МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Физтех-школа радиотехникик и компьютерных наук

Исследование продвинутых ститистических методов на примере обработки физического эксперимента

Автор:
Шипилов Степан Юрьевич
Б01-301

Долгопрудный
28 января 2025 г.

Содержание

1. Введение и постановка задачи

Во время обработки результатов, полученных на лабораторных практикумах по физике, часто неправильное применение статистических методов может приводить к неправильным выводам. Некорректный учёт тех или иных погрешностей, или вовсе пренебрежением какими-то из них может исказить конечный результат.

В своё очередь, если применять более продвинутые статистические методы, то зачастую можно получить не только просто более точный результат, но и во все, в местах, где при первичной оценке результаты не соответствовали ожиданиям, получить соответствие теории.

2. Различные подходы к учёту погрешностей

В данном разделе будут рассмотрены различные подходы к рассмотрению учёта погрешностей.

2.1. Экспериментальные данные

В качестве данных использовались результаты полученные на лабораторной работе по термодинамике 2.4.1. Данные приведены в таблице 1

T ^0c	PПа,
24	2416
25	2537
27	2770
26	2950
28	3238
29	3498
30	3665
31	3917
32	4185
33	4466
34	4759
35	4946
36	5287
37	5660
38	6082

Таблица 1: Данные из лабораторной работы 2.4.1

Теоретическая зависимость будет иметь вид:

$$d(ln(p)) = -\frac{L}{R} \cdot d(\frac{1}{T})$$

Будем искать коэффициент наклона k, из которого будем искать теплоту парообразования L. Результат будем сравнивать с табличным значением:

$$L_{\text{табл}} = 40{,}68 \frac{\text{кДж}}{\text{моль}}$$

2.2. Расчёт коэфициента наклона и его погрешности методом наименьших квадратов

Воспользуемся формулой для расчёта наклона наилучшей прямой:

$$a = \frac{< xy > - < x >< y >}{< x^2 > - < x >^2} \qquad \sigma_a = \sqrt{\frac{1}{n-1}(\frac{D_{yy}}{D_{xx}} - a^2)}$$

Посчитаем это воспользовавшись программой написанной на языке Python:

$$a = 6057 \qquad \sigma_a = 77$$

Так как полученная погрешность отображает случайную погрешность, уточним её, добавив систематическую погрешность

$$\sigma_{\text{сист}} = k \cdot \sqrt{\varepsilon^2_{\frac{1}{T}} + \varepsilon^2_{ln(p)}}$$

Тогда итоговая погрешность, которая учитывает как случайную так и систематичскую погрешности:

$$\sigma_{total} = \sqrt{\sigma^2_{\text{случ}} + \sigma^2_{\text{сист}}} = 85$$

Тогда если мы посчитаем теплоту парообразования:

$$L_{\text{exp}} = (50{,}33 \pm 0{,}71) \frac{\text{кДж}}{\text{моль}} \qquad L_{\text{табл}} = 40{,}68 \frac{\text{кДж}}{\text{моль}}$$

Как видим, значение не совпадает в пределах погрешности, и достаточно далеко от истинного значения. Рассмотрим проблемы которые возникают при таком методе обработки:

- Метод МНК подразумевает, что для каждой точки σ одинаково, то есть мы не учитываем, что для каждой точки может быть разная погрешность
- Значение для систематической погрешности выбирается единственным образом для каждого из параметров(либо медиана погрешности, либо среднее значение)

- При проведении регрессионной прямой, учитываются только погрешности по оси y, а по оси абсцисс погрешностью пренебрегаем
- Погрешность наклона прямой отображает погрешность "разброса"точек, относительно регрессионной прямой, а не показывает неточность получении коэфициента a
- С одинаковым вкладом учитываются значения как с большой погрешностью, так и с маленькой

