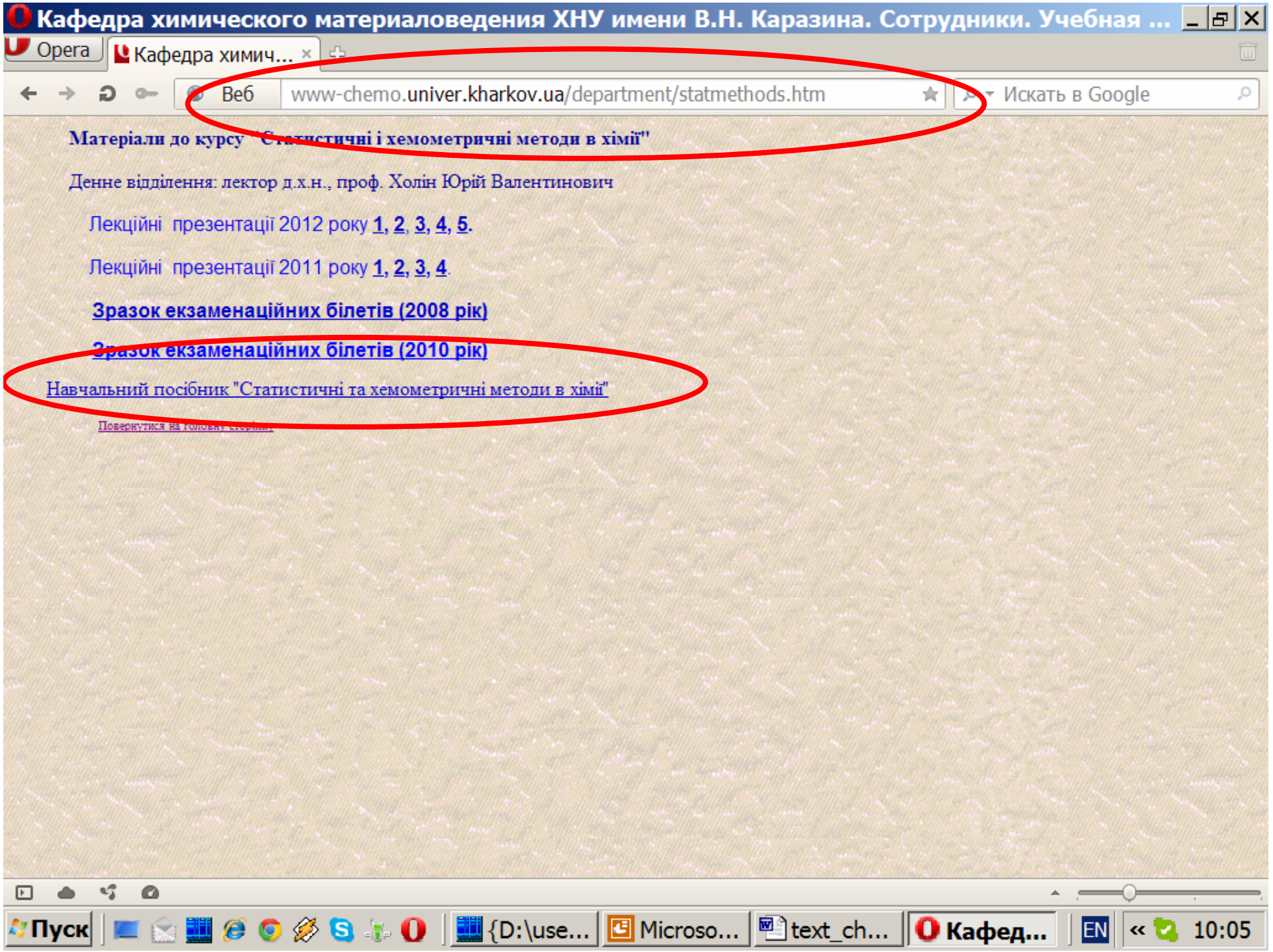
СТАТИСТИЧНІ ТА ХЕМОМЕТРИЧНІ МЕТОДИ В ХІМІЇ

18 тижнів, 18 годин лекційного, 36 годин лабораторного, 108 годин самостійного навантаження

Екзамен

www-chemo.univer.kharkov.ua


Кафедра химического материаловедения ХНУ имени В.Н. Каразина. Сотрудники. Учебная ...
Opera
Кафедра химич...
Веб
www-chemo.univer.kharkov.ua/department/statmethods.htm
Искать в Google
Матеріали до курсу "Статистичні і хемометричні методи в хімії"
Денне відділення: лектор д.х.н., проф. Холін Юрій Валентинович
Лекційні презентації 2012 року 1, 2, 3, 4, 5.
Лекційні презентації 2011 року 1, 2, 3, 4.
Зразок екзаменаційних білетів (2008 рік)
Зразок екзаменаційних білетів (2010 рік)
Навчальний посібник "Статистичні та хемометричні методи в хімії"
Повернутися на головну сторінку
Пуск
{D:\use...
Microso...
text_ch...
Кафед...
EN
10:05

Тема 1. Експериментальні дані. Фактографічна та бібліографічна хімічна інформація. «Дані»: визначення, типи. Шкали: порядкова; відношень; інтервальна. Значення хемометрії та інформатики для хімії (хімічний аналіз, параметрична ідентифікація моделей, QSAR, молекулярна інформатика, автоматизація обробки даних експерименту тощо). Джерела хімічної інформації, бази даних і пакети прикладних програм (Science Citation Index, CSD тощо). Хімічні періодичні видання. Друковані та електронні версії. Імпакт-фактори журналів. Пошук інформації в мережі Інтернет. Статистика, хімічна інформатика, хемометрія.

Тема 2. Представлення та статистична обробка даних. Первинне представлення даних. Дескриптивне представлення даних. Гістограми. Результат вимірювань як випадкова величина. Генеральна сукупність та вибірка. Вибіркові оцінки. Моменти. Середнє. Стандартне відхилення. Дисперсія. Коваріаційні матриці. Коефіцієнти кореляції. Перетворення даних (масштабування, автомасштабне перетворення). Задачі обробки первинних експериментальних даних: дослідження однорідності вибірки, визначення функції розподілу, кореляцій між змінними, класифікація, факторний аналіз. Структурна і параметрична ідентифікація моделей, перевірка адекватності. Статистичні розподіли випадкової величини. Дискретні й неперервні випадкові величини. Біноміальний розподіл. Розподіли неперервних величин: рівномірний, Гауса, Лапласа, Пуассона, $\chi 2$. Центральна гранична теорема. Метод максимуму правдоподібності. Функція правдоподібності. Правдоподібні оцінки параметрів генеральної сукупності при нормальному та Лапласівському розподілах похибок.

Тема 3. Перевірка статистичних гіпотез. Задача перевірка статистичних гіпотез. Схема перевірки гіпотези. Помилки I та II родів. Потужність критеріїв. Перевірка гіпотез про функції розподілу. Критерій $\chi 2$, графічні способи перевірки гіпотез про функції розподілу.

Тема 4. Основи кореляційного та регресійного аналізу. Кореляційний аналіз. Приклади кореляцій в хімії, значення кореляцій. Принцип лінійності вільних енергій як основа багатьох хімічних кореляцій. Теоретичні засади методу найменших квадратів (МНК) та статистичні властивості оцінок МНК. Розрахункова схема МНК. Вибір найкращого набору регресорів: методи всіх регресій, покрокової регресії, вилучення регресорів. Приклади використання МНК у хімічних задачах. Лінійний та нелінійний МНК як приклад некоректної задачі (теоретичний аналіз та приклади), мультиколінеарність. Її формальні та неформальні причини. Способи подолання мультиколінеарності: α регулярізація за Тихоновим, застосування ортогональних поліномів, сингулярний розклад. Типові приклади математично некоректних задач в хімії. Вплив викидів на оцінки МНК. Уявлення про робастні оцінки.

Тема 5. Класифікація та кластерний аналіз. Види класифікацій та їх значення для аналізу даних. Типи ознак. Міри схожості об'єктів. Класифікація без навчання. Ієрархічна класифікація, дендрограми. Найпростіші алгоритми ієрархічного кластерного аналізу сукупності об’єктів. Проблема стійкості класифікації. Факторний аналіз. Характеристика моделей з латентними змінними. Кореляційна та коваріаційна матриці – об'єкт факторного аналізу. Формулювання задачі факторного аналізу. Матриця навантажень, вектори характерностей і загальностей. Основна факторна теорема. Експлораторний та конфірматорний факторний аналіз. Алгоритми факторного аналізу. Факторний аналіз хроматографічних даних.

Тема 6. Елементарні засоби апроксимації експериментальних залежностей. Робота з програмними засобами.
Тема 7. Вивчення властивостей деяких неперервних розподілів випадкових величин.
Тема 8. Перевірка нормальності розподілу випадкових величин за критерієм $\chi 2$.
Тема 9. Апроксимація концентраційних частот виявлення неспадними функціями.
Тема 10. Сингулярний розклад.
Тема 11. Підсумкове заняття.

