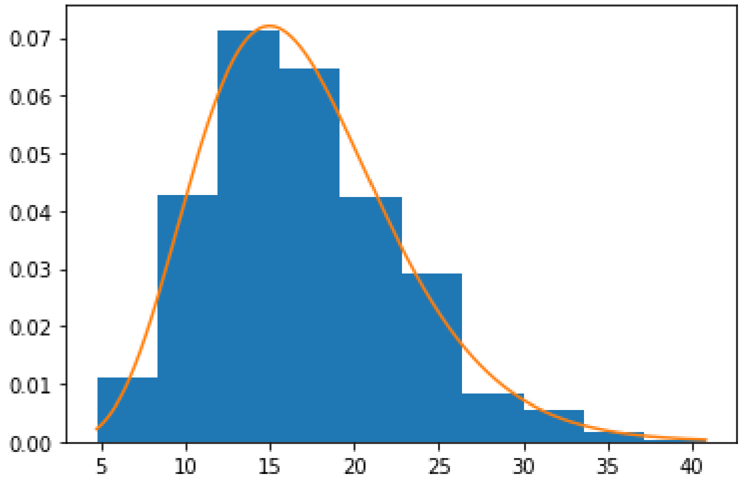

Рисунок 4.3 — Построенноераспределение χ^2по определению

Проверим по критерию χ^2-Пирсона с n–1 степенями свободы. При вычислении было получено значение 6.2683, критическое значение равно 17.5346, можно с уверенность утверждать, что случайная величина распределена по χ^2.

Сгенерируем выборку распределения Фишера на основе определения. На основе двух случайных реализаций Y_1, Y_2 случайных величин, распределенных по χ^2-распределению со степенями свободы $d1$, $d2$ соответственно, сгенерируем выборку, распределенную по распределению Фишера $S \sim F\ (d1,\ d2)$ в соответствии с определением по формуле (4.3).

Для получившейся выборки построим гистограмму, визуализируем на гистограмме теоретическую плотность $F\ (d1,\ d2)$ распределения и проведем тест на распределение Фишера с помощью критерия χ^2-Пирсона.

Сгенерируем две χ^2 распределённых величин, со случайными параметрами. Получим случайную величину с распределением Фишера по определению, при этом S ~ F(df1,df2) (Рисунок 4.4).

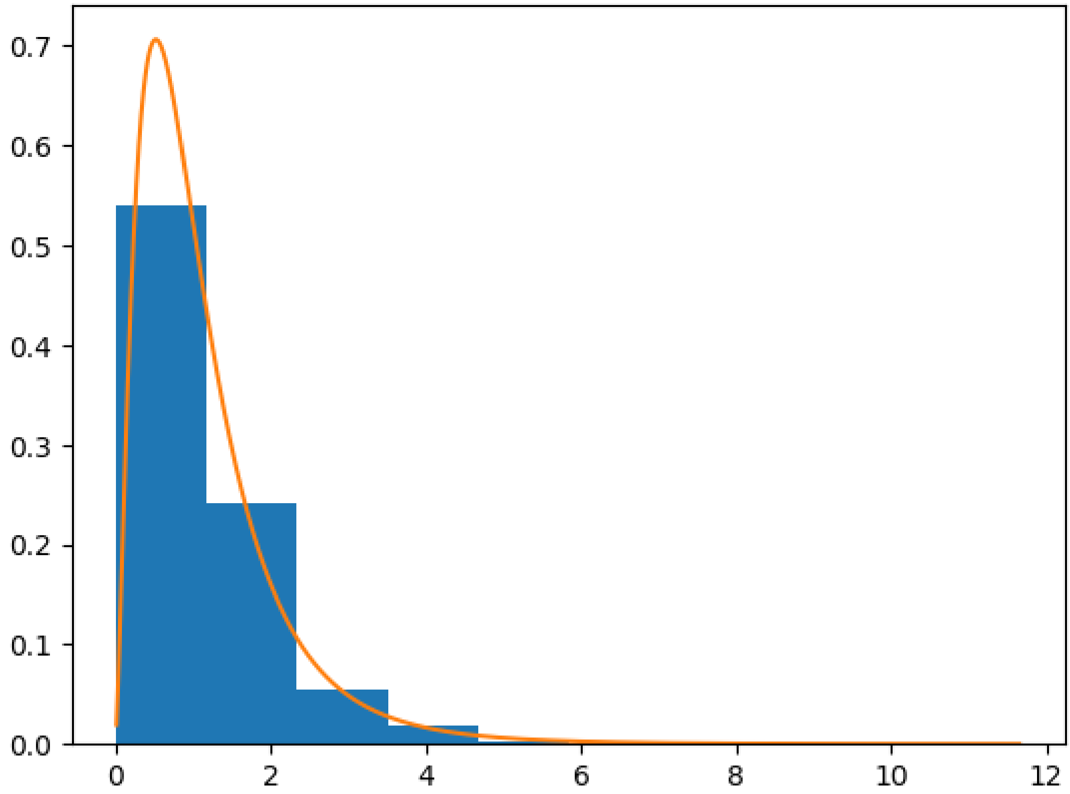

Рисунок 4.4 — Построенное распределение Фишера по определению

Проверим по критерию χ^2-Пирсона. При вычислении было получено значение 9.8481, критическое значение равно 16.0128, можно с уверенность утверждать, что случайная величина имеет распределение Фишера.

Сгенерируем выборку $T \sim t(n)$, распределенную по t-распределению Стьюдента с $df = n$ степенями свободы в соответствии с определением по формуле (4.4).

Для получившейся выборки построим гистограмму, визуализируем на гистограмме теоретическую плотность $t(n)$, полученную через аналитическую формулу с использованием бета-функции по формуле (4.5).

Для получившейся выборки проведем тест на t-распределение Стьюдента с помощью критерия χ^2-Пирсона.

Сгенерируем помощью n = 8 нормально распределённых величин со случайными параметрами. Воспользуемся формулой для нахождения случайной величины с t-распределением. Получим случайную величину сt-распределением по определению (Рисунок 4.5).

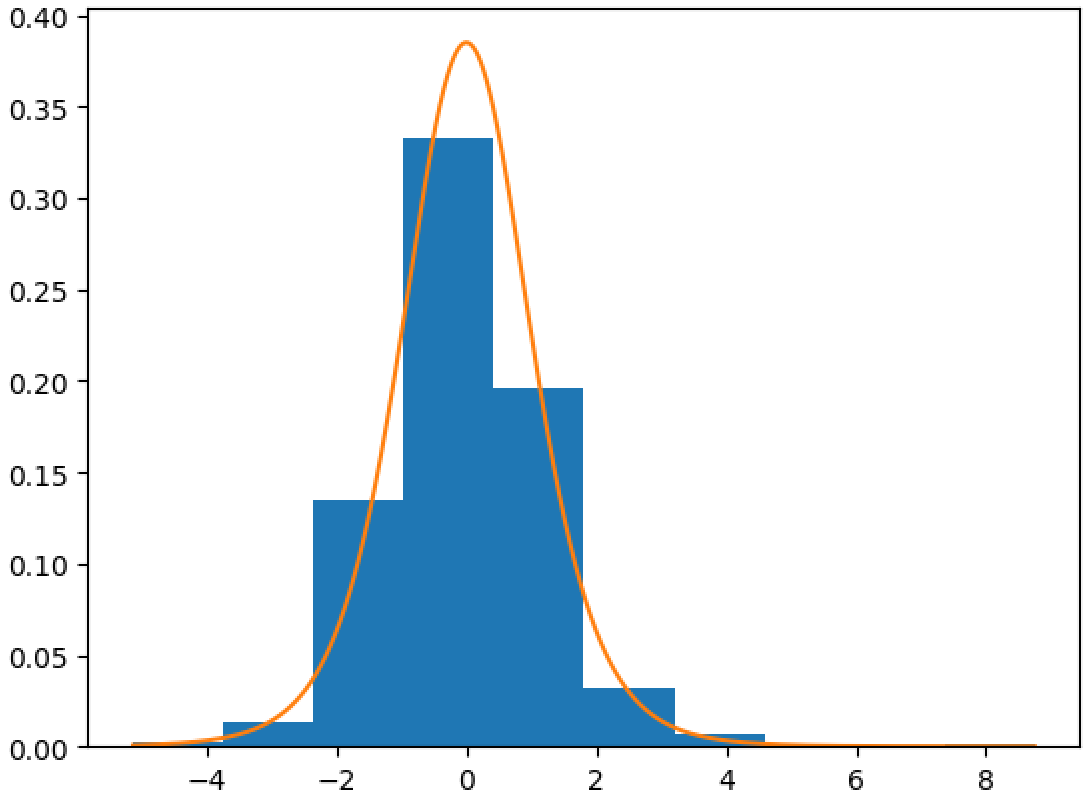

Рисунок 4.5 — Построенное t-распределение по определению

Проверим по критерию χ^2-Пирсона. При вычислении было получено значение 10.1846, критическое значение равно 17.5345, случайная величина имеет t-распределение.

Нами сгенерированы четыре выборки распределённых по следующим законам: нормальное распределения, χ^2-распределение, распределение Фишера, а также t-распределение Стьюдента. После генерации выборок нами проведены тесы на верность распределений.

5 ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ, ОЦЕНКА ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ ПАРАМЕТРОВ

5.1 Постановка задачи

Для выполнения практического задания необходимо:

1. Открыть папку, соответствующую своей группе.
2. Открыть папку с вариантом, совпадающим с вашим номером в списке.

В папке 3 файла с данными.

1. 1-ый файл содержит 2 ряда данных. Первый столбец x содержит факторную переменную, второй столбец y результирующую. Для первого файла необходимо:

- оценить коэффициент корреляции Пирсона $r(x,\ y)$ между двумя переменными в первом и втором столбце;
- по шкале Чеддока оценить характеристику корреляционной связи между величинами;
- проверить статистическую значимость коэффициента корреляции Пирсона с помощью t-статистики;
- построить доверительный интервал для $r(x,\ y)$ с надежностью γ = 0.95;
- построить линейную регрессию между столбцами, оценить значение коэффициентов линейной зависимости;
- оценить адекватность модели с использованием критерия Фишера.
- оценить значимость полученных коэффициентов прямой;

- построить доверительные интервалы для полученных коэффициентов;
- оценить интервал прогноза для линейной модели на 3 значения вперед.