2.3. Расчёт коэфициента наклона и его погрешности методом χ^2

Так как мы исследуем данные на линейную зависиомость $y = a + bx$, то определим функцию χ^2 следующим видом:

$$\chi^2(a,b) = \sum_{i=1}^{n} \frac{(y_i - (a + bx_i))^2}{\sigma_{y_i}^2}$$

Будем минимизировать эту функцию, тем самым получим параметр χ^2 наилучшей прямой. Тогда получим:

$$a = -5985$$

Найдем погрешность коэфициента, найдя взвешенные остатки:

$$\sigma^2_{\text{остатки}} = \frac{1}{n-2} \sum_{i=1}^{n} \frac{(y_i - (a + bx_i))^2}{\sigma_{y_i}^2} \qquad \sigma_b = \sigma_{\text{остатки}} \cdot \sqrt{\frac{1}{\sum \frac{(x_i - <x>)^2}{\sigma_{y_i}^2}}}$$

Получим:

$$\sigma_b = 77 \qquad \sigma_{total} = \sigma_{total} = \sqrt{\sigma^2_{\text{случ}} + \sigma^2_{\text{сист}}} = 80$$

Тогда получим:

$$L_{\text{exp}} = (49{,}74 \pm 0{,}67)\frac{\text{кДж}}{\text{моль}} \qquad L_{\text{табл}} = 40{,}68\frac{\text{кДж}}{\text{моль}}$$

Рассмотрим минусы использования данного метода:

- Метод χ^2 хоть и учёл погрешность по оси ординат, но всё еще не учитывает погрешность по оси абсцисс.
- Значение для систематической погрешности по оси абсцисс всё также выбирается единственным образом(медиана или среднее значение)

2.4. Расчёт коэфициента наклона методом двумерной взвешенной регрессии

Определим функцию χ^2 похожим образом, но учтем при этом погрешность по оси абсцисс:

$$\chi^2(a,b) = \sum_{i=1}^{n} \frac{(y_i - (a + bx_i))^2}{\sigma_{y_i}^2 + b^2\sigma_{x_i}^2}$$

Для того, чтобы получить коэфициент b, воспользуемся одним из предыдущих методом, чтобы оценить параметр, после этого проведем повторную подгонку.

$$a = -6003$$

Вычислим погрешности для коэфициента наклона:

$$\sigma^2_{\text{остатки}} = \sqrt{\frac{\chi^2}{n-2}} \qquad \sigma_a = \sigma_{\text{остатки}} \cdot \sqrt{\frac{1}{\sum \frac{(x_i - <x>)^2}{\sigma_{y_i}^2 + b^2\sigma_{x_i}^2}}}$$

Получим в итоге значения:

$$L_{\text{exp}} = (49{,}88 \pm 3{,}58)\frac{\text{кДж}}{\text{моль}}$$

Рассмотрим результат

- При вычислении наклона регресионной прямой мы наконец-то учли погрешности по всем направлениям, что достаточно сильно увеличило нашу погрешность
- Значение теплоты парообразования уточнилось и приблизилось к табличному значению по сравнению с прошлыми методами

Однако осталась проблема с тем, что мы учли систематическую погрешность, но не учли случайную. Рассмотрим как это можно исправить.

2.5. Метод Монте-Карло

Метод Монте-Карло состоит в том, чтобы многократно "воспроизводить"эксперимента данные добавляя случайные отклонения к кажому значению x_i и y_i. Чтобы учесть не только систематическую, но и случайную погрешность, вычислим погрешность на оснвое остаточных откланений.

$$\Delta_{\text{остаток}} = y_i - (a + b \cdot x_i) \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - (a + b \cdot y_i))^2}$$

Чтобы можно было корректно применить метод Монте-Карло, должно выполняться два условия:

1) Остатки должны быть распределены нормально

2) Гомоскедатичность остатков

Для того чтобы оценить распределение остатков, построим гистограмму (рис.1) Распределение остатков сильно похоже на нормальное. Проверим это