<table>
<tr><td rowspan="4">Поточне тестування та самостійна робота</td><td>Модуль 1</td><td colspan="6">Модуль 2</td></tr>
<tr><td rowspan="3">Теми 1-5</td><td>Т6</td><td>Т7</td><td>Т8</td><td>Т9</td><td>Т10</td><td>Т11</td></tr>
<tr><td colspan="5">контрольна робота</td><td></td></tr>
<tr><td colspan="2">10</td><td colspan="2">25</td><td colspan="2">25</td><td></td></tr>
<tr><td>Підсумковий семестровий контроль (екзамен)</td><td colspan="7">40</td></tr>
<tr><td>Сума</td><td colspan="7">100</td></tr>
</table>

Приклад екзаменаційного білету

- 1. 14 балів. У паралельних вимірюваннях знаходили масу монети.
- № маса, г № маса, г
- 1 3.025 6 3.111
- 2 3.024 7 3.022
- 3 3.028 8 3.029
- 4 3.027 9 3.021
- 5 3.028 10 3.022
- 1.1. Знайдіть середнє значення, медіану, стандартне відхилення, відносне стандартне відхилення та стандартне відхилення середнього значення. (5 балів)
- 1.2. Приймаючи, що результати вимірювання маси розподілені за законом Гауса, а середнє значення та вибіркова дисперсія є добрими наближеннями до параметрів генеральної сукупності – математичного сподівання μ та дисперсії $\sigma 2$, розрахуйте частки результатів вимірювань, що дають значення маси в інтервалах: а) 3.022-3.025, б) >3.106. (5 балів)
- 1.3. З урахуванням відповіді на питання 1.2.б, поясніть, яку з величин (середнє значення чи медіану) доцільно використовувати як оцінку математичного сподівання μ? (4 бали)

- 2. 9 балів. Молярну масу ідеального газу (М) можна визначити за рівнянням Клапейрона–Менделєєва:

$$M = \frac{mRT}{PV}$$

де m – маса, г; R – універсальна газова стала 0.082056 л·атм/(моль·К), Р – тиск, атм, V – об’єм, л. У досліді одержали такі дані:

- m = 0.118 г (sm = 0.001 г); Т = 298.2 К (sT = 0.05 К); Р = 0.724 атм (sP = 0.005 атм); V = 0.250 л (sV = 0.005 л); всі виміряні величини є незалежними.
- 2.1. Розрахуйте молярну масу газу, її стандартне відхилення та відносне стандартне відхилення. (6 балів)
- 2.2. Яка з виміряних величин дає найбільший внесок у похибку (стандартне відхилення) молярної маси газу? Відповідь обґрунтуйте. (3 бали)

- 3. 10 балів. При вимірюванні масової частки нітрогену в добриві одержали такі результати.

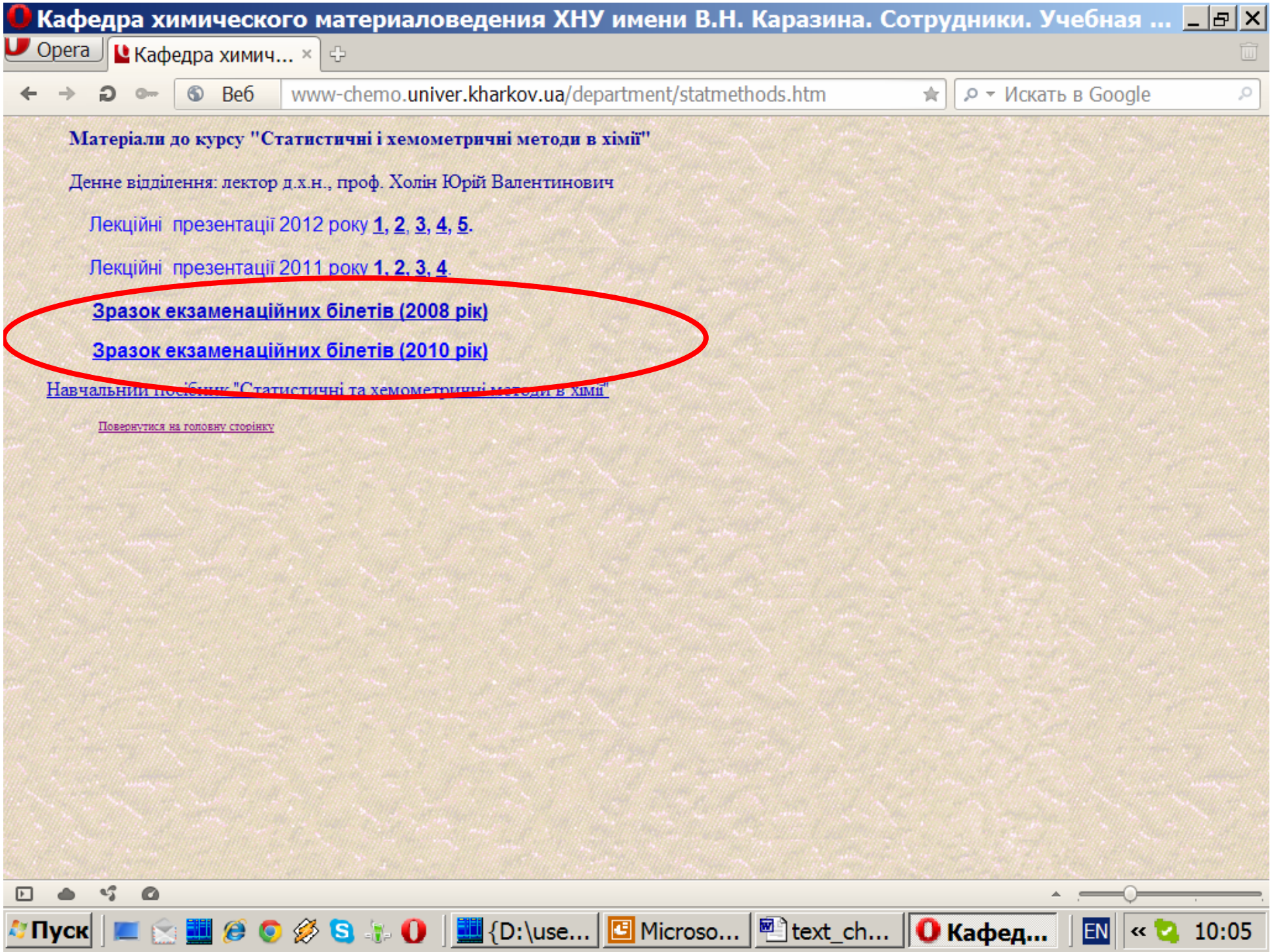
Масова частка, %	Кількість дослідів	Масова частка, %	Кількість дослідів
менше 12.00	1	12.07-12.08	4
12.00-12.02	3	12.08-12.09	3
12.02- 12.03	3	12.09-12.10	4
12.03- 12.04	5	12.10-12.11	2
12.04- 12.05	5	12.11-12.12	2
12.05-12.06	6	12.12-12.13	1

- 3.1. Побудуйте гістограму частот результатів аналізу. (3 бали)
- 3.2. За критерієм $\chi 2$ перевірте гіпотезу про узгодження експериментальних даних з нормальним розподілом. (7 балів)

- 4. 10 балів. В цьому завданні ви повинні оцінити адекватність регресійної моделі за локальними та глобальним критерієм адекватності. В кінетичних дослідженнях за залежністю швидкості реакції (V, моль·л-1·с-1) від концентрації реагентів методом найменших квадратів знайшли кінетичне рівняння. В таблиці наведено швидкості реакції, виміряні та розраховані за знайденим регресійним рівнянням (містить два підгоночні параметри). Відносне стандартне відхилення виміряних швидкостей реакції становить 2.0%.
- № досліду Vексперимент Vрозрахунок № досліду Vексперимент Vрозрахунок
- 1 3.03 2.98 5 4.84 4.70
- 2 4.04 4.11 6 5.12 5.14
- 3 4.56 4.60 7 5.36 5.39
- 4 4.73 4.69 8 5.43 5.37
- 4.1. Розрахуйте статистичні ваги wk та локальні критерії адекватності – зважені залишки ξк = wk1/2·(Vрозрахунок – Vексперимент). (4 бали)
- 4.2. Побудуйте графік залежності ξк від Vрозрахунок та зробіть висновок щодо адекватності регресійної моделі. (3 бали)
- 4.3. Перевірте гіпотезу про адекватність регресійної моделі на основі порівняння з відповідним критичним значенням (для довірчої ймовірності 5%). (3 бали)

- 5. 7 балів Карбон має два стабільні нукліди, 12C та 13C. Мольні частки цих нуклідів становлять, відповідно, 98.89% і 1.11%.
- 5.1. Визначте середнє значення та стандартне відхилення кількості атомів 13C в молекулі холестерину C27H44O. (2 бали)
- 5.2. Якою є ймовірність зустріти в зразку холестерину молекулу, що не містить жодного атома 13C? (5 балів)

Бали	Оцінка	
100-90	Відм.	A
89-80	Добре	B
79-70	Добре	C
69-60	Задовільно	D
59-50	Задовільно	E
<50	Незадовільно	FX


Кафедра химического материаловедения ХНУ имени В.Н. Каразина. Сотрудники. Учебная ...
Opera
Кафедра химич...
Веб
www-chemo.univer.kharkov.ua/department/statmethods.htm
Искать в Google
Матеріали до курсу "Статистичні і хемометричні методи в хімії"
Денне відділення: лектор д.х.н., проф. Холін Юрій Валентинович
Лекційні презентації 2012 року 1, 2, 3, 4, 5.
Лекційні презентації 2011 року 1, 2, 3, 4.
Зразок екзаменаційних білетів (2008 рік)
Зразок екзаменаційних білетів (2010 рік)
Навчальний посібник "Статистичні та хемометричні методи в хімії"
Повернутися на головну сторінку
Пуск
{D:\use...
Microso...
text_ch...
Кафед...
EN
10:05

Література

1. Худсон Д. Статистика для физиков / Д. Худсон. – М. : Мир, 1970. – 295 с.
2. Тейлор Дж. Введение в теорию ошибок / Дж. Тейлор. – М. : Мир, 1985. – 272 с.
3. Уилкс С. Математическая статистика / С. Уилкс. – М. : Наука, 1967. – 632 с.
4. Налимов В. В. Применение математической статистики при анализе вещества / В. В. Налимов. – М. : Гос. изд-во физ.-мат. лит-ры, 1960. – 430 с.
5. Доерфель К. Статистика в аналитической химии / К. Доерфель. – М. : Мир, 1969. – 248 с.
6. Демиденко Е. З. Линейная и нелинейная регрессия / Е. З. Демиденко. – М. : Финансы и статистика, 1981.
7. Родионова О. Е. Хемометрика: достижения и перспективы / О. Е. Родионова, А. Л. Померанцев // Успехи химии. – 2006. – Т. 75, № 4. – С. 302-321.
8. Большев Л. Н. Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – М. : Наука, 1983. – 416 с.
9. Ван дер Варден Б. Л. Математическая статистика / Б. Л. Ван дер Варден. – М. : Изд-во иностр. л-ры, 1960. – 434 с.