2. 2-ой файл содержит 4 ряда данных. Первый ряд (столбец) содержит количественную факторную переменную, следующие два —качественную факторную переменную, последний — результирующую переменную. Для второго файла данных необходимо:

- необходимо с помощью теста Чоу обосновать необходимость деления выборки по одной из качественных факторных переменных;
- произвести разбиение и построить две линейных регрессии, оценить коэффициенты моделей.

3. 3-ий файл содержит 2 ряда данных. Для третьего файла данных необходимо:

- необходимо двумя способами (тест Спирмена и тест Гольдфельда-Квандта) определить, присутствует ли в данных гетероскедастичность;
- построить линейную регрессию, оценить значения коэффициентов модели;
- оценить значимость полученных коэффициентов и адекватность модели;
- все расчеты проводить для уровня значимости $\alpha = 0.05$.

Коэффициент корреляции Пирсона r(x, y) рассчитывается по формуле (5.1):

$$r(x,y) = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} \tag{5.1}$$

Статистическая значимость коэффициента корреляции Пирсона рассчитывается по формуле (5.2):

$$t_r = |r| \cdot \sqrt{\frac{n-2}{1-r^2}} \qquad (5.2)$$

Преобразования Фишера осуществляется по формуле (5.3):

$$z = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \qquad (5.3)$$

Ошибка среднего рассчитывается по формуле (5.4):

$$m_z = \frac{1}{\sqrt{N-3}} \qquad (5.4)$$

Доверительный интервал для z-оценке коэффициента корреляции Пирсона рассчитывается по формуле (5.5):

$$z - t(n-1)_{1-\frac{\alpha}{2}} \cdot m_z \leq z \leq z - t(n-1)_{1-\frac{\alpha}{2}} \cdot m_z \qquad (5.5)$$

Подсчёт коэффициентов парной линейной регрессии производится по формулам (5.6 — 5.7):

$$a = r(x,y) \cdot \frac{\sigma_y}{\sigma_x}, b = \bar{y} - a \cdot \bar{x} \qquad (5.6 — 5.7)$$

Сумма квадратов ошибок модели относительно значений результирующей переменной рассчитывается по формуле (5.8):

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}(x_i)\right)^2 \tag{5.8}$$

Полная сумма квадратов рассчитывается по формуле (5.9):

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5.9}$$

F-статистика для оценки адекватности модели рассчитывается по формуле (5.10):

$$F = \frac{RSS\ /\ (n-2)}{TSS\ /\ (n-1)} \tag{5.10}$$

Оценка статистической значимости коэффициента наклона прямой осуществляется по формуле (5.11):

$$T_a = \frac{a}{m_a} \tag{5.11}$$

Коэффициент m_a рассчитывается по формуле (5.12):

$$m_a = \frac{s_{AD}}{\sigma_x \cdot \sqrt{n}} \tag{5.12}$$

Среднеквадратичное отклонение данных относительно модели регрессии рассчитывается по формуле (5.13):

$$s_{AD} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}(x_i)\right)^2}{n-2}} \tag{5.13}$$

Оценка статистической значимости коэффициента смешения прямой осуществляется по формуле (5.14):

$$T_b = \frac{b}{m_b} \tag{5.14}$$

Коэффициент m_a рассчитывается по формуле (5.15):

$$m_b = \frac{s_{AD} \cdot \sqrt{\sum_{i=1}^{n} x_i^2}}{n \cdot \sigma_x} \tag{5.15}$$

Доверительный интервал для коэффициентов прямой считается по формуле (5.16 —5.17):

$$\begin{gathered} \hat{a} \in (a - m_a \cdot t, a + m_a \cdot t) \\ \hat{b} \in (b - m_b \cdot t, b + m_b \cdot t) \end{gathered} \tag{5.16 —5.17}$$

Тест Чоу для определения необходимости делить выборку по одной из качественных факторных переменных провидится по формуле (5.18):

$$F = \frac{(RSS - RSS_1 - RSS_2) \ / \ k}{(RSS_1 + RSS_2) \ / \ (n - 2k)} \tag{5.18}$$

Остатки считаются по формуле (5.19):

$$e_i = y_i - \hat{y}(x_i) \tag{5.19}$$

Коэффициент ранговой корреляции Спирмена считается по формуле (5.20):

$$r_s(e, x) = 1 - \frac{6 \sum_{i=1}^{n} (rank(e_i) - rank(x_i))}{n \cdot (n^2 - 1)} \tag{5.20}$$

Выполним задачу в соответствии с заданием, в ходе выполнения воспользуемся собранным инструментарием.

5.2 Выполнение задачи

Решим задачу для первого файла данных, содержащего 2 ряда данных. Первый столбец содержит факторную переменную x, второй столбец результирующую переменную y. Начертим график зависимости y от x (Рисунок 5.1).

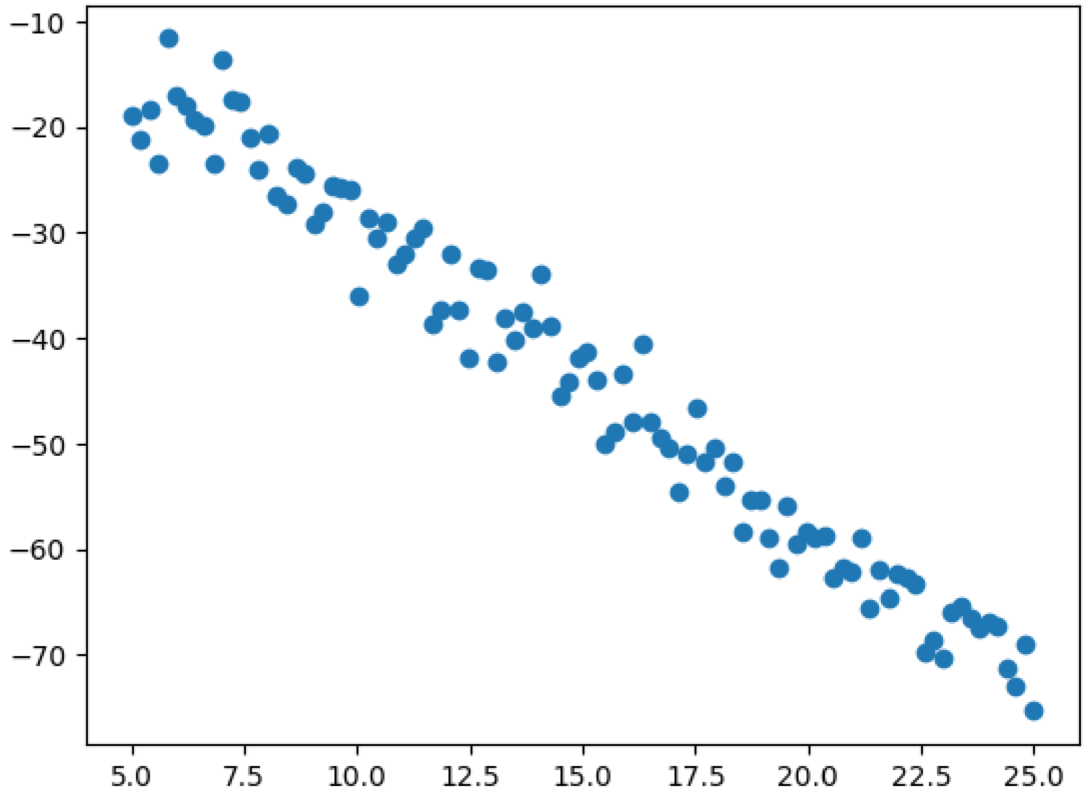

Рисунок 5.1 — Зависимость данных первого файла

Введём 2 гепотизы о связи переменных x и y между собой:

- H1: значение коэффициента линейной корреляции Пирсона значительно отличается от нуля;
- H0: значение коэффициента линейной корреляции Пирсона незначительно отличается от нуля.

Оценим коэффициент корреляции Пирсона $r(x,y)$ между двумя переменными x и y соответственно по формуле(5.1). Получим $r(x, y) = 0.9845$

По шкале Чеддока (Таблица 5.1) оценим характеристику корреляционной связи между величинами.

Таблица 5.1—Шкала Чеддока.

Количественная мера тесноты связи	Качественная характеристика силы связи
0.1 — 0.3	слабая
0.3 — 0.5	умеренная
0.5 — 0.7	заметная
0.7 — 0.9	высокая
0.9 — 1	весьма высокая

Значение r(x,y) = 0.9845 сигнализирует о весьма высокой силе связи между переменными.

Проверим статистическую значимость коэффициента корреляции Пирсона с помощью t-статистики по формуле (5.2). Получим $\mathrm{t_r} = 55.7865.$

При этом критическое значение t = 1.98446, так как посчитанное значение значительно превосходит критическое, принимается гипотеза H1.

Построим доверительный интервал для $r(x,y)$ с надежностью γ = 0.95.

Преобразуем коэффициент корреляции в z-оценку с помощью преобразования Фишера по формуле (5.3). Получим z= 2.4299, а $m_z = 0.1015$ по формуле (5.4).

Преобразуем z в r обратно и получим доверительный интервал для r(x,y).

$$r(x, y) \in [0.9897, 0.97707]$$

Построим линейную регрессию связи между переменными (Рисунок 5.2)