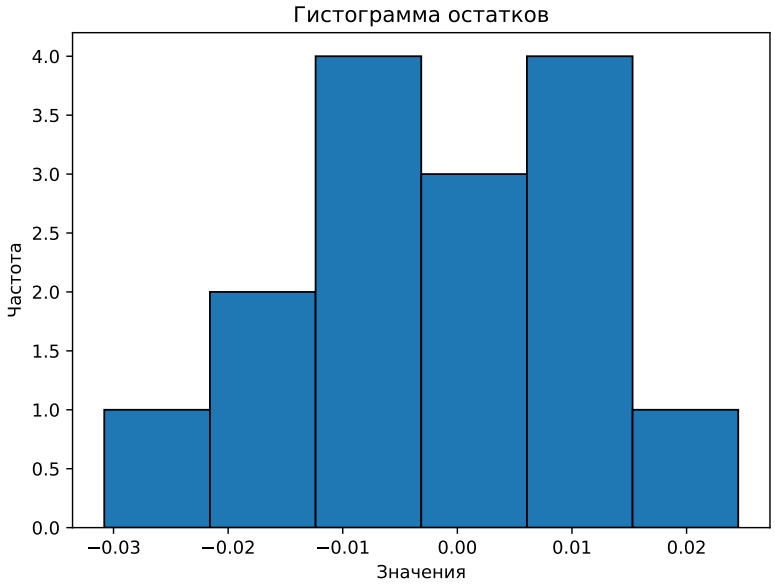

Рис. 1: Гистограмма распределения остатков

точно, воспользовавшись методом Шапиро Уилка:

$$W = \frac{(\sum_{i=1}^{n} \omega_i \cdot x_i)^2}{\sum_{i=1}^{n}(x_i - < x >)^2}$$

где x_i - упорядоченные по возрастанию значения в выборке, ω_i - весовые коэффициенты, которые зависят от ожидаемых значений упорядоченной нормально распределенной выборки, $< x >$ - среднее значение выборки.

Пусть H_0 - гипотеза, что остатки распределены нормально

Посчитаем значение статистики W:

$$W = 0.95 \qquad p_{\text{значение}} = 0{,}47$$

Тогда при 5% уровне значимости можем утверждать в соответствие с тестом Шапиро-Уилка, что данные распределены нормально(не отвергаем H_0)

Проверим данные на гомоскедатичность. Воспользуемся тестом Уайта. Определим статистику:

$$\text{White's statistic} = n \cdot R^2 = 3{,}27 \qquad p_{\text{значение}} = 0{,}2$$

Получаем что для 5% уровня значимости $p_{\text{значение}} > 0{,}05$, что позволяет нам не отклонить гипотезу H_0, и сделать вывод о гомоскедатичности остатков.

Итак, теперь, убедившись в выполнении необходимых условий, можем вычислить параметры регрессии методом Монте-Карло. Моделировать точки будем следующим образом:

$$x_i^{\text{новое}} = x_i + \mathcal{N}(0, \sigma_{x_i})$$

$$y_i^{\text{новое}} = y_i + \mathcal{N}(0, \sqrt{\sigma_{y_i}^2 + RMSE^2})$$

Будем проводить $N = 1000$ итераций. Гистограмма распределения представлена на (рис.2). А сам результат:

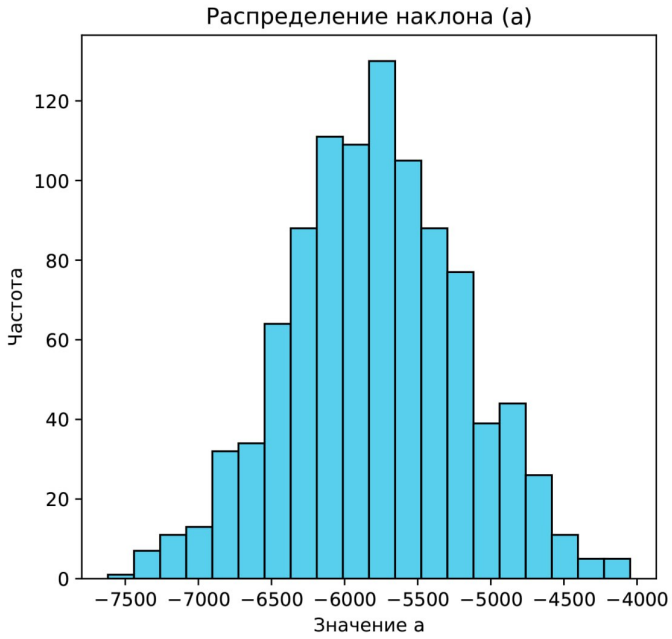

Рис. 2: Гистограмма коэфициента наклона в методе Монте-Карло

$$a = -5779.29616.83 \qquad L = 48{,}02 \pm 5{,}12$$

В итоге, получается, что в данном методе, мы исправили все минусы предшествующих методов.