1. Хемометрия как междисциплинарная научная дисциплина. Хемометрия и химия, хемометрия и прикладная статистика

«Кто владеет информацией - тот владеет миром»

-

Mayer Amschel Bayern Rothschild;
1744 - 1812

«Каждая попытка применить математические методы для исследования химических проблем должна рассматриваться как абсолютно абсурдная и противоречащая самому духу химии. Если математический анализ когда-либо займет сколько-нибудь значительное место в химии — извращение, которое по счастью почти невероятно — это повлечет за собой повсеместно быстрое вырождение этой науки».

1825 г., Огюст Конт

Эконометрия (~1930), биометрия (~1938), социометрия, психометрия, **хемометрия (1974)**

Основные разделы

- Создание и управление базами данных по химии
- Прогнозирование свойств химических соединений и материалов
- Фармакофоры и фармакофорный поиск
- Молекулярное подобие и поиск по молекулярному подобию
- Виртуальный скрининг
- Компьютерный синтез
- Визуализация и исследование химического пространства
- Молекулярный дизайн химических соединений с заданными свойствами

Хемоинформатика это научная дисциплина, возникшая за последние 40 лет в пограничной области между химией и вычислительной математикой. Было осознано, что во многих областях химии огромный объем информации, накопленный в ходе химических исследований, может быть обработан и проанализирован только с помощью компьютеров. Более того, многие из проблем в химии настолько сложны, что для их решения требуются новые подходы, основанные на применении методов информатики. Исходя из этого, были разработаны методы для построения баз данных по химическим соединениям и реакциям, для прогнозирования физических, химических и биологических свойств соединений и материалов, для поиска новых лекарственных препаратов, анализа спектральной информации, для предсказания хода химических реакций и планирования органического синтеза.

The Obernai Declaration,
assembled on May 29-31, 2006 in Obernai , France 100 scientists from 18 European countries as well as from USA and Canada

Хемоинформатика, наряду с квантовой химией и молекулярным моделированием, является ветвью *теоретической химии* (*theoretical chemistry*) и областью вычислительной (компьютерной) химии.

Хемоинформатика тесно связана с биоинформатикой, и между ними нет четкой границы. Биоинформатику можно считать частным случаем хемоинформатики для биологических макромолекул, а хемоинформатику — распространением биоинформатики на небиологические молекулы. Есть ряд областей, например, *хемогеномика* (*chemogenomics*), которые в равной степени относятся к биоинформатике и хемоинформатике.

На пересечении **хемоинформатики** и фармакологии находится медицинская (фармацевтическая) химия.

На пересечении **хемоинформатики** и аналитической химии находится *хемометрика* (*chemometrics*).

Математическими основами **хемоинформатики**, связанными с представлением химических соединений в виде молекулярных графов, занимается *математическая химия* (*mathematical chemistry*).

ХЕМОМЕТРИЯ

- химическая дисциплина, применяющая математические, статистические и другие методы, основанные на формальной логике, для построения или отбора оптимальных методов измерения и планов эксперимента, а также для извлечения наиболее важной информации при анализе экспериментальных данных.

Д. Массарт

- решает следующие задачи в области химии: как получить химически важную информацию из химических данных, как организовать и представить эту информацию, и как получить данные, содержащую такую информацию.

С. Волд

То, что делают хемометрики

Применение хемометрии в химии

Комарь Н.П. Основы качественного химического анализа. I. Ионные равновесия. – Харьков: Изд-во Харьковского университета, 1955. – 448 с.

- **Аналитическая химия.**
- Физическая химия – исследование кинетики.
- Органическая химия – QSAR.
- Химия полимеров.
- Теоретическая и квантовая химия.

Мировая периодика

Journal of Chemometrics

Chemometrics and Intelligent Laboratory Systems

Analytical Chemistry

Analytica Chimica Acta

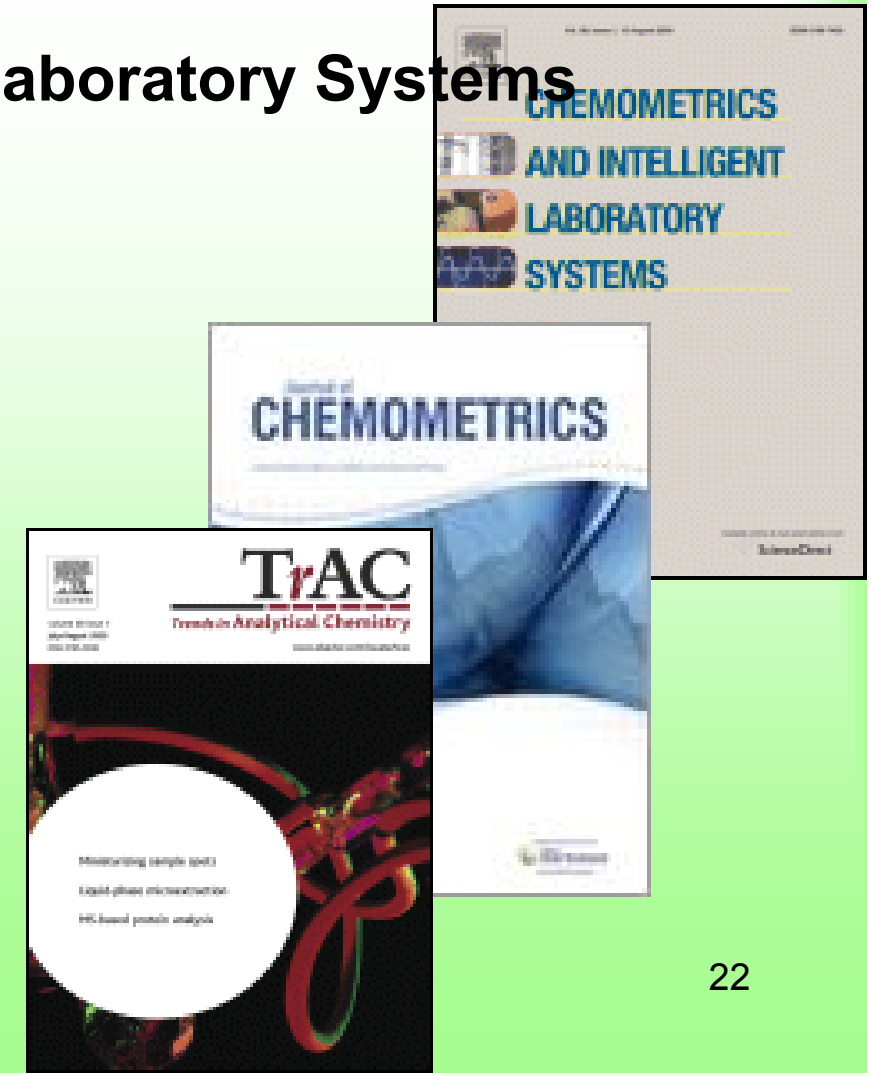
Analyst

Talanta

Trends in Analytical Chemistry

…

www.chemometrics.ru



Прикладная статистика

решает задачи

- описания данных;
- оценивания;
- проверки гипотез.

↓

Построение моделей

(формальных и содержательных)

Хемометрия

• имеет дело с **данными** (зачастую с очень большими), поэтому хемометрия - подраздел информатики (Data mining).

• данные по большей части происходят из **химии**, поэтому хемометрия - подраздел химии.

Решает следующие задачи

• как получить информацию из химических данных.

• как представить эту информацию.

• как получить данные, содержащие такую информацию.

Замена прямых измерений косвенными и их обработка

Основные программные пакеты, используемые в курсе

- Origin
- Excell
- MatLab
- Mathcad
- Statistica

- **Data** is a set of values of qualitative or quantitative variables; restated, data are individual pieces of information. Data in computing (or data processing) are represented in a structure that is often tabular (represented by rows and columns), a tree (a set of nodes with parent-children relationship), or a graph (a set of connected nodes). Data are typically the results of measurements and can be visualised using graphs or images.

Обработка данных включает операции:

- ввод (сбор) данных — накопление данных с целью обеспечения достаточной полноты для принятия решений;
- формализация данных — приведение данных, поступающих из разных источников, к одинаковой форме, для повышения их доступности;
- фильтрация данных — это отсеивание «лишних» данных, в которых нет необходимости для повышения достоверности и адекватности;
- сортировка данных — это упорядочивание данных по заданному признаку с целью удобства их использования;
- архивация — это организация хранения данных в удобной и легкодоступной форме;
- защита данных — включает меры, направленные на предотвращение утраты, воспроизведения и модификации данных;
- транспортировка данных — приём и передача данных между участниками информационного процесса;
- преобразование данных — это перевод данных из одной формы в другую или из одной структуры в другую.

Огляд розрахункових методів

1. Лінеарізація функцій.
2. Аппроксимация функцій.
3. Оптимізация функцій.
4. Чисельні методи (інтегрування, диференціювання).
5. **Застосування вищезгаданого**

Действия с функциями

- Линеаризация — один из методов приближённого представления замкнутых нелинейных систем, при котором исследование нелинейной системы заменяется анализом линейной системы, эквивалентной исходной

Методы линеаризации:
1) Логарифмирования;
2) Обратного преобразования;
3) Комплексный.

Действия с функциями

- Интерполяция – разновидность аппроксимации, при которой кривая построенной функции проходит точно через имеющиеся точки данных;
- Экстраполяция – функция аппроксимируется не между заданными значениями, а вне заданного интервала.