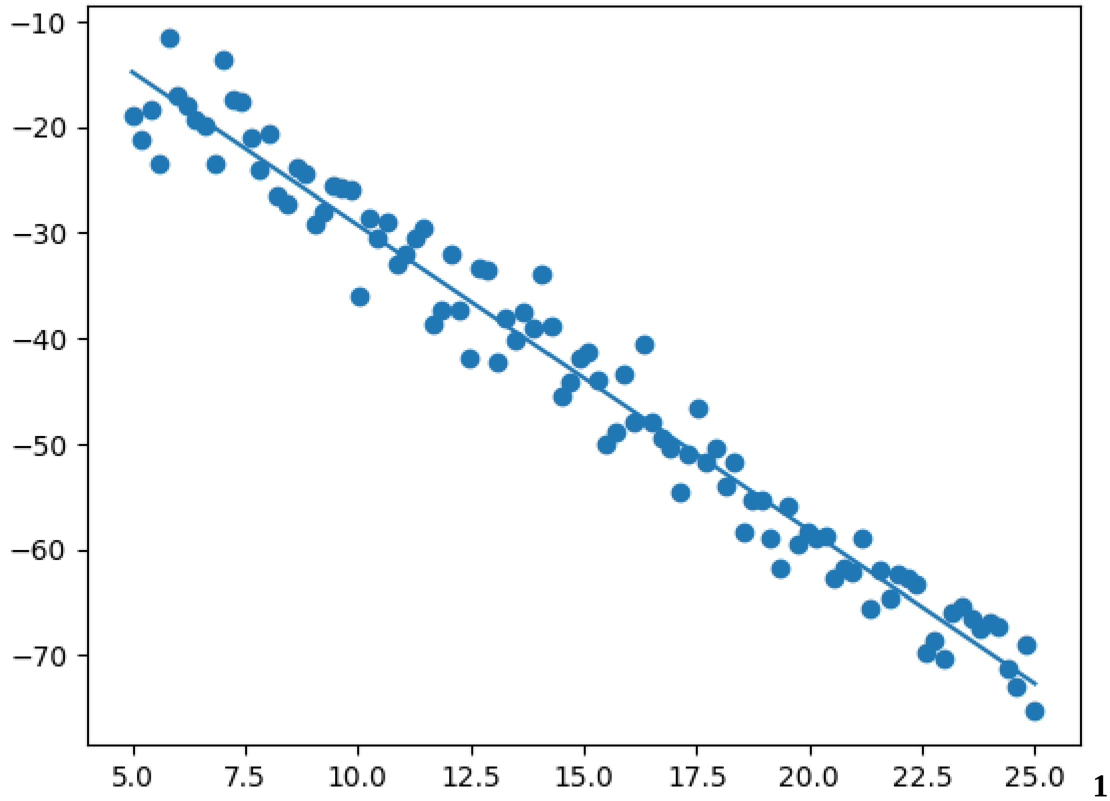

Рисунок 5.2 — Построенная линейная регрессия

Оценим коэффициенты линейной регрессии. По формуле (5.6 — 5.7) получим $a = 1.4875,\ b = 0.03084$

Оценим адекватность модели с использованием критерия Фишера по формуле (5.10):

- $RSS = 896.6910$ по формуле (5.8);
- $TSS = 29372.43$ по формуле (5.9);
- $F = 0.02192844$ по формуле (5.10).

Так же рассчитаем критическое значение $f(n-2, n-1) = 1.4675$. Так как рассчитанная статистика значительно меньше критического значения, модель признается адекватной.

Оценим значимость полученных коэффициентов прямой по формулам (5.11 — 5.15).

- $m_a = 0.08518$;
- $m_b = 0.83479$;

- $s_{AD} = 3.02488$;
- $T_a = 55.7860$
- $T_b = 0.41629$;
- $t = t(n-2)_{1-\frac{\alpha}{2}} = 1.98446$.

Так как T_a больше критического значения, a — статистически значимый коэффициент. T_b меньше критического значения, следовательно b— статистически не значимый коэффициент.

Построим доверительные интервалы по формуле (5.16 — 5.17) для полученных коэффициентов.

$$\hat{a} \in (2.9966, 2.7908)$$

$$\hat{b} \in (-2.0041, 1.3091)$$

Оценим интервал прогноза для линейной модели на 3значения вперед.

Спрогнозируем значения, подставив значения x в построенную модель:

- $y_1(25.20202) = -73.2749$;
- $y_2(25.40404) = -73.8595$;
- $y_3(25.60606) = -74.4440$.

Оценим интервал прогноза:

- $y_1 \in (\hat{y}(x) - 6.1234, \hat{y}(x) + 6.1234)$ = (58.548, 61.6019);
- $y_2 \in (\hat{y}(x) - 6.1279, \hat{y}(x) + 6.1270)$ = (59.01994, 62.08296);
- $y_3 \in (\hat{y}(x) - 6.1307, \hat{y}(x) + 6.1307)$ = (59.49181, 62.5641).

Заметим, что чем дальше прогноз тем больше становится интервал.

Проведем исследование второго набора данных. Он содержит 4 ряда данных. Первый ряд содержит количественную факторную переменную, следующие два — качественную факторную переменную, последний — результирующую переменную.

Построим графики зависимость y от x, при этом в зависимости от категориальной переменной окрасим значения. Построим график, где окраска наблюдений зависит от первой категориальной переменной(Рисунок 5.3).

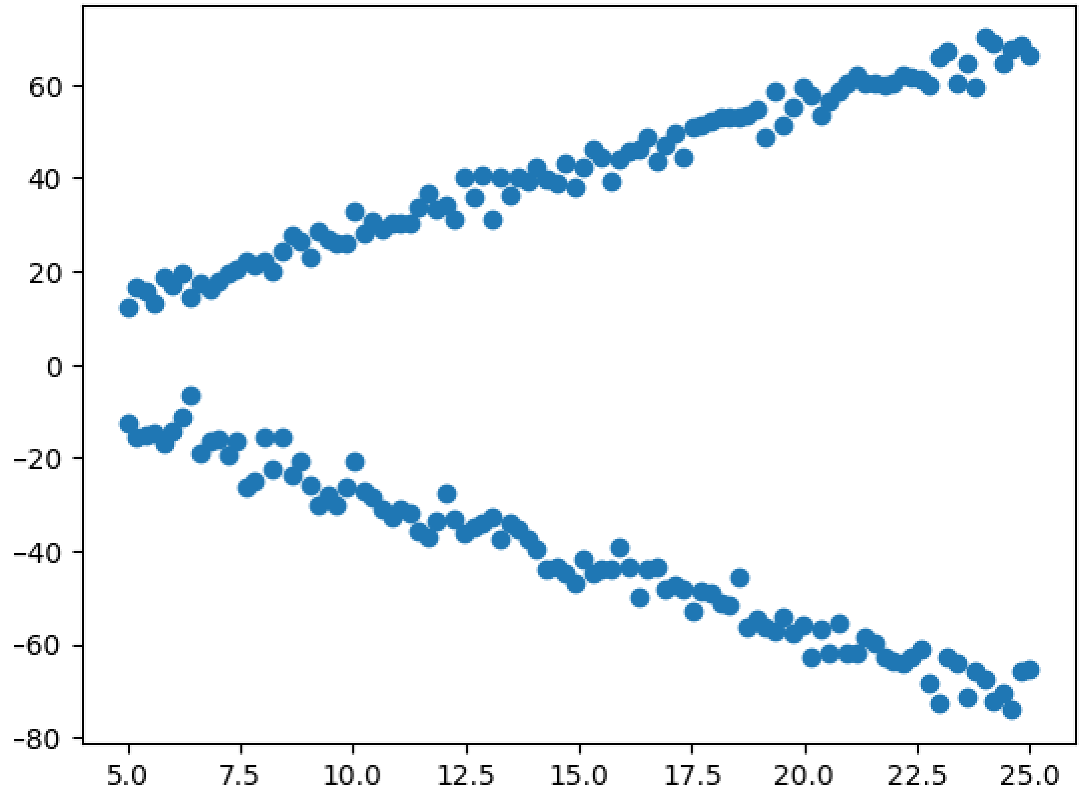

Рисунок 5.3 — Визуализация данных с учетом первой категориальной переменой

Построим график, где окраска наблюдений зависит от второй категориальной переменной (Рисунок 5.4).

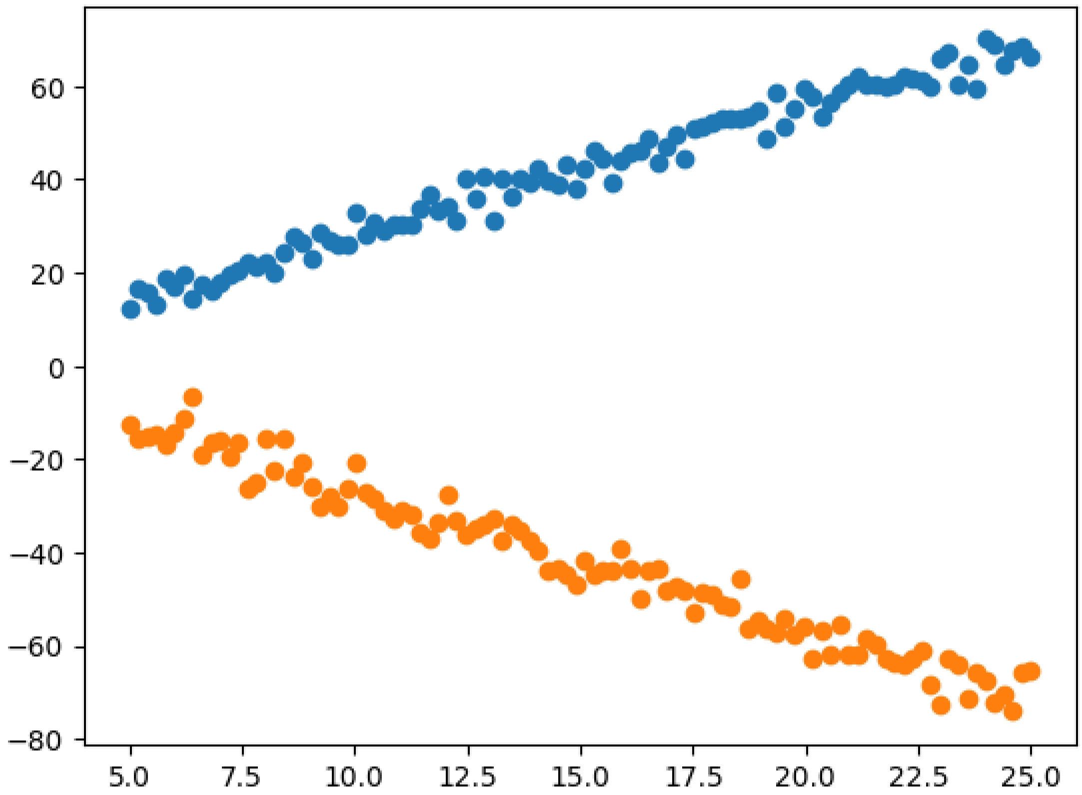

Рисунок 5.4 — Визуализация данных с учетом второй категориальной переменой

Необходимо с помощью теста Чоу обосновать необходимость деления выборки по одной из качественных факторных переменных. Для этого рассчитаем коэффициент F по формуле (5.18):

Рассчитаем элементы для подсчёта F-статистики для первой категориальной переменной:

- RSS = 409800.130 — RSS для общей модели;
- RSS_1 = 1036.8899 — RSS для модели где категориальная переменная равна 0;
- RSS_2 = 598.5439 — RSS для модели где категориальная переменная равна 1;
- k = 2 — количество элементов построенной модели;
- n = 200 — количество строк в оцениваемой таблице.

Тогда F = 48856.67, при этом критическое значение f(k, n–2k) = 3.7591. Рассчитанная F-статистика значительно больше критического значения, поэтому принимается гипотеза о том, что следует строить отдельные модели в соответствии с данной категориальной переменной.