2.6. Метод Бутсрэпа

Данный метод заключается в том, что на основе данных создаётся бустрэп-выборка, то есть из данных многократно генерируются новые подвыборки с повторением, и по каждой подвыборке строится линейная модель. После этого среднее все коэфициентов наклона усредняется, а в качесвте погрешности берется стандратное отклонение.

В нашем случае получаются следующие параметры:

$$a_{mean} = -5756 \qquad a_{std} = 475$$

А итоговое значение

$$L = 47{,}83 \pm 3{,}95$$

Таким образом, полученное значение оказалось чуть менее точным чем в методе Монте-Карло, однако нам не потребовалось делать дополнительных действий по оценке распределения остатков.

2.7. Результаты

Как мы видим, применение более продвинутых статистических методов не позволило значительно улучшить значение исследуемой величины, так как прирост в точности составил около 5%. Тем не менее, мы получили намного более точную погрешность, которая позволила в пределах 1,5 - 2 σ согласовать табличное значение с экспериментальным. Таким образом, мы видим, что применение более точных методов, позволило изменить вывод о согласии теории с экспериментом. Также рассмотрим, насколько такая большая погрешность справедлива. Построим регрессионную прямую (рис.3): Оценим, в

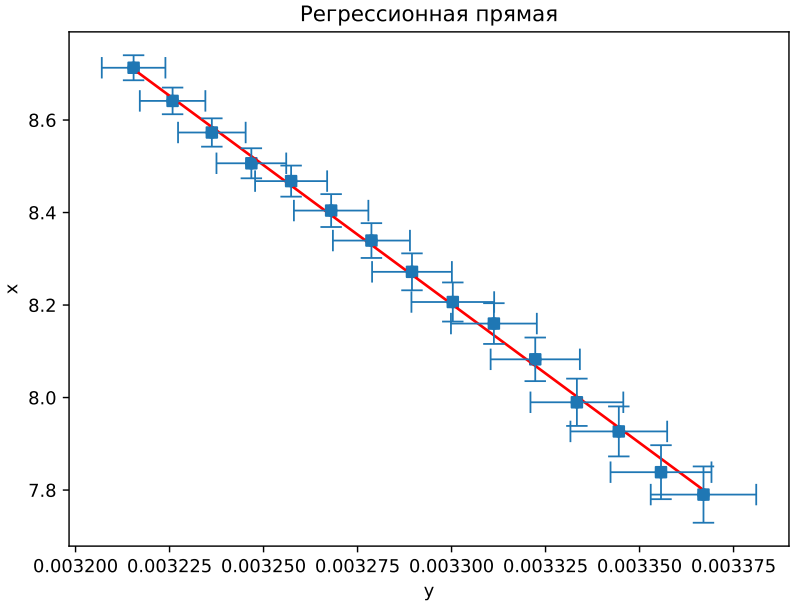

Рис. 3: График линейной регрессии.

какие предельные прямые можно провести через планки погрешностей (рис. 4) Значения прямых лежат лежат в диапазоне:

$$-63000 < k < -5300$$

Это даём нам основания предполагать, что погрешность коэфициента наклона в самом деле может принимать те значения, которые мы получили в ходе обработки.