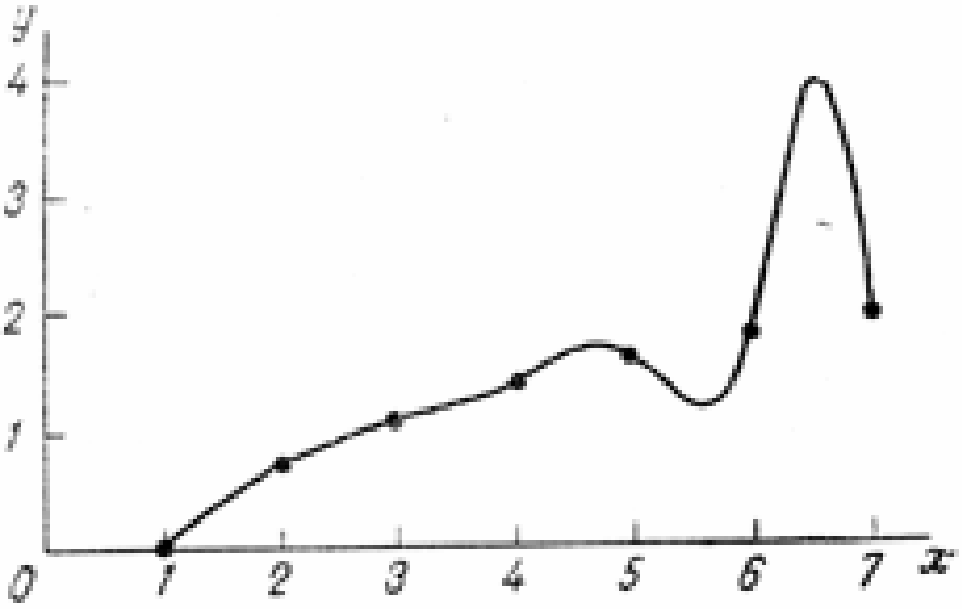
Действия с функциями

- Аппроксимация — замена одних объектов другими, в том или ином смысле близкими к исходным, но более простыми.

Для вычисления значений сложных функций часто используется вычисление значения отрезка ряда, аппроксимирующего функцию

Действия с функциями

- Построение сплайна – функции, совпадающей с функциями более простой природы на каждом элементе разбиения своей области определения

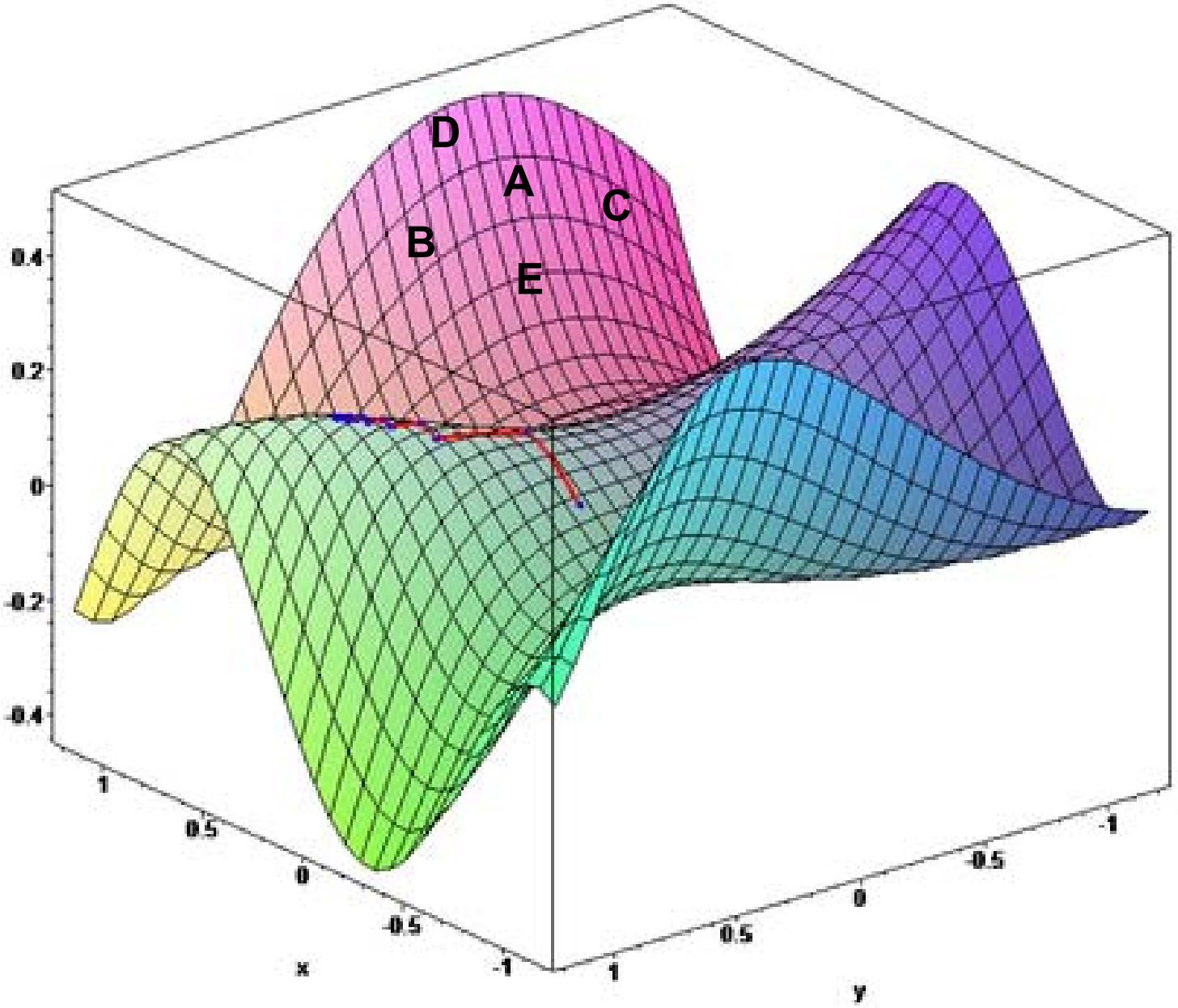


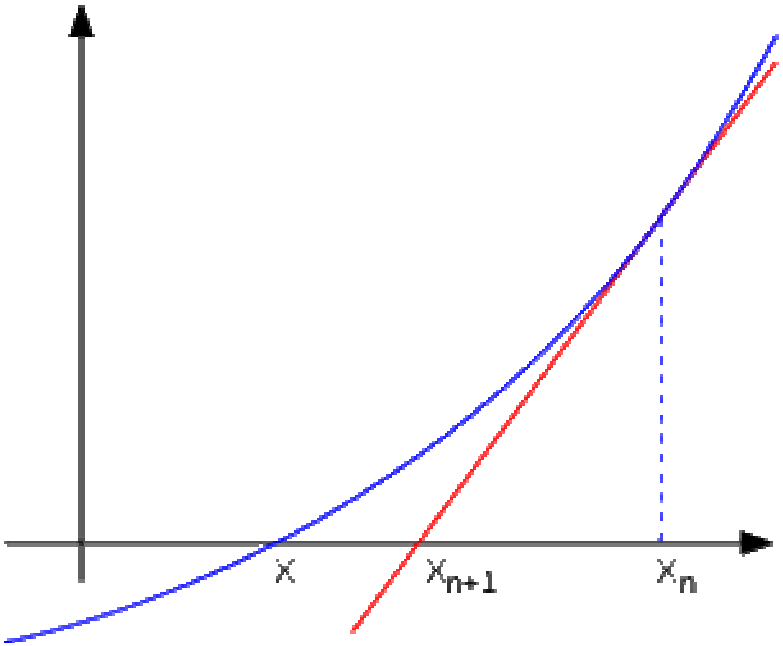
Оптимизация

- Задачей оптимизации в математике называется задача о нахождении экстремума (минимума или максимума) вещественной функции в некоторой области.
- задача определения наилучших, в некотором смысле, структуры или значения параметров объектов.

Постановка задачи

- Формирование допустимого множества.
- Формирование целевой функции.
- Направление поиска оптимума (min, max).


D
A
C
B
E
0.4
0.2
0
-0.2
-0.4
1
0.5
0
-0.5
-1
x
1
0.5
0
-0.5
-1
y



задаётся начальное приближение вблизи предположительного корня,
после чего строится касательная к исследуемой функции в точке
приближения, для которой находится пересечение с осью абсцисс.
Эта точка и берётся в качестве следующего приближения.
И так далее, пока не будет достигнута необходимая точность

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Пример

$$f(x) = x^3 - 2x + 2.$$

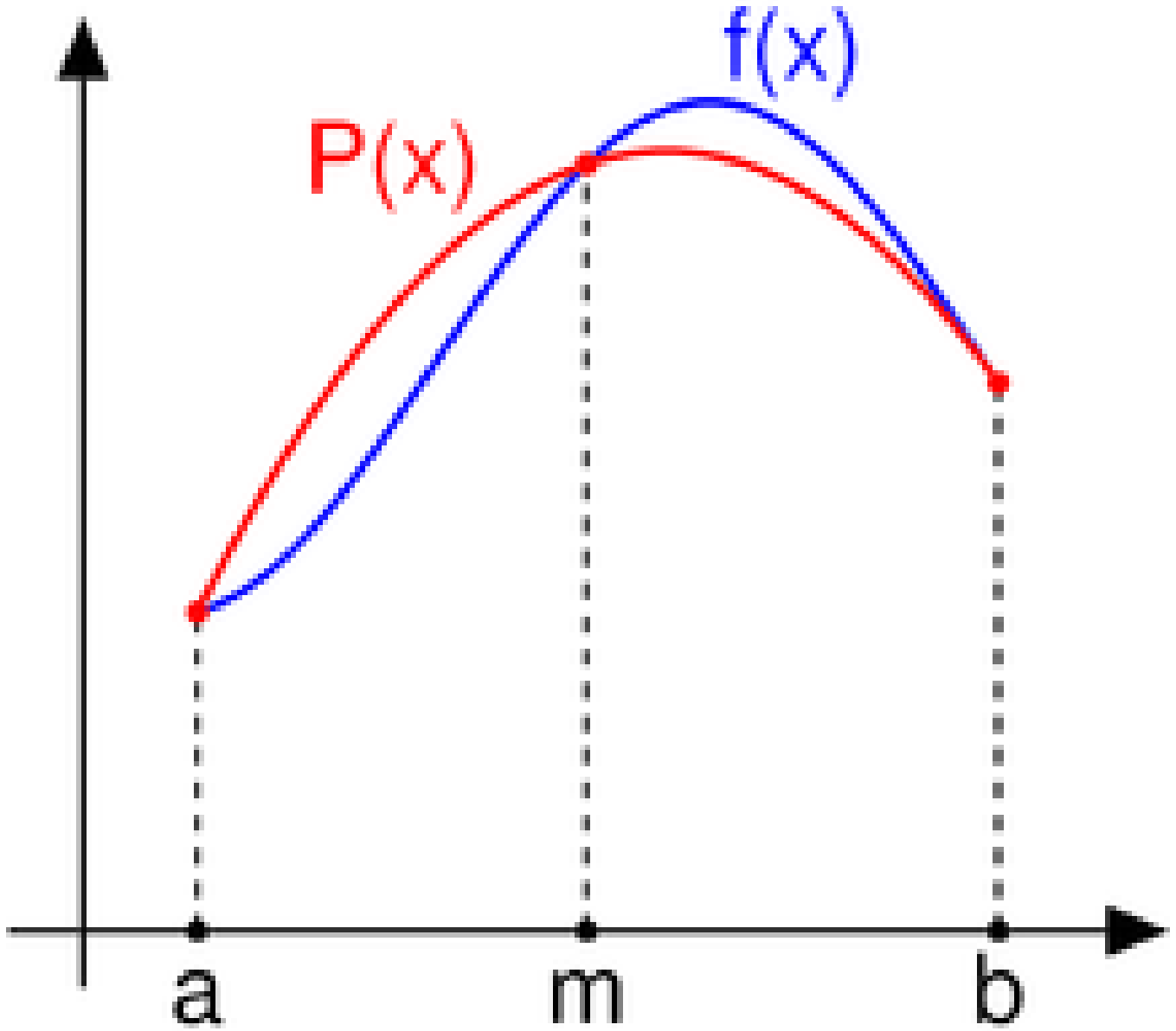
$$x_{n+1} = x_n - \frac{x_n^3 - 2x_n + 2}{3x_n^2 - 2}.$$