Рассчитаем элементы для подсчёта F-статистики для второй категориальной переменной:

- RSS = 409800.130 — RSS для общей модели;
- RSS_1 = 205395.498 — RSS для модели где категориальная переменная равна 0;
- RSS_2 = 201865.8795 — RSS для модели где категориальная переменная равна 1;
- k = 2 — количество элементов построенной модели;
- n = 200 — количество строк в оцениваемой таблице.

Тогда F = 0.6246 при этом критическое значение $f(k, n-2k) = 3.7591$. Рассчитанная F-статистика значительно меньше критического значения, поэтому принимается гипотеза о том, что следует строить общую модель.

Произведем разбиение и построим две линейных регрессии, оценив коэффициенты моделей (Рисунок 5.5).

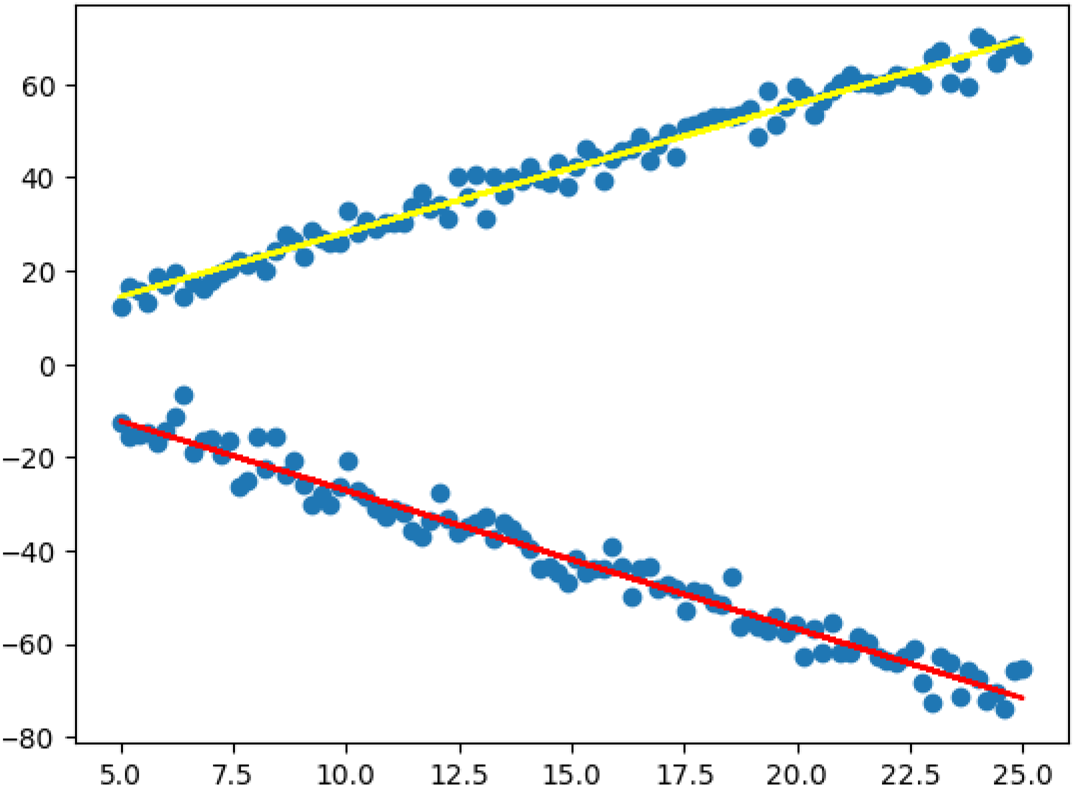

Рисунок 5.5 — Визуализация построенных моделей

Построим первую модель, где категориальная переменная равна нулю:$a = -2.97165,\ b = 2.618$

Построим вторую модель, где категориальная переменная равна единице: $a = 2.756, b = 0.65317$

Исследуем 3 набор данных, содержащий 2 ряда данных. Необходимо двумя способами (тест Спирмена и тест Гольдфельда-Квандта) определить, присутствует ли в данных гетероскедастичность.

Воспользуемся тестом Спирмена построим линейную регрессию и посчитаем остатки по формуле (5.19).

Отранжируем рассчитанные остатки, атакже отранжируем переменную, относительно которой ожидается гетероскедастичность. Произведем вычисления коэффициента ранговой корреляции Спирмена по формуле (5.20). Получим:

- $r_s(\text{e}, x) = 1$;

- $\left|r_S(l,x)\cdot\sqrt{n-1}\right| = 9.94987;$
- $N(0,1)_{1-\frac{\alpha}{2}} = 1.959964.$

Рассчитанная статистика больше чес критическое значение, поэтому принимаем гипотезу о наличии гетероскедастичности.

Воспользуемся тестом Гольдфельда-Квандта.

Отсортируем выборку по первой переменной и возьмем по $m1 = m2 = \lfloor 0.4 * n \rfloor$ записей с каждой стороны. Тогда можем рассчитать F-статистику: для n = 100, $m1$ = $m2$ = 40, k = 1.

Проведя расчёты при помощи вычислительных средств, получим F =5.0616, при критическом значении $F\ (39, 39)_{1-0.05/2} \approx 1.890719$.

Расчётная статистика намного больше критического значения распределения Фишера, что говорит о принятии гипотезы о наличии гетероскедастичности в данной зависимости.

Необходимо построить линейную регрессию, оценить значения коэффициентов модели (Рисунок 5.6).

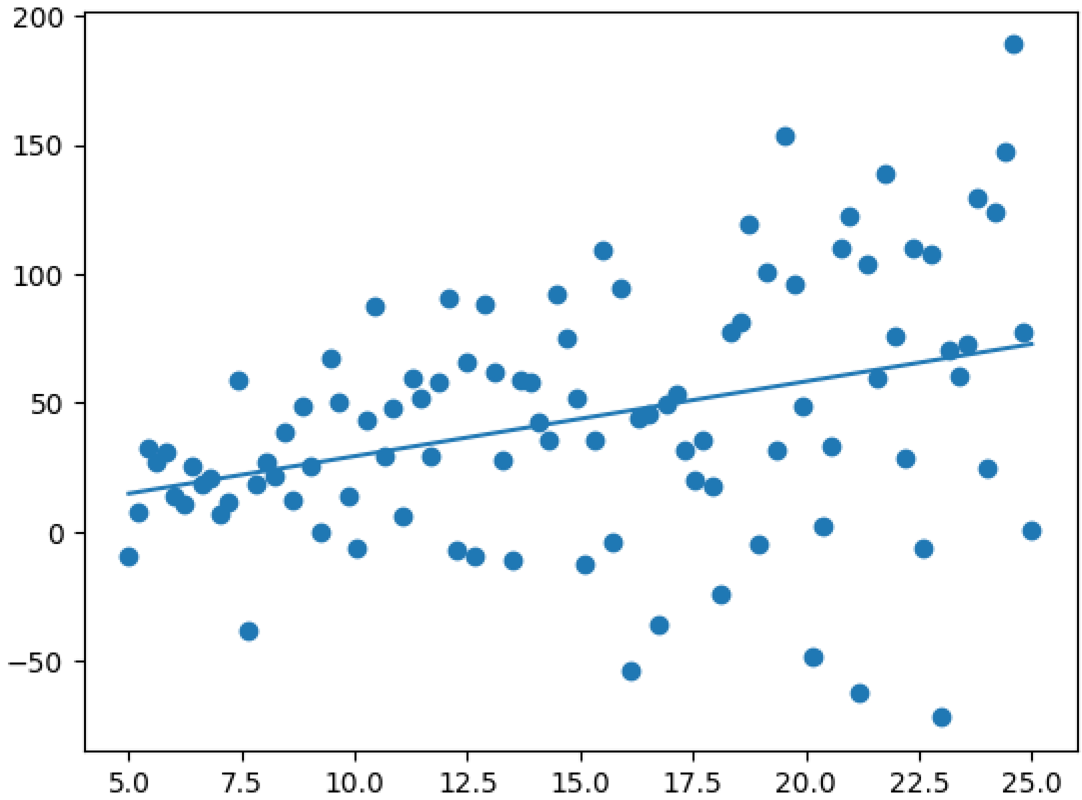

Рисунок 5.6 — Визуализация построенной модели

Оценим кэфициенты линейной регрессии $a = 2.901, b = 0.245$.

Оценим значимость полученных коэффициентов и адекватность модели.

Оценим адекватность модели с использованием критерия Фишера по формуле (5.10). Получим:

- $RSS = 204430$;
- $TSS = 233047.5$;
- $F = 0.886154$.

Так же рассчитаем критическое значение f(n–2, n–1) = 1.487451. Так как рассчитанная статистика меньше критического модель признается адекватной.

Оценим значимость полученных коэффициентов прямой по формуле (5.11 — 5.15). Получим:

- $m_a = 1.886587$;
- $m_b = 30.36214$;
- $s_{AD} = 45.67297$;
- $T_a = 1.537648$;
- $T_b = 0.008072322$;
- $t = t(n-2)_{1-\frac{\alpha}{2}} = 1.984467$;

Так как T_a меньше критического значения, a — статистически не значимый коэффициент. T_b меньше критического значения, b — статистически не значимый коэффициент.

Мы оценили коэффициент корреляции между двумя рядами данных, после подсчета оценили характеристику корреляционной связи между величинами, проверили статистическую значимость найденного коэффициента корреляции, а также построили доверительный интервал для коэффициента Пирсона. Мы изучили аппарат для определения необходимости деления выборки по одной из качественной переменной. Воспользовавшись тестом Чоу, произвели разбиение выборки по одной из качественной переменной. Оценили коэффициенты построенной модели. Изучили способы определения наличия гетероскедастичности в данных.

6 СГЛАЖИВАНИЕ ВРЕМЕННЫХ РЯДОВ

6.1 Постановка задачи

Для выполнения практического задания необходимо:

- открыть папку, соответствующую своей группе;
- открыть папку с вариантом, совпадающим с вашим номером в списке.

В папке два файла, которые содержат разные временные ряды. В первом файле находится ряд с синусоидальным трендом. Во втором —с линейным.

Необходимо выделить тренд, используя 4 метода:

1. простое скользящее среднее (SMA);
2. взвешенное скользящее среднее (WMA) особо типа;
3. экспоненциальное сглаживание (EMA);
4. двойное экспоненциальное сглаживание (DEMA).