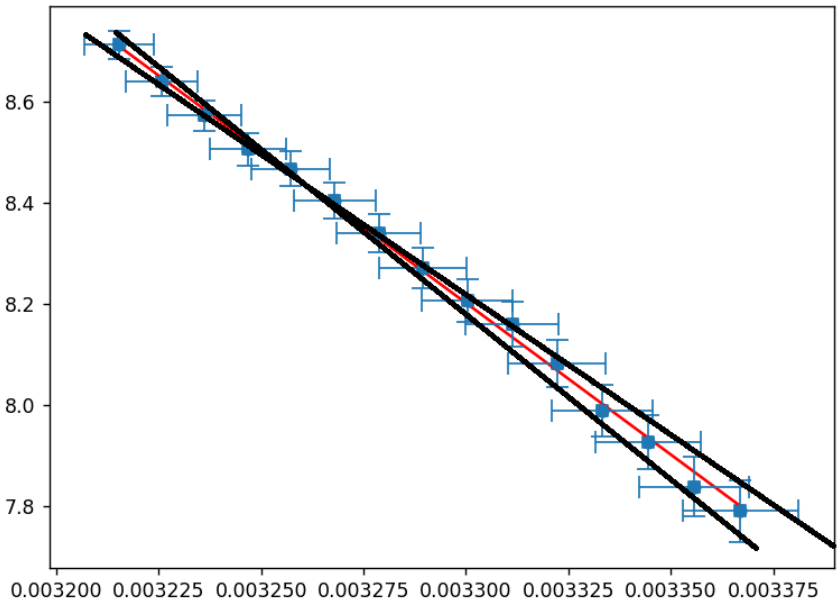

Рис. 4: Предельные прямые

3. Влияние статистического метода на значение искомой величины

Рассмотрим как применение тех или иных способов для вычисления регрессионных параметров может повлиять на результаты исследования

3.1. Экспермиентальные данные

В качестве экспериментальных данных возьмем результатты измерения постоянной времени τ интегрирующей цепочки в лабораторной работе 3.6.1

K	ν, кГц
0,164	300
0,094	600
0,055	900
0,039	1200
0,035	1500
0,027	1800
0,024	2100
0,015	2400

Таблица 2: Данные из лабораторной работы 2.4.1

В данном разделе будет в большей степени показано влияние различных

методов на само значение исследуемой величины, а также показано, как некорректное применение методов при невыполнении тех или иных условий может дать неправильный результат.

3.2. Метод МНК

Воспользуемся формулой для расчёта наклона наилучшей прямой:

$$a = \frac{< xy > - < x >< y >}{< x^2 > - < x >^2} \qquad \sigma_a = \sqrt{\frac{1}{n-1}(\frac{D_{yy}}{D_{xx}} - a^2)}$$

Посчитаем численно значение коэфициента наклона

$$a = 3{,}13 \cdot 10^{-6} \qquad \sigma_a = 0{,}12 \cdot 10^{-6}$$

Так как полученная погрешность отображает случайную погрешность, уточним её, добавив систематическую погрешность

$$\sigma_{\text{сист}} = a \cdot \varepsilon_K$$

Тогда итоговая погрешность, которая учитывает как случайную так и систематичскую погрешности:

$$\sigma_{total} = \sqrt{\sigma^2_{\text{случ}} + \sigma^2_{\text{сист}}} = 0{,}44 \cdot 10^{-6}$$

Тогда если мы посчитаем теплоту парообразования:

$$\tau_{\text{exp}} = (3{,}13 \pm 0{,}44) \cdot 10^{-6} \text{ с} \qquad \tau_{\text{табл}} = 3 \cdot 10^{-6} \text{ с}$$

Как видим, значение совпадает в пределах погрешности, однако само значение отличается от истинного на $\approx 5\%$. Попробуем это исправить.

3.3. Одномерный метод χ^2

Так как мы исследуем данные на линейную зависиомость $y = a + bx$, то определим функцию χ^2 следующим видом:

$$\chi^2(a, b) = \sum_{i=1}^{n} \frac{(y_i - (a + bx_i))^2}{\sigma^2_{y_i}}$$

Будем минимизировать эту функцию, тем самым получим параметр χ^2 наилучшей прямой. Тогда получим:

$$a = 3{,}07 \cdot 10^{-6}$$

Найдем погрешность коэфициента, найдя взвешенные остатки:

$$\sigma^2_{\text{остатки}} = \frac{1}{n-2} \sum_{i=1}^{n} \frac{(y_i - (a + bx_i))^2}{\sigma^2_{y_i}} \qquad \sigma_b = \sigma_{\text{остатки}} \cdot \sqrt{\frac{1}{\sum \frac{(x_i - < x >)^2}{\sigma^2_{y_i}}}}$$

Получим:

$$\sigma_b = 0{,}17 \cdot 10^{-6} \qquad \sigma_{total} = \sqrt{\sigma^2_{\text{случ}} + \sigma^2_{\text{сист}}} = 0{,}46$$