Начальное приближение x_n=0. Чему равно x_{n+1}?

Численные методы

набор алгоритмов, позволяющих получать приближенное (численное) решение математических задач

- Дифференцирования;
- Интегрирования;
- Разложения матриц.


f(x)
P(x)
a
m
b

1. Оптимизация

Регрессионный анализ и задача МНК

1) Задача построения калибровочных зависимостей, по которым можно прогнозировать значение условий, которые обеспечивают определенный результат.
2) Определение важных физико-химических характеристик системы – например, зависимость свободной энергии системы от температуры будет линейной, а ее наклон равен изменению энтропии системы с отрицательным знаком.
3) Еще одна задача – сжатие данных, когда подобранная функция используется для описания (аппроксимации) огромного массива экспериментальных данных (например, зависимость теплоемкости от температуры можно представить в виде функции, и в справочниках приводят не экспериментальную таблицу зависимости Ср от Т, а коэффициенты выбранной функции).
4) Аппроксимация экспериментальных данных функцией с целью поиска точек минимума (максимума).

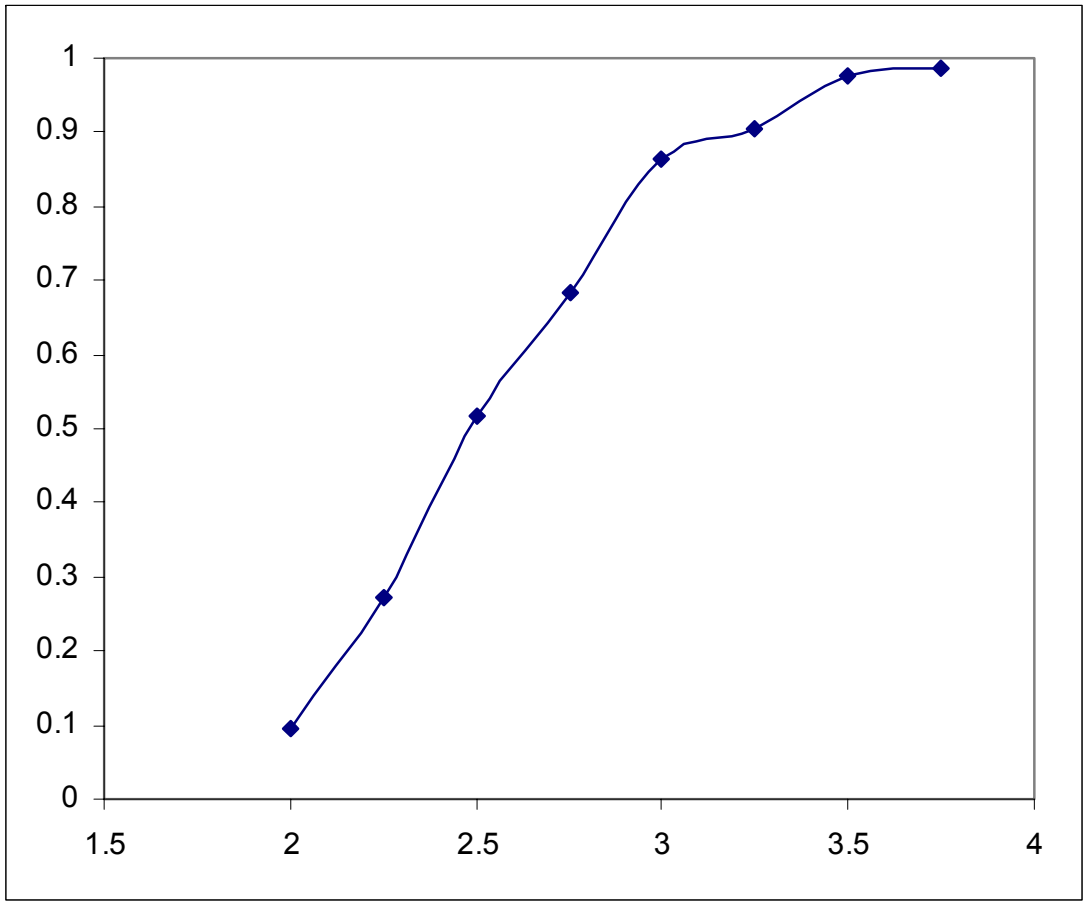
2. Аппроксимация

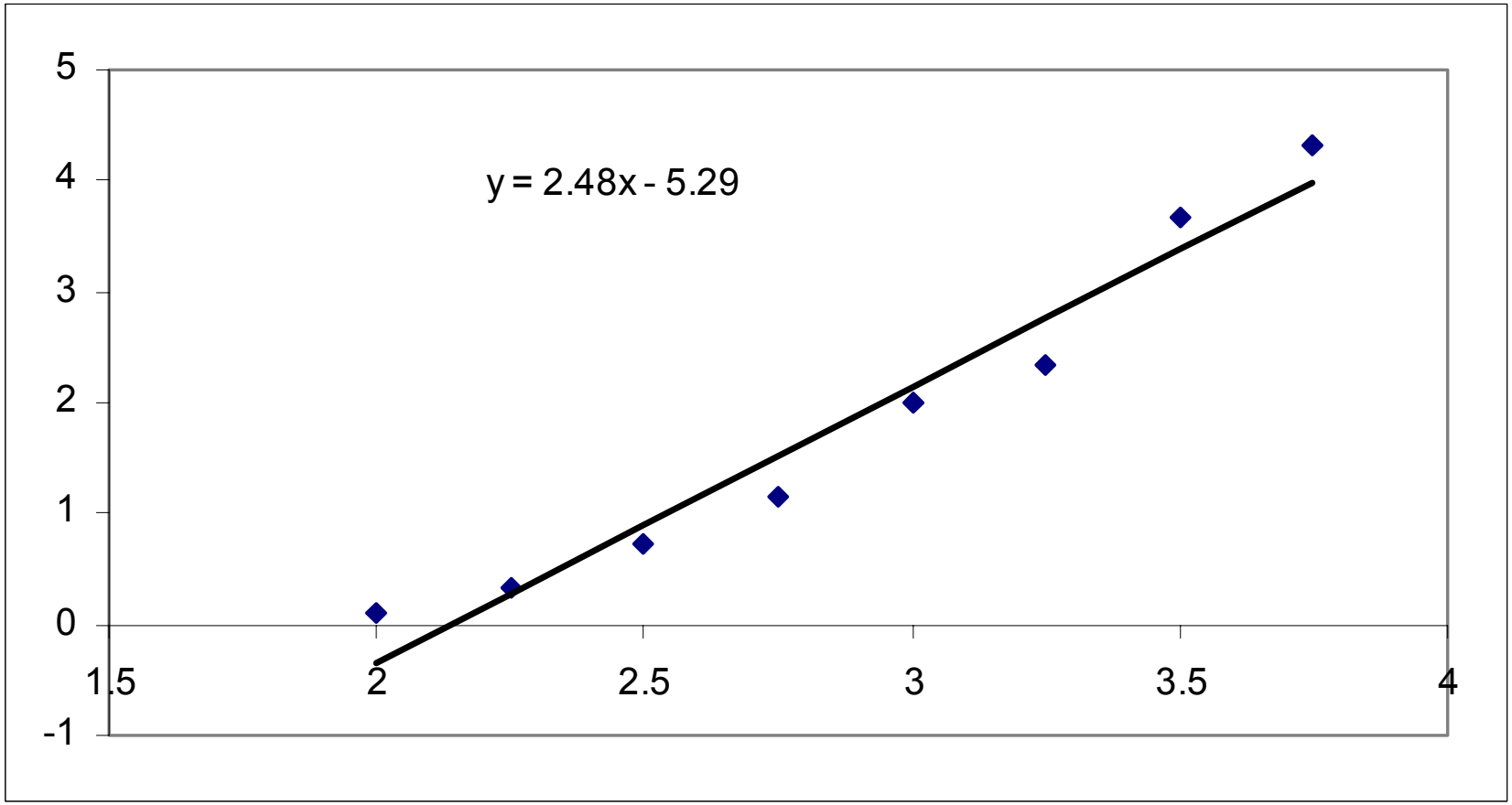
Регрессионный анализ

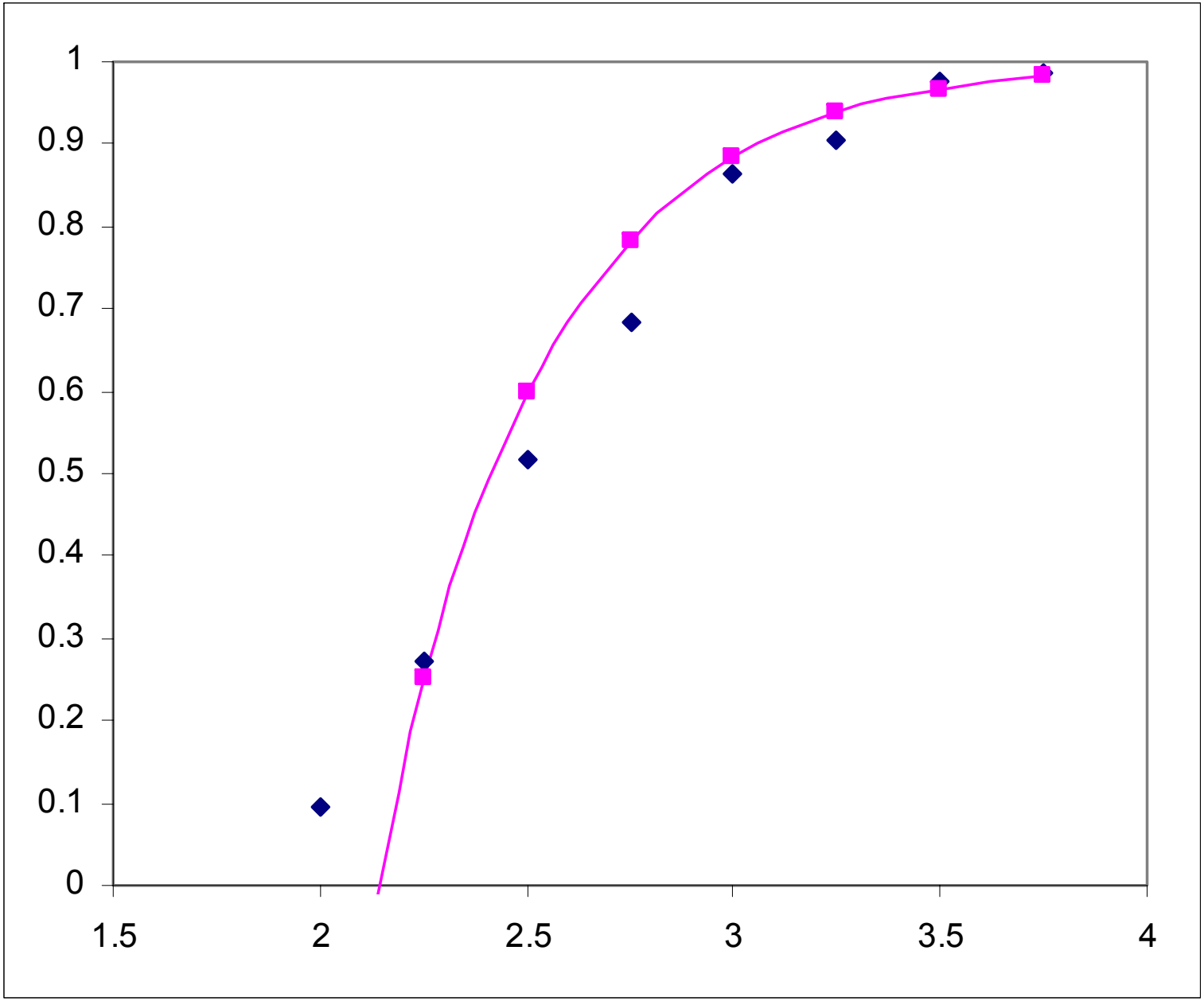
Приближение градуировочной функции прямой позволяет упростить вычисления и просто и наглядно представить градуировочные характеристики.

3. Линеаризация

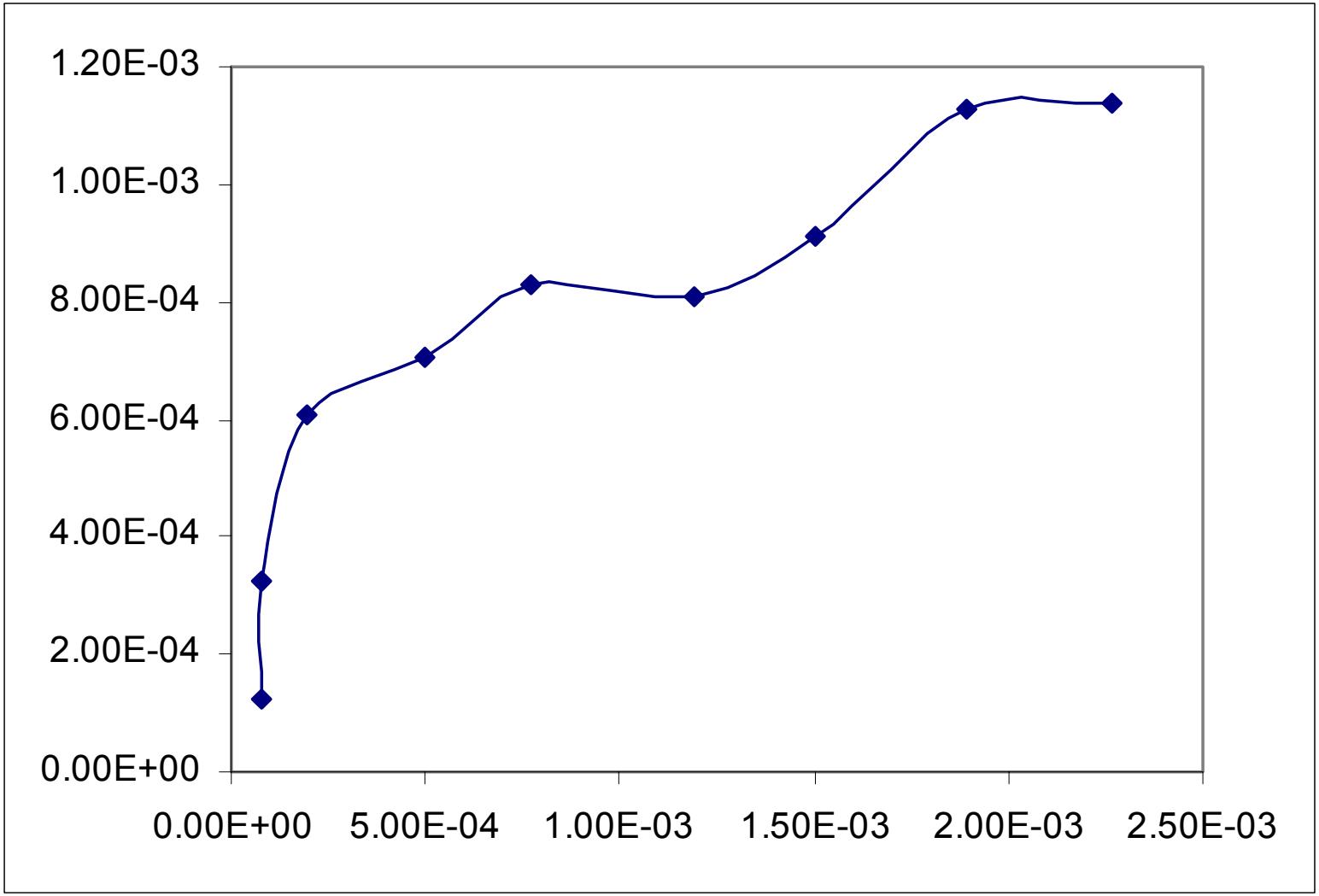
$$P(c) = 1 - \exp\left(\frac{-(c-a)}{b}\right)$$