Каждый метод требует подбора некоторых параметров:

1. SMA и WMA —размер окна, EMA —параметр сглаживания A, DEMA —параметр сглаживания вокруг тренда A и параметр сглаживания самого тренда B.
2. Необходимо подобрать оптимальные значения соответствующих параметров, используя Q-статистику Льюнг-Бокса при m = 5. Оптимальными параметрами будем считать те, что минимизируют приведенную статистику.
3. В качестве размеров окна $w = 2 *m + 1$ перебрать значения $m = 3$, 5, 7, 9; в качестве параметров сглаживания: $\alpha, \gamma = 0.1, 0.2, \dots, 0.9$. Обратите внимание, что метод DEMA двухпараметрический, что

требует выбрать оптимальную комбинацию сразу двух параметров α, γ.

4. После подбора оптимальных параметров провести тест Дарбина-Уотсона ($m = 1$, $\alpha = 0.95$) на данных после исключения выделенного тренда для каждого метода и каждого ряда.
5. В отчете изобразить графики исходных данных, графики трендов при оптимальных параметрах у каждого метода для каждого ряда, расчетные формулы, а также результаты тестов Дарбина-Уотсона.

Размер окна определяется по формуле (6.1):

$$\omega = 2 \cdot m + 1 \tag{6.1}$$

Веса для SMA рассчитываются по формуле (6.2):

$$w_i = \frac{1}{2 \cdot m + 1} \tag{6.2}$$

Экспоненциальная весовая функция (6.3) используется для определения весов в WMA:

$$w_i = \frac{e^{-\varepsilon \cdot |-i|}}{\sum_{j=-m}^{m} e^{-\varepsilon \cdot |j|}} \tag{6.3}$$

Модель скользящего среднего определяется по формуле (6.4):

$$\tilde{y}_t = \sum_{i=-m}^{m} w_i \cdot y_{t+i} \tag{6.4}$$

Модель экспоненциального сглаживания определяется рекурсивной формулой (6.5):

$$\tilde{y}_t = \begin{cases} y_t, t = 1 \\ \tilde{y}_{t-1} + \alpha \cdot (y_t - \tilde{y}_{t-1}), t > 1 \end{cases} \tag{6.5}$$

Модель двойного экспоненциального сглаживания определяется рекурсивной формулой (6.6):

$$\begin{gathered} \tilde{y}_t = \alpha \cdot y_t + (1 - \alpha) \cdot (\tilde{y}_{t-1} + b_{t-1}) \\ b_t = \gamma \cdot (\tilde{y}_t - \tilde{y}_{t-1}) + (1 + \gamma) \cdot b_{t-1} \\ \tilde{y}_1 = y_1, b_1 = y_2 - y_1 \end{gathered} \tag{6.6}$$

Для поверки отсутствия автокорреляции остатков используется Q-статистика Льюнг-Бокса, которая вычисляется по формуле (6.7):

$$Q = n \cdot (n - 2) \cdot \sum_{k=1}^{m} \frac{r(k)^2}{n - k} \tag{6.7}$$

Критерий Дарбина-Уотсона рассчитывается по формуле (6.8):

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{6.8}$$

Выполним задачу в соответствии с заданием, в ходе выполнения воспользуемся собранным инструментарием.

6.2 Выполнение задачи

Начнём решение задачи с временного ряда из первого файла. Построим временной ряд с не линейным трендом (Рисунок 6.1).

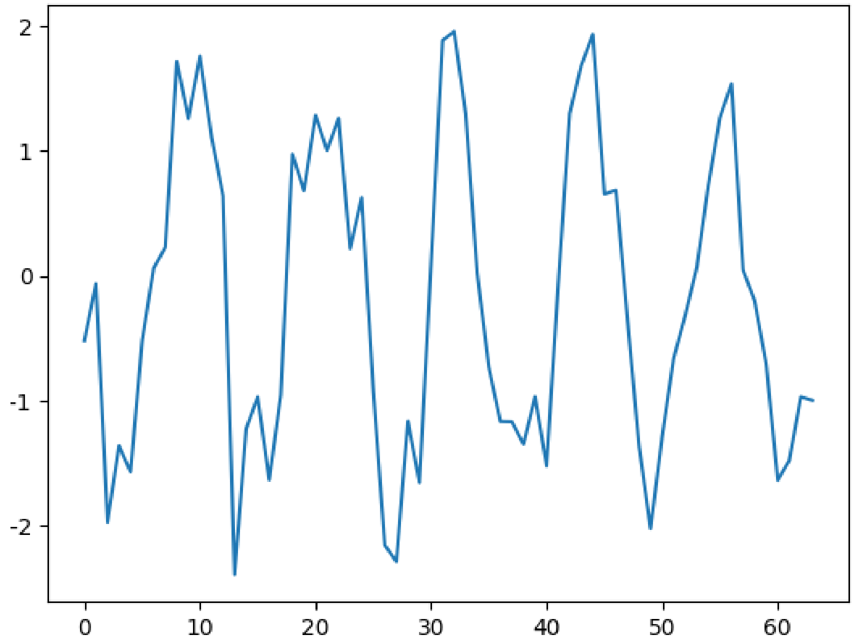

Рисунок 6.1 — График первого временного ряда

Воспользуемся методом сглаживания SMA по формуле (6.4). Для этого необходимо выбрать параметр m— количество членов ряда в сумме по одной стороне от центрального значения. Возможные значения параметра m = 2, 3, 5, 7, 9, из которых возьмем m = 2. Подробности выбора параметра будут описаны ниже. Определим размер окна по формуле (6.1), получим ω = 5. Также обходимо рассчитать веса по формуле (6.2), SMA имеет одинаковые веса для всех значений в окне. Получим $w_i = 1/5$.

Построим сглаживание по полученным параметрам (Рисунок 6.2).

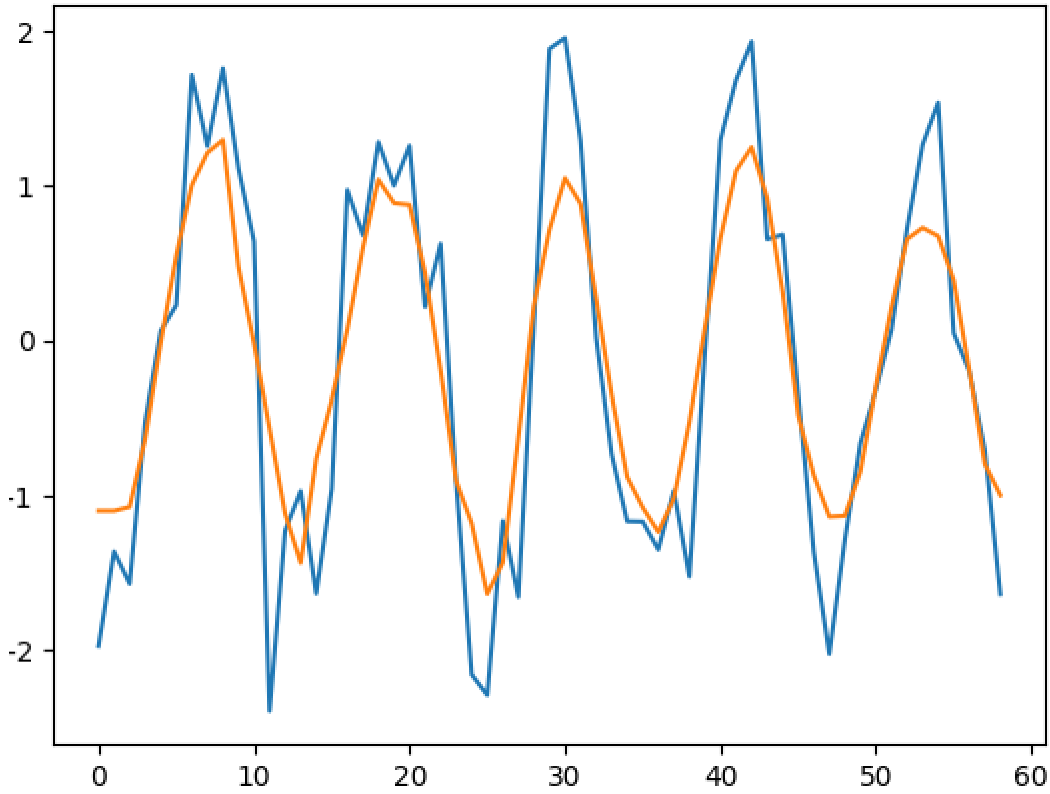

Рисунок 6.2 — Сглаживание первого временного ряда методом SMA

Вычислим Q-статистику Льюнг-Бокса при m = 5 по формуле (6.7) для проверки автокорреции остатков и выяснения, являются ли остатки случайными. Получим значение Q-статистики, равное 13.368. Q-статистика распределена по χ^2 при критическом значении, равном 12.832. Таким образом, можно сделать вывод, что остатки не являются случайной величиной.

Параметр m = 2 был выбран путём перебора и был выбран таким, чтобы Q-статистика при данном значении параметра была минимальна.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Для того чтобы определить автокорреляцию первого порядка для случайной составляющей временного ряда. Получаемое в результате теста значение равно 1.720 при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Реализуем метод сглаживания WMA по формуле (6.4) для этого необходимо выбрать параметр m — количество членов ряда в сумме по одной стороне от центрального значения. По аналогичным предыдущему решению причинам, возьмём m = 2. Определим размер окна по формуле (6.1), получим ω = 5. Также необходимо рассчитать веса по формуле (6.3). В отличие от WMA имеет разные веса для всех значений в окне. Веса будут рассчитаны с помощью вычислительных средств.

Построим сглаживание по полученным параметрам (Рисунок 6.3).