Тогда если мы посчитаем временную постоянную

$$\tau_{\text{exp}} = (3{,}07 \pm 0{,}46) \cdot 10^{-6} \text{ с}$$

Отметим несколько особенностей использованного метода:

- По сравнению с прошлым метоодом(МНК), погрешность сильно не изменилось
- Однако, на целых 3% изменилось само исследуемое значение. Это говорит о более высокой точности оценки параметра

Результат апроксимации приведён на (рис. 5)

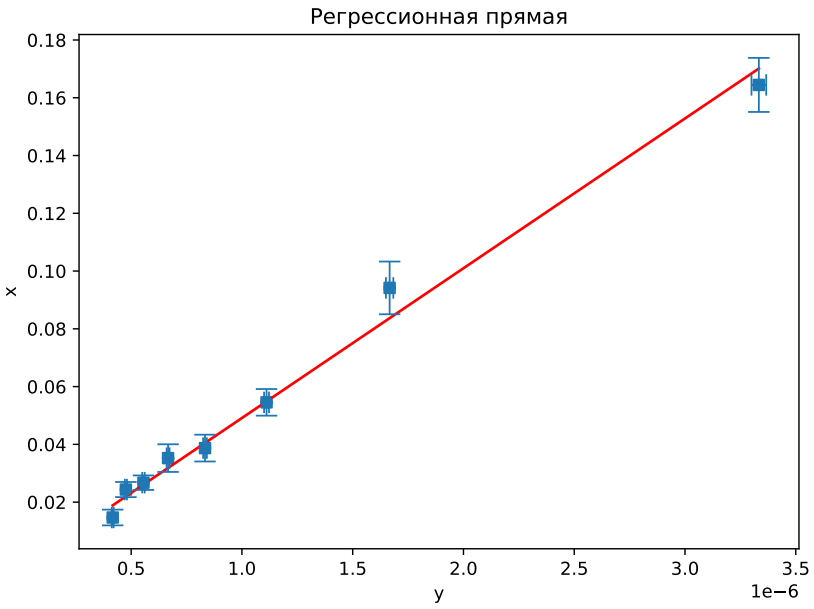

Рис. 5: Линейная апроксимация

3.4. Расчёт коэфициента наклона методом двумерной взвешенной регрессии

Определим функцию χ^2 в случае двумерной взвешенной регрессии

$$\chi^2(a,b) = \sum_{i=1}^{n} \frac{(y_i - (a + bx_i))^2}{\sigma^2_{y_i} + b^2\sigma^2_{x_i}}$$

Для того, чтобы получить коэфициент b, воспользуемся одним из предыдущих методом, чтобы оценить параметр, после этого проведем повторную подгонку.

$$a = -3{,}07$$

Вычислим погрешности для коэфициента наклона:

$$\sigma^2_{\text{остатки}} = \sqrt{\frac{\chi^2}{n-2}} \qquad \sigma_a = \sigma_{\text{остатки}} \cdot \sqrt{\frac{1}{\sum \frac{(x_i - <x>)^2}{\sigma^2_{y_i} + b^2\sigma^2_{x_i}}}}$$

Получим в итоге значения:

$$\tau_{\text{exp}} = (3{,}07 \pm 0{,}46) \cdot 10^{-6} \text{ с}$$

Как видим, с точностью до второго знака, значения совпали с предыдущим пунктом. Поймем причины этого

- Данный пример нагялдно показывает, что применение метода χ^2 порой черевато огромными неточностями. Мы отчётливо видим, что в случае, когда погрешность по оси абсцисс была ощутимой, наши значение в предыдущем разделе изменили оцень внушительно. Сейчас же, когда мы можем пренебречь погрешностями по оси абсцисс, отличие между одномерной и двумерной регрессией практически отсутствуют(как минимум с точностью до второго знака после запятой) Это показывает, что что хоть метод χ^2 и даёт большую точность по сравнению с привычным МНК, но использование его "упрощенной"версии, не всегда можеть быть корректно, а порой может привести и к большим неточностям при обработке результатов.