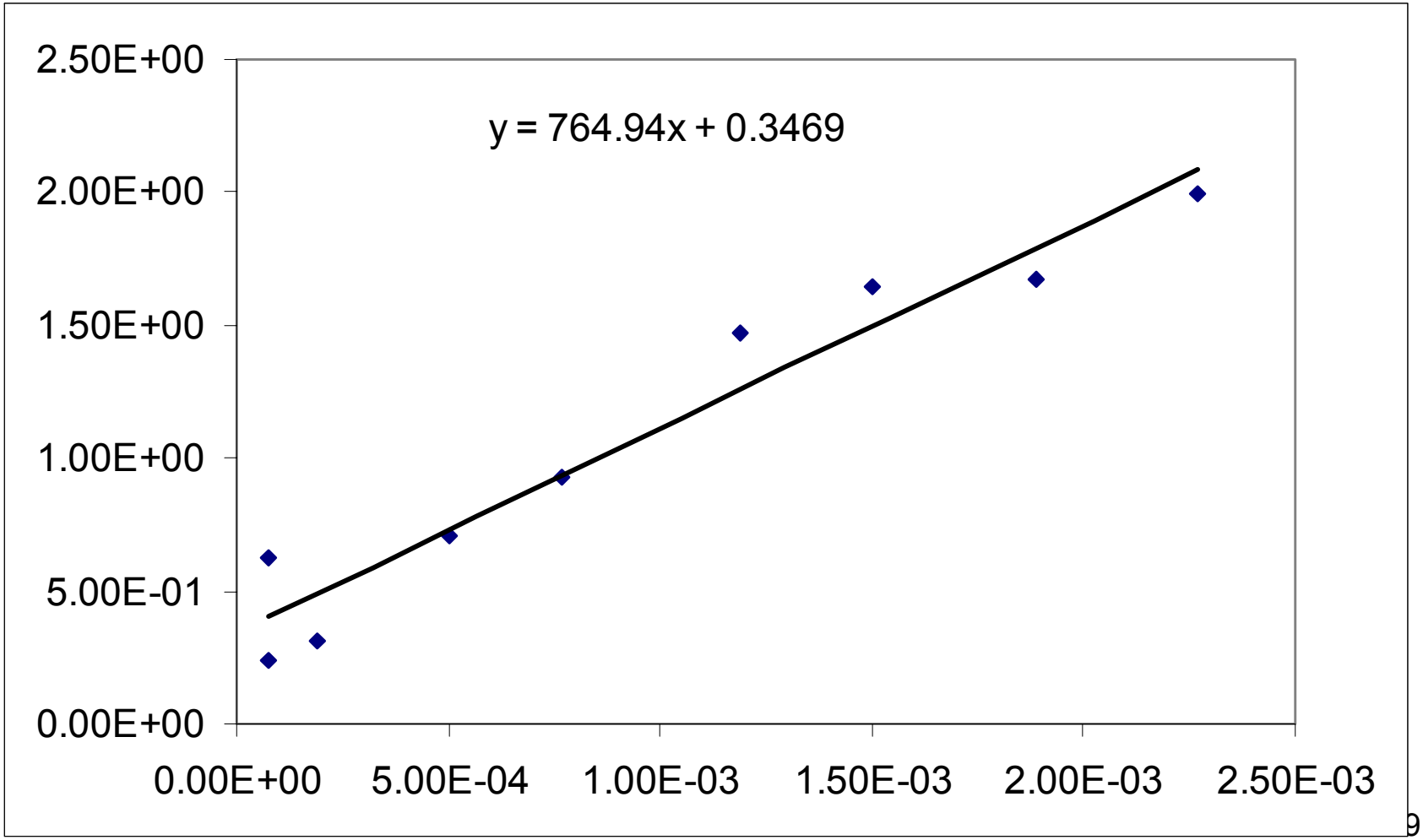

y = 2.48x - 5.29
5
4
3
2
1
0
-1
1.5
2
2.5
3
3.5
4

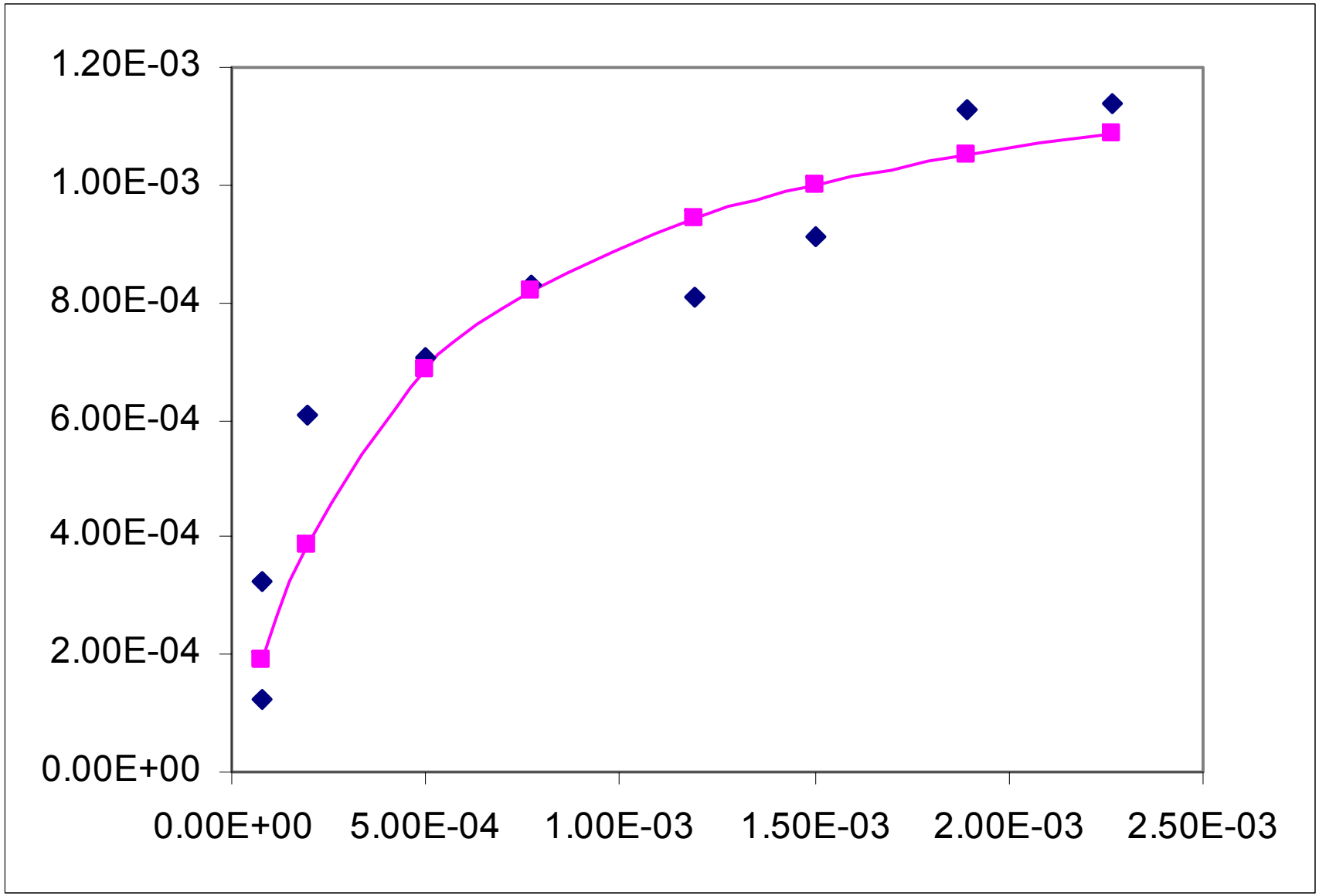

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.5
2
2.5
3
3.5
4

$$A = t_Q \frac{\beta[S]}{1 + \beta[S]}$$

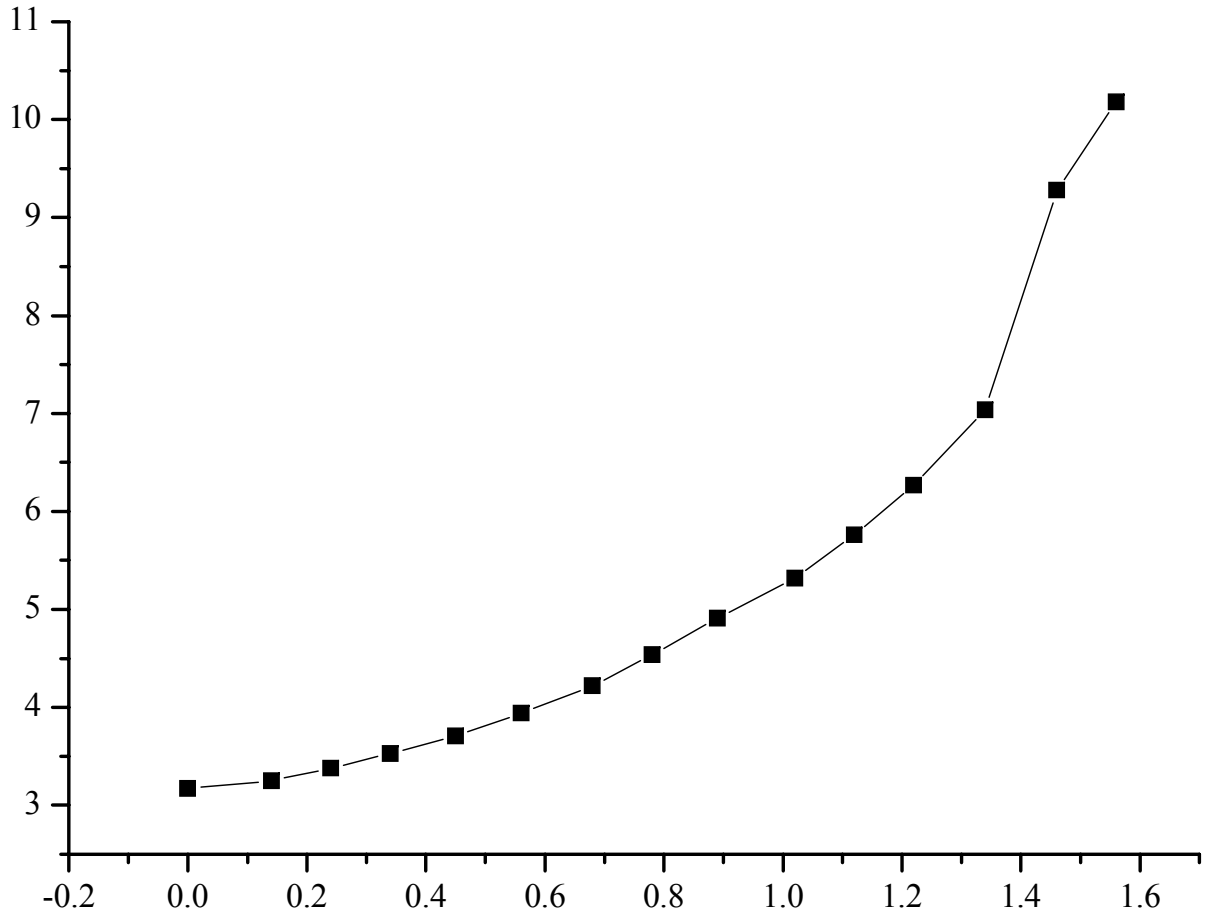


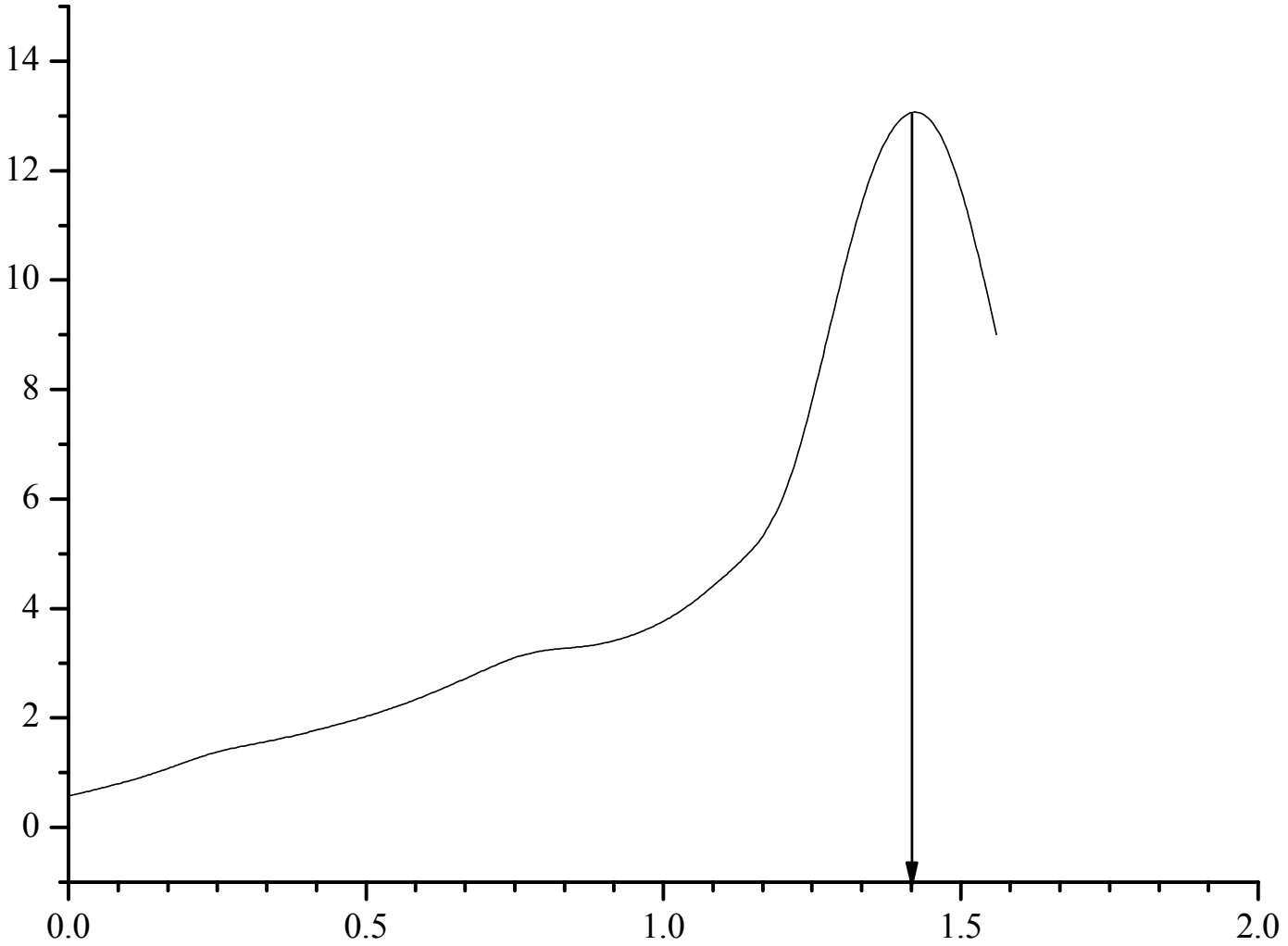
$$\frac{1}{D} = \frac{[S]}{[\overline{SQ}]} = \frac{1}{\beta\, t_Q} + \frac{1}{t_Q}[S]$$



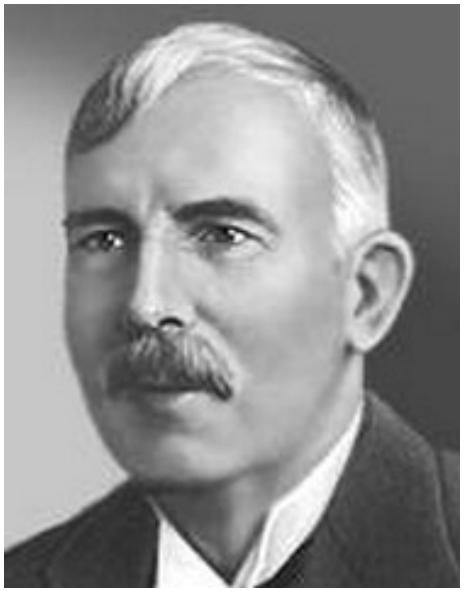

1.20E-03
1.00E-03
8.00E-04
6.00E-04
4.00E-04
2.00E-04
0.00E+00
0.00E+00
5.00E-04
1.00E-03
1.50E-03
2.00E-03
2.50E-03

4. Численные методы

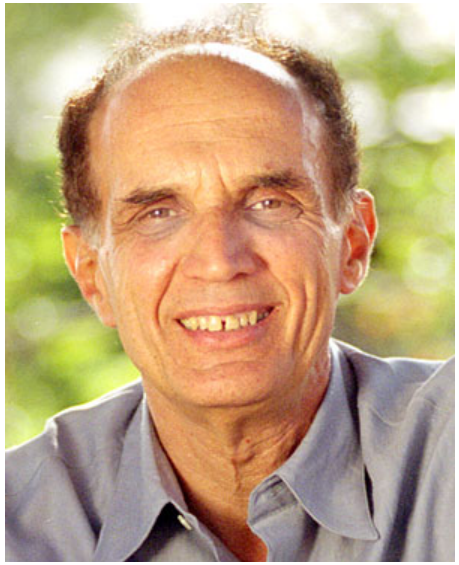



14
12
10
8
6
4
2
0
0.0
0.5
1.0
1.5
2.0

Представлення та статистична обробка даних. Первинне представлення даних. Дескриптивне представлення даних. Гістограми. Результат вимірювань як випадкова величина. Генеральна сукупність та вибірка. Вибіркові оцінки. Моменти. Середнє. Стандартне відхилення. Дисперсія. Коваріаційні матриці. Коефіцієнти кореляції.



Если для Вашего эксперимента требуется статистика, то Вы должны переделать его более тщательно
(Эрнест Резерфорд)



Те, кто игнорируют статистику, обречены изобрести ее заново
(Бредли Эфрон).

Задачи
о бросании монеты,
Шевалье Де Мере,
о Леди, пробующей чай
…

D. Sasburg. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century (2001)

Представление данных

Табличное, матричное, графическое.

Измерение - получение любых количественных характеристик материальных объектов опытным путем.

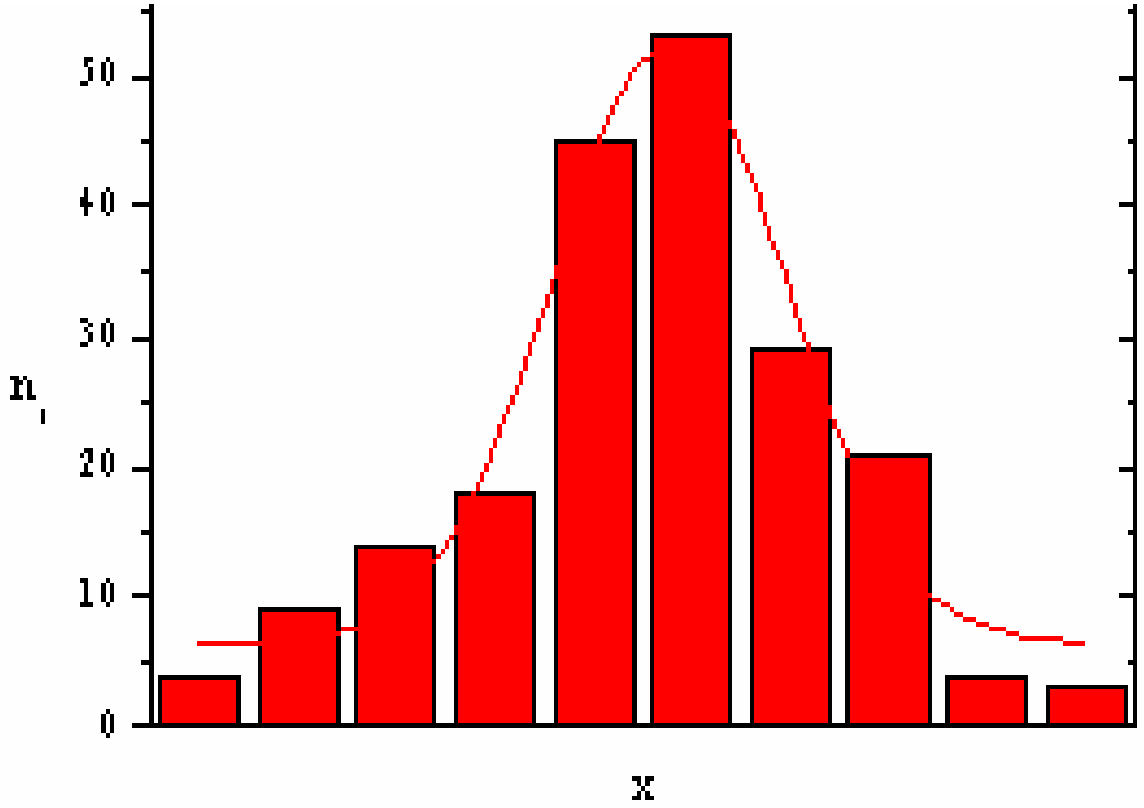