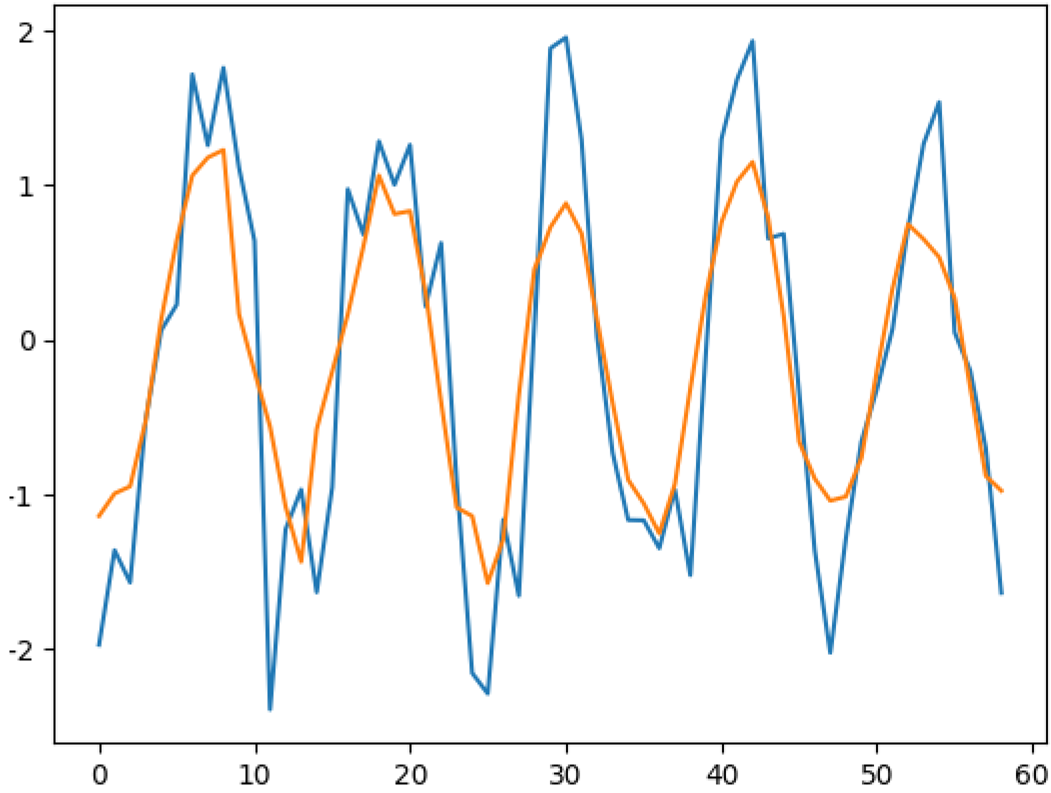

Рисунок 6.3 — Сглаживание первого временного ряда методом WMA

Вычислим Q-статистику Льюнг-Бокса по формуле (6.7) при m = 5. Получим значение Q-статистики, равное 16.045. Критическое значение Q-статистики равно 12.832, следовательно, остатки не являются случайной величиной.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Получаемое в результате теста значение равно 1.571, при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Реализуем метод сглаживания EMA по формуле (6.5), для этого необходимо выбрать параметр сглаживания a. По описанным в дальнейшем причинам возьмём a = 0.9.

Построим сглаживание по полученным параметрам (Рисунок 6.4).

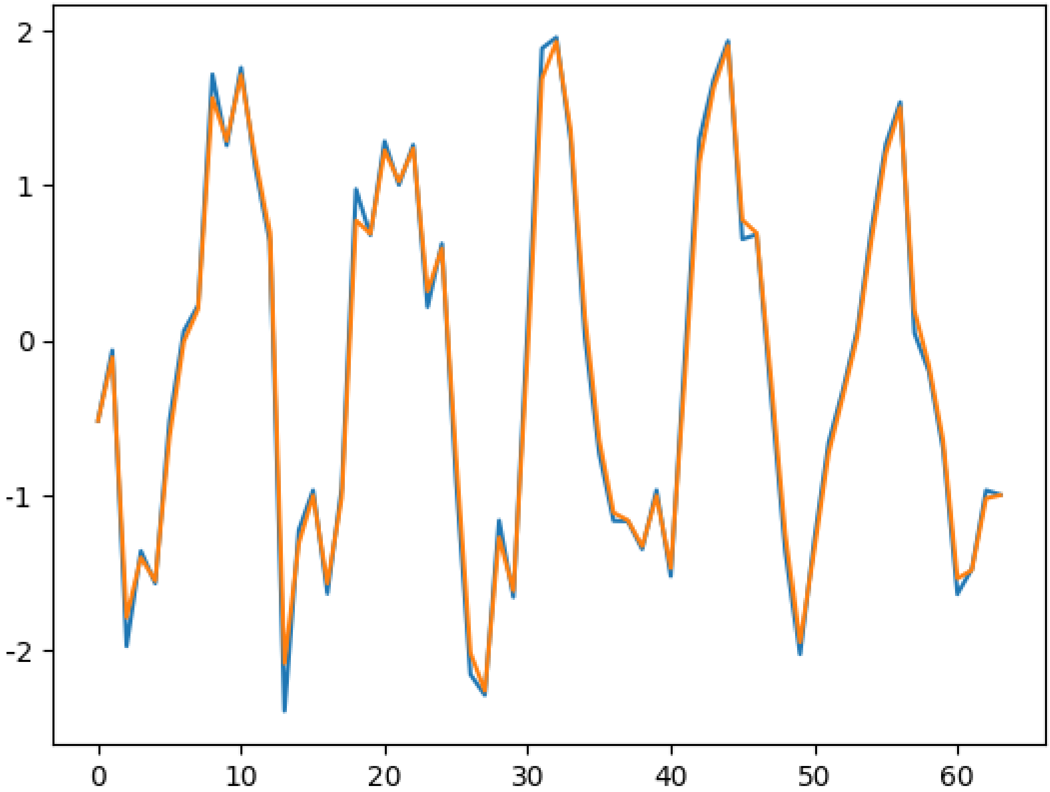

Рисунок 6.4 — Сглаживание первого временного ряда методом EMA

Вычислим Q-статистику Льюнг-Бокса по формуле (6.7) при m = 5. Получим значение Q-статистики, равное 22.851. Критическое значение Q-статистики равно 12.832, следовательно, остатки не являются случайной величиной.

Параметр a = 0.9 был выбран путём перебора и был выбран таким, чтобы Q-статистика при данном значении параметра была минимальна. Параметр выбирался из значений a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Получаемое в результате теста значение равно 1.6222, при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Реализуем метод сглаживания DEMA. Для этого необходимо выбрать параметр сглаживания вокруг тренда a и параметр сглаживания самого тренда b. По описанным в дальнейшем причинам возьмем a = 0.8, b = 0.9. DEMA — двойное экспоненциальное сглаживание и позволяет рекурсивно оценить новое значение сглаженного ряда по исходным данным по формуле (6.6).

Построим сглаживание по полученным параметрам (Рисунок 6.5).

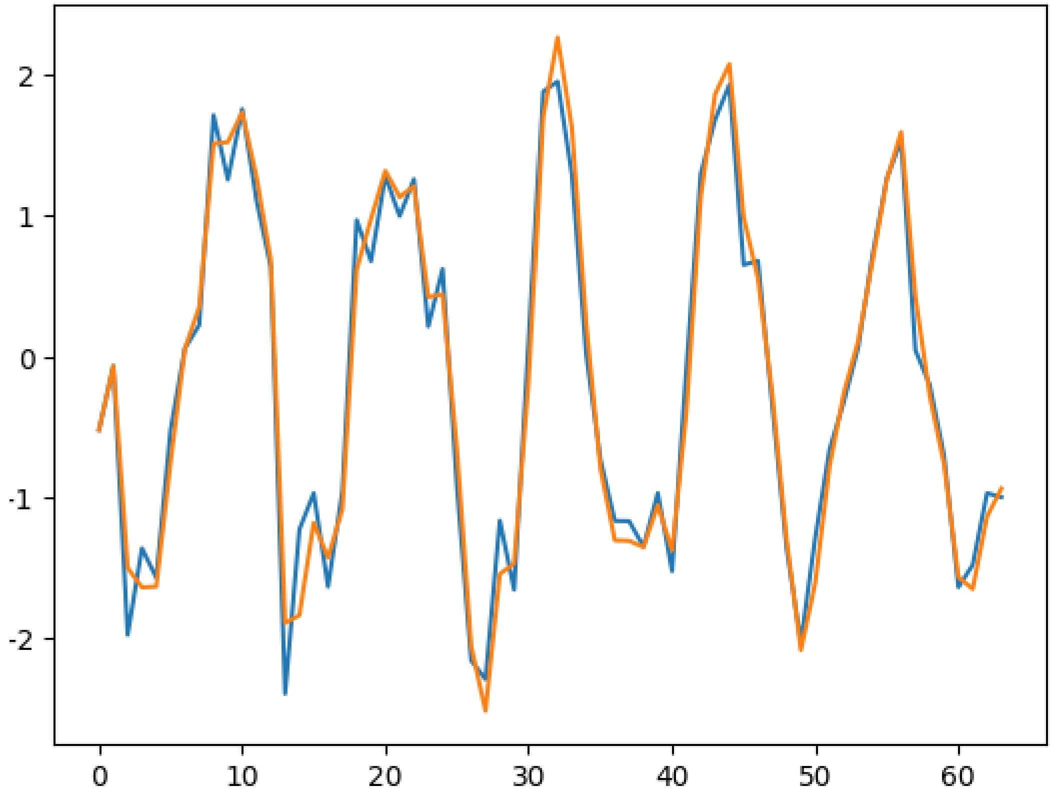

Рисунок 6.5 — Сглаживание первого временного ряда методом DEMA

Вычислим Q-статистику Льюнг-Бокса по формуле (6.7) при m = 5. Получим значение Q-статистики, равное 8.9023. Критическое значение Q-статистики равно 12.832, следовательно, остатки являются случайной величиной.

Параметры a = 0.8 и b = 0.9 были выбраны путём перебора и были выбраны такими, чтобы Q-статистика при данных значениях параметров минимальна. Параметры выбирались из значений a, b = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Получаемое в результате теста значение равно 2.2406, при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Теперь выполним задачу для временного ряда из второго файла. Построим временной ряд с не линейным трендом (Рисунок 6.6).