3.5. Метод Монте-Карло

Попробуем применить метод Монте-Карло в данном случае. Аналогично, чтобы можно было применить его корректно, должны выполянться два условия.

1) Остатки должны быть распределены нормально

2) Гомоскедатичность остатков

Для того чтобы оценить распределение остатков, построим гистограмму (рис.1) Как и в предыдущем разделе, распределение очень похоже на нормальное. Однако, давайте проведем тест Шапиро-Уилка.

$$W = \frac{(\sum_{i=1}^{n} \omega_i \cdot x_i)^2}{\sum_{i=1}^{n}(x_i - < x >)^2}$$

где x_i - упорядоченные по возрастанию значения в выборке, ω_i - весовые коэффициенты, которые зависят от ожидаемых значений упорядоченной нормально распределенной выборки, $< x >$ - среднее значение выборки.

Пусть H_0 - гипотеза, что остатки распределены нормально

Посчитаем значение статистики W:

$$W = 0.81 \qquad p_{\text{значение}} = 0{,}04$$

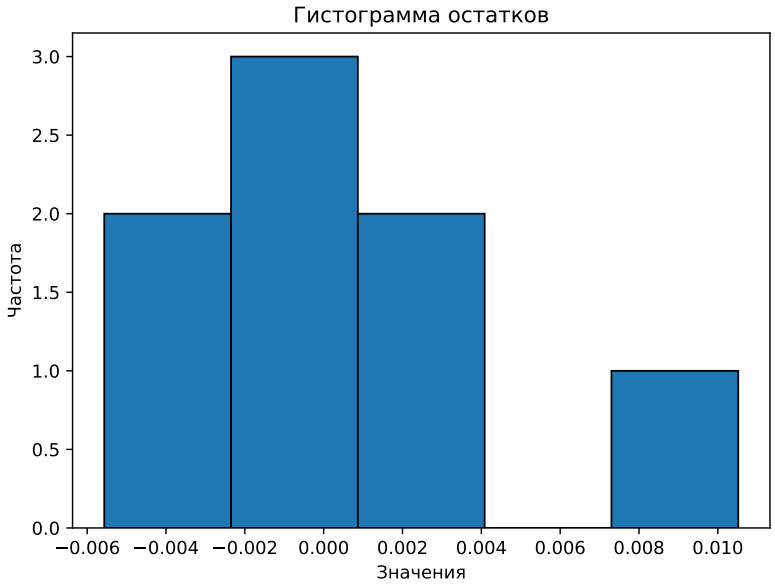

Рис. 6: Гистограмма распределения остатков

Однако в отличие от предыдущего случая, при 5% уровне значимости НЕЛЬЗЯ утверждать, что данные распределены нормально(отвергаем H_0)

Проверим данные на гомоскедатичность. Воспользуемся тестом Уайта. Определим статистику:

$$\text{White's statistic} = n \cdot R^2 = 3{,}72 \qquad p_{\text{значение}} = 0{,}16$$

Получаем что для 5% уровня значимости $p_{\text{значение}} > 0{,}05$, что позволяет нам не отклонить гипотезу H_0, и сделать вывод о гомоскедатичности остатков.

Так как одно из условий невыполняется, то применять корректно метод Монте-Карло нельзя. Посмотрим, что в самом деле результаты не совсем будут соответствовать нашим ожиданиям.

$$x_i^{\text{новое}} = x_i + \mathcal{N}(0, \sigma_{x_i})$$
$$y_i^{\text{новое}} = y_i + \mathcal{N}(0, \sqrt{\sigma_{y_i}^2 + RMSE^2})$$

$$\tau_{\text{exp}} = (3{,}5 \pm 0{,}5) \cdot 10^{-6} \text{ с}$$

Как мы видим, значение в сравнение с прошлыми результатами очень сильно отклонилось от ожидаемого, хотя метод Монте-Карло должен работать точнее, чем предыдущие. Этот пример, показывает важность соблюдения учёта необходимых условий при применении того или иного метода.