Измерения бывают прямыми (когда объект непосредственно сопоставляется с носителем единицы измерения, например, измерение длины линейкой) и косвенными (когда измеряемая величина рассчитывается из других измеренных величин, например, измерение глубины с помощью эхолота)

Выборка (выборочная совокупность) - конечное число значений одной случайной величины

Генеральная совокупность - полное (бесконечное) множество значений (т.е. она включает все возможные значения измеряемой величины и ничего добавить туда уже нельзя).

Представление результатов измерений

2.5617
2.52109
2.5605
2.77961
….
2.6637
2.59694
2.74882
2.51967



200 измерений
случайной величины

Гистограмма – характеристика выборки,
функция распределения – характеристика ГС

- Под случайной величиной (СВ) понимается величина, которая в результате опыта со случайным исходом принимает то или иное значение, причем заранее, до опыта, неизвестно, какое именно.
- Ω – множество возможных значений величины X.
- Опыт – бросок кубика; случайные величины X – число выпавших очков; Ω = {1, 2, 3, 4, 5, 6}.
- Опыт – работа ЭВМ до первого отказа; случайные величины X – время наработки на отказ; Ω = (0, ∞].

- Случайная величина (СВ) X называется дискретной, если множество Ω – счетное, т.е. его элементы можно расположить в определенном порядке и пронумеровать.
- Случайная величина X называется непрерывной (недискретной), если множество Ω – несчетное.

Выборочные моменты распределения случайной величины

$$m_r' = \frac{1}{n}\sum_n x_i^r \, p(x_i)$$ Момент относительно начала координат

$$m_r = \sum_n (x_i - \bar{x})^r$$ Момент относительно среднего

Центральные моменты высших порядков