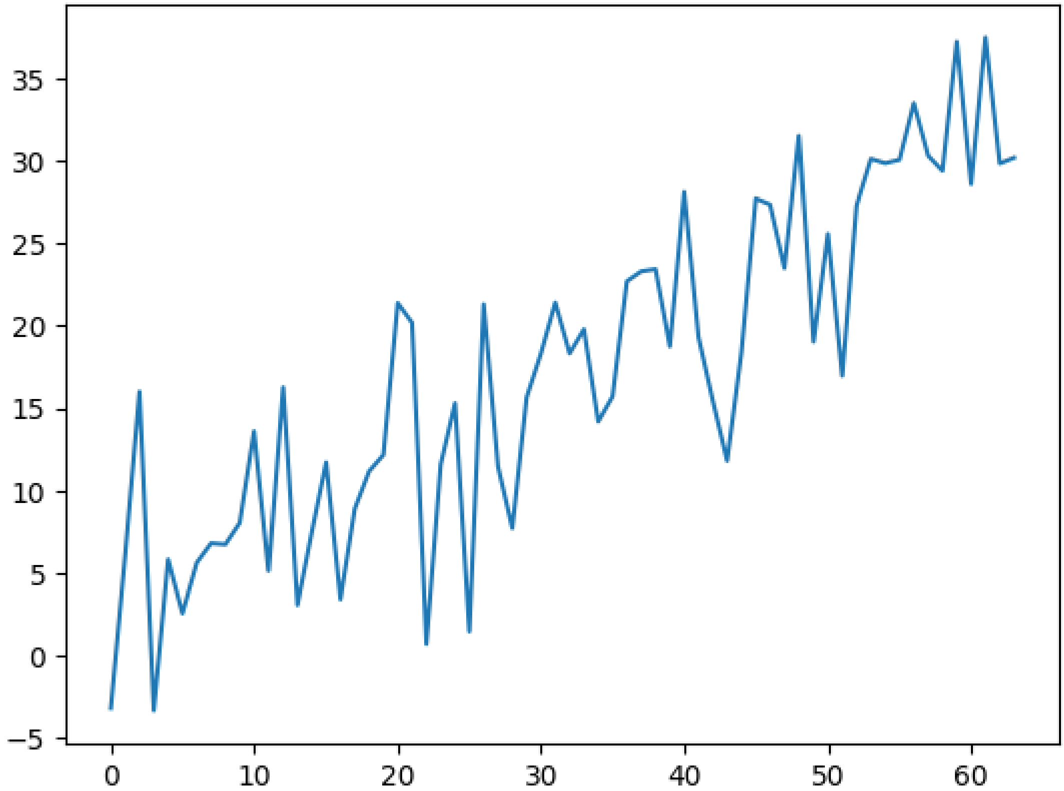

Рисунок 6.6 — График второго временного ряда

Воспользуемся методом сглаживания SMA по формуле (6.4). Для этого необходимо выбрать параметр m— количество членов ряда в сумме по одной стороне от центрального значения. Возможные значения параметра m = 2, 3, 5, 7, 9, из которых возьмем m = 3. Подробности выбора параметра будут описаны ниже. Определим размер окна по формуле (6.1), получим ω = 7. Также обходимо рассчитать веса по формуле (6.2), SMA имеет одинаковые веса для всех значений в окне. Получим w_i = 1/7.

Построим сглаживание по полученным параметрам (Рисунок 6.7).

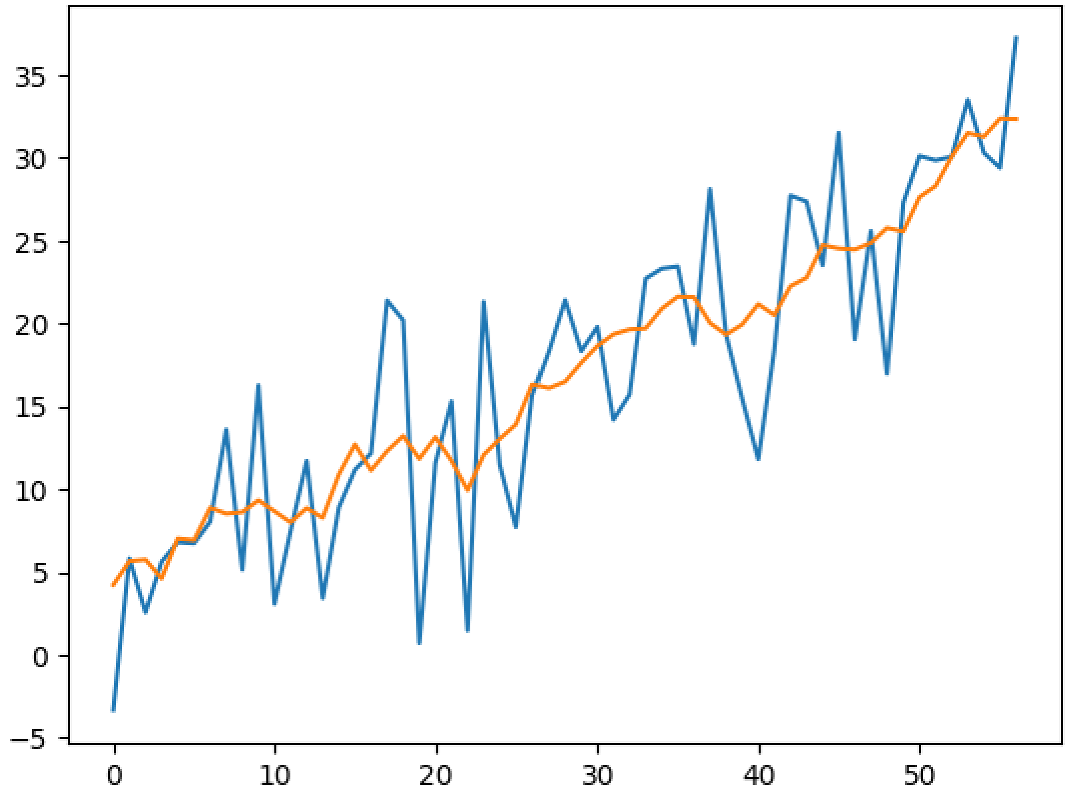

Рисунок 6.7 — Сглаживание второго временного ряда методом SMA

Вычислим Q-статистику Льюнг-Бокса при m = 5 по формуле (6.7) для проверки автокорреции остатков и выяснения, являются ли остатки случайными. Получим значение Q-статистики, равное 6.2587. Q-статистика распределена по chi^2 при критическом значении, равном 12.832. Таким образом, можно сделать вывод, что остатки являются случайной величиной.

Параметр m = 3 был выбран путём перебора и был выбран таким, чтобы Q-статистика при данном значении параметра была минимальна.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Для того чтобы определить автокорреляцию первого порядка для случайной составляющей временного ряда. Получаемое в результате теста значение равно 2.2619 при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Реализуем метод сглаживания WMA по формуле (6.4). Для этого необходимо выбрать параметр m — количество членов ряда в сумме по одной стороне от центрального значения. По аналогичным предыдущему решению причинам, возьмём m = 9. Определим размер окна по формуле (6.1), получим ω = 19. Также необходимо рассчитать веса по формуле (6.3). В отличие от WMA имеет разные веса для всех значений в окне. Веса будут рассчитаны с помощью вычислительных средств.

Построим сглаживание по полученным параметрам (Рисунок 6.8).

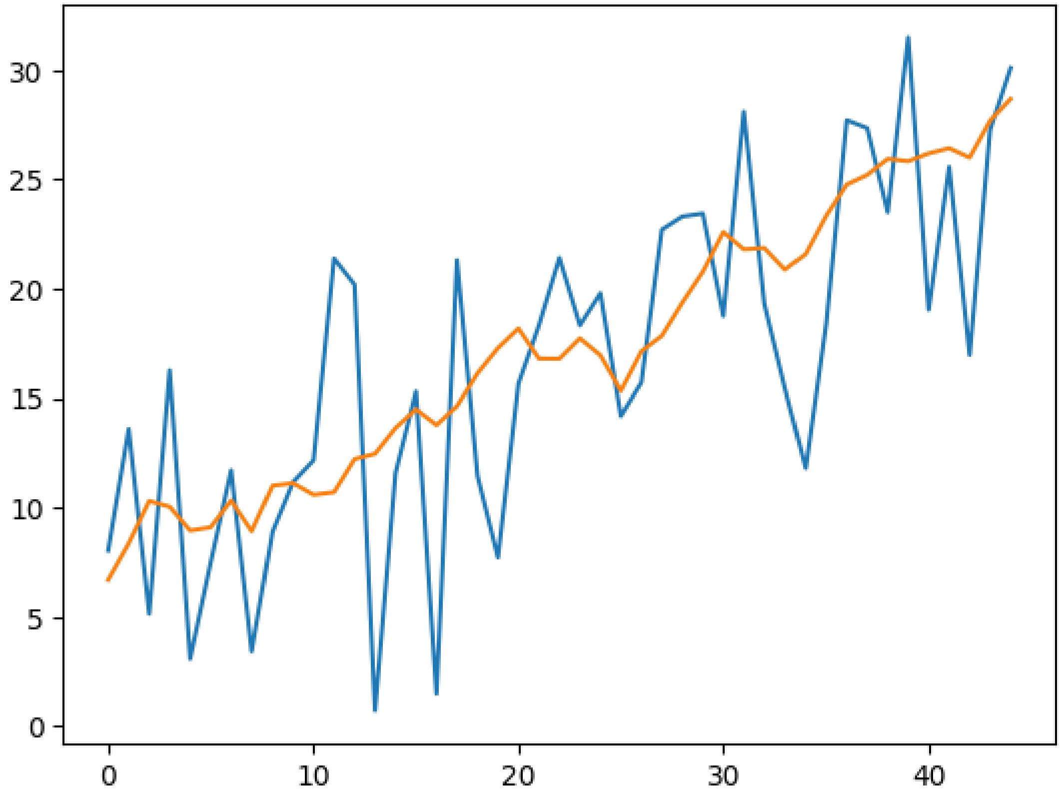

Рисунок 6.8 — Сглаживание второго временного ряда методом WMA

Вычислим Q-статистику Льюнг-Бокса по формуле (6.7) при m = 5. Получим значение Q-статистики, равное 4.6627. Критическое значение Q-статистики равно 12.832, следовательно, остатки являются случайной величиной.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Получаемое в результате теста значение равно 2.0876, при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Реализуем метод сглаживания EMA по формуле (6.5), для этого необходимо выбрать параметр сглаживания a. По описанным в дальнейшем причинам возьмём a = 0.1.

Построим сглаживание по полученным параметрам (Рисунок 6.9).