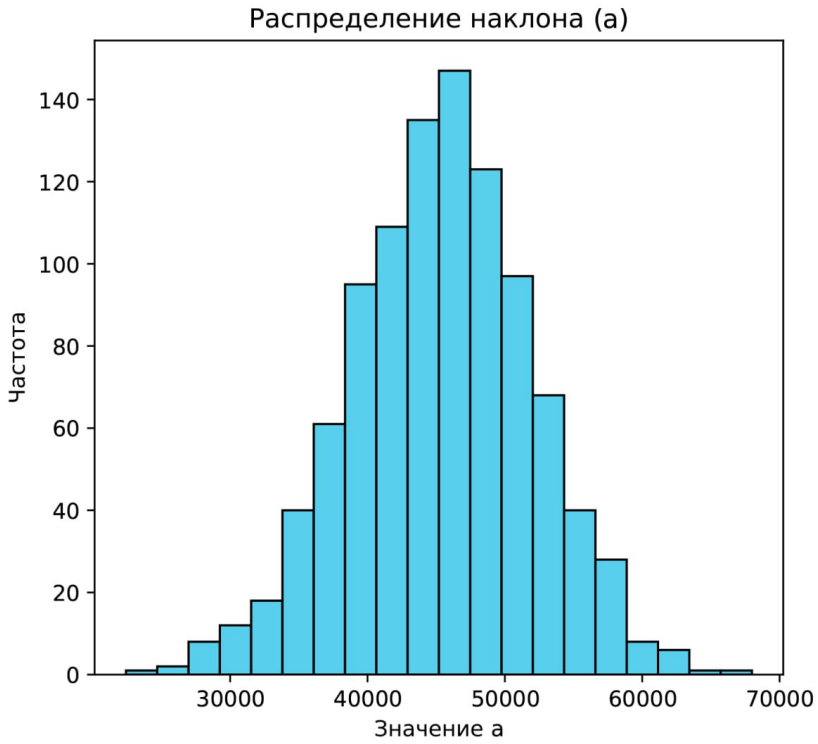

Рис. 7: Гистограмма наклона в методе Монте-Карло

4. Влияние на статстического метода на малые погешности

Бывают случаи, когда величина погрешностей по каждой из осей мала по отноешнию к величие, и может сложиться впечатление, что и погрешность регрессионной прямой также будет пренебрижима мала. Однако, не всегда это может быть правдой, и применение более сложных статистических методов может выявить наличие более значимой ошибки коэфициентов.

4.1. Экспериментальные данные

Возьмём в качестве примера такой набор данных:

x	y	σ_x	σ_y
1	3,9	0,00001	0,05
2	7,9	0,00001	0,05
3	11,9	0,00001	0,05
4	15,9	0,00001	0,05
5	19,9	0,00001	0,05
6	23,9	0,00001	0,05
7	27,9	0,00001	0,05

Таблица 3: Данные из лабораторной работы 2.4.1

4.2. Сравнение статистических методов

Посчитаем коэфициент наклона для приведённого набора данных различными методами. Все результаты занесем в таблицу

Статистический метод	a	σ_a
МНК	4,0	$4{,}2 \cdot 10^{-8}$
Одномерный χ^2	3,9999	0,0067
Двумерный χ^2	3,9999	0,0094
Метод Монте-Карло	3,9999	0,0014
Бутстрэп	3,9999	0,0095

Таблица 4: Данные из лабораторной работы 2.4.1

Как можно заметить, использование простого метода наименьших квадратов не даёт представление об истинной погрешности, поскольку значением порядка 10^{-8} можно принебречь. Исходя из этих данных можно было бы сделать неверный вывод о высокой точности измерений. Однако, применяя остальные статистические методы, можно заметить, что относительная неточность соствляет порядка 0,25%, что также не является большой величиной, но она показывает что ошибки в измерениях присутствуют и они вносят погрешность в итоговые измерения.

5. Результаты

В ходе проделанной работы, мы рассмотрели как применение разлчных статистических методом позволяет улучшить интерпретацию данных полученных при физическом эксперименте. На примере трёх наборов данных, мы получили сначала более корректный учёт ошибок, во втором более корректный учёт значений, а в третьем показали, что важно использовать более сложные статистические методы, чтобы получать корректное представление о погрешностях. Сравнили методы между собой, показали их основные достоинства и недостатки. На реальном примере убедились в важности соблюдения тех или иных условий при применении определённых методов.