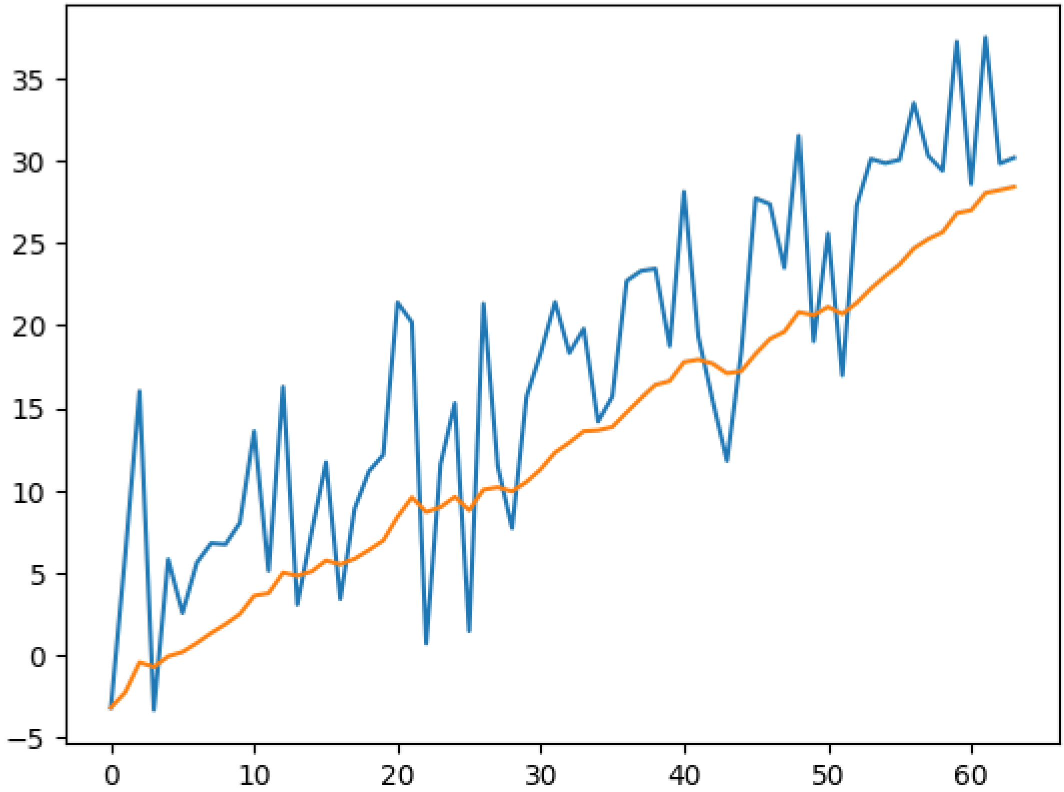

Рисунок 6.9 — Сглаживание второго временного ряда методом EMA

Вычислим Q-статистику Льюнг-Бокса по формуле (6.7) при m = 5. Получим значение Q-статистики, равное 22.851. Критическое значение Q-статистики равно 12.832, следовательно, остатки не являются случайной величиной.

Параметр a = 0.1 был выбран путём перебора и был выбран таким, чтобы Q-статистика при данном значении параметра была минимальна. Параметр выбирался из значений a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Получаемое в результате теста значение равно 1.2132, при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Реализуем метод сглаживания DEMA. Для этого необходимо выбрать параметр сглаживания вокруг тренда a и параметр сглаживания самого тренда b. По описанным в дальнейшем причинам возьмем a = 0.5, b = 0.2. DEMA — двойное экспоненциальное сглаживание и позволяет рекурсивно оценить новое значение сглаженного ряда по исходным данным по формуле (6.6).

Построим сглаживание по полученным параметрам. (Рисунок 6.10).

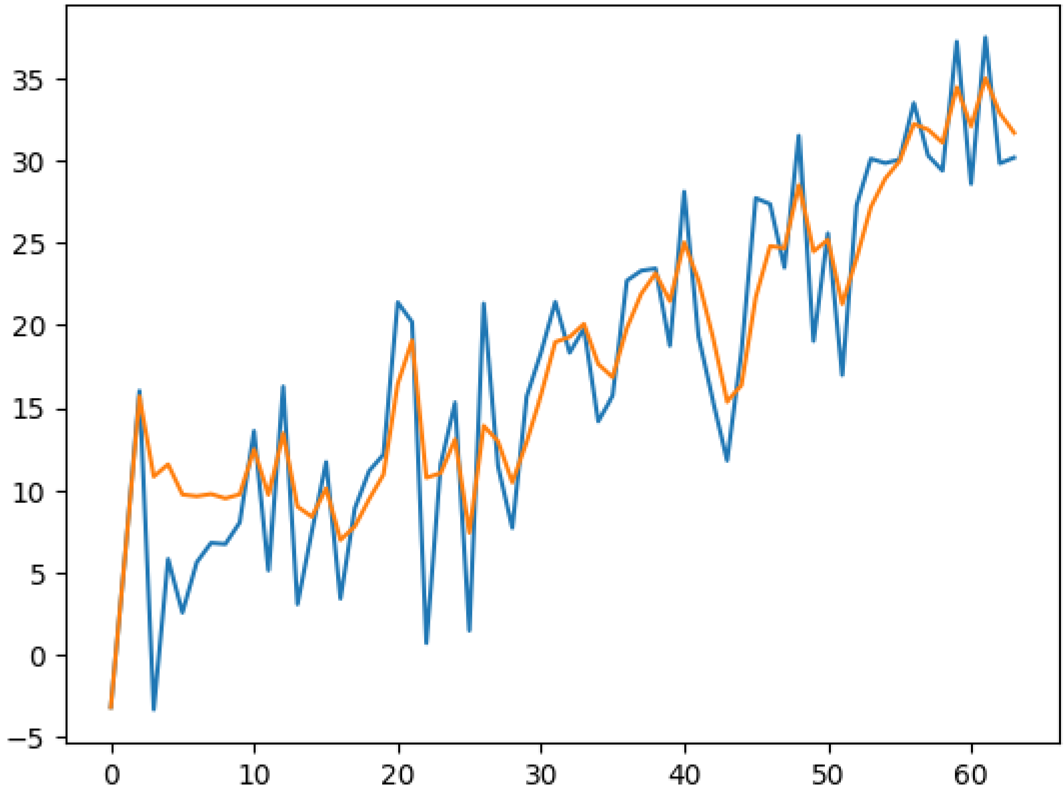

Рисунок 6.10 — Сглаживание второго временного ряда методом DEMA

Вычислим Q-статистику Льюнг-Бокса по формуле (6.7) при m = 5. Получим значение Q-статистики, равное 2.9555. Критическое значение Q-статистики равно 12.832, следовательно, остатки являются случайной величиной.

Параметры a = 0.5 и b = 0.2 были выбраны путём перебора и были выбраны такими, чтобы Q-статистика при данных значениях параметров минимальна. Параметры выбирались из значений a, b = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

Также проведём тест Дарбина-Уотсона по формуле (6.8) (m = 1, a = 0.95). Получаемое в результате теста значение равно 1.8903, при критическом значении, равном 5.024. Таким образом, можно сделать вывод, что выделенная случайная величина случайна.

Мы изучили методы для выделения тренда. Нами были рассмотрены два метода, скользящего среднего, которые отличаются способами подсчета весов. Также были рассмотрены два метода экспоненциального сглаживания. Для каждого из методов подобрали необходимые параметры. Также провели тест Дарбина-Уотсона и тест Льюнг-Бокса.

ЗАКЛЮЧЕНИЕ

В ходе выполнении курсовой работы был рассмотрен теоретический аспект расчета статистических параметров, построения доверительных интервалов точечных оценок, идентификации распределений, генерации распределений, оценок адекватности модели, сглаживания временных рядов, а также были построены регрессионные модели и сделаны статистические выводы о различных объектах исследования.

Для изучения предмета статистики разработаны и применяются специфические приемы, которые в совокупности образуют методологию статистики. Понятие статистики можно определить как совокупность приемов, применяемых ею для познания своего предмета. Сама статистика также выступает методом познания и для других общественных наук.

Задачи, решаемые в данной курсовой работе:

- расчет статистических параметров;
- построение доверительного интервала точечных оценок;
- идентификация распределений;
- генерация распределений;
- оценка адекватности моделей;
- сглаживание временных рядов.

Цель курсовой работы — применение знаний, полученных на курсе «Прикладные задачи математической статистики» — достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 РАСЧЕТ СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ

1.1 С. Е. Демин, Е. Л. Демина, М-во образования и науки РФ, ФГАОУ ВО «УрФУ им. первого Президента России Б.Н.Ельцина», Нижнетагил. технол. ин-т (фил.). — Нижний Тагил : НТИ (филиал) УрФУ, 2016. – 284 с.

1.2 О.И. Хайруллина, О.В. Баянова; Министерство сельского хозяйства Российской Федерации, федеральное государственное бюджетное образовательное учреждение высшего образования «Пермский аграрно-технологический университет имени академика Д.Н. Прянишникова». — Пермь : ИПЦ «Прокростъ», 2019 —176 с.

2 ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ТОЧЕЧНЫХ ОЦЕНОК

2.1 К. О. Кизбикенов. — Барнаул :АлтГПУ,2017 — 115 с.Вероятностные разделы математики / Под ред. Максимова Ю.Д. — Спб.: «Иван Фёдоров», 2001. — С. 400. — 592 с.

2.2 Карасёва Л. А. Статистика // Всемирная история экономической мысли: В 6 томах / Гл. ред. В. Н. Черковец. — М.: Мысль, 1987. — Т. I. От зарождения экономической мысли до первых теоретических систем политической жизни. — С. 484—494. — 606 с. — 20 000 экз. — ISBN 5-244-00038-1.

3 ИДЕНТИФИКАТОР РАСПРЕДЕЛЕНИЙ

3.1 Миклашевский И. Н. Статистика теоретическая // Энциклопедический словарь Брокгауза и Ефрона : в 86 т. (82 т. и 4 доп.). — СПб., 1890—1907.

3.2 Норман Дрейпер, Гарри Смит. Прикладной регрессионный анализ. Множественная регрессия — 3-е изд. — М.: «Диалектика», 2007. — С. 912. — ISBN 0-471-17082-8.

4 ГЕНЕРАЦИЯ РАСПРЕДЕЛЕНИЙ

4.1 Орлов А. И. Прикладная статистика. Учебник. — М.: Экзамен, 2006. — 671 с.

4.2 ДареллХафф. Как лгать при помощи статистики. — М.: Альпина Паблишер, 2015. — 163 с. — ISBN 978-5-9614-5212-9.

5 ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ, ОЦЕНКА ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ ПАРАМЕТРОВ

5.1 Глинский В. В., Ионин В. Г. Статистический анализ. — М.: Инфра-М, 2002. — 241 с. — (Высшее образование). — 5000 экз. — ISBN 5-16-001293-1.

5.2 Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.

6 СГЛАЖИВАНИЕ ВРЕМЕННЫХ РЯДОВ

6.1 White C. Unkind cuts at statisticians (англ.) // The American Statistician. — 1964. — Vol. 18, no. 5. — P. 15—17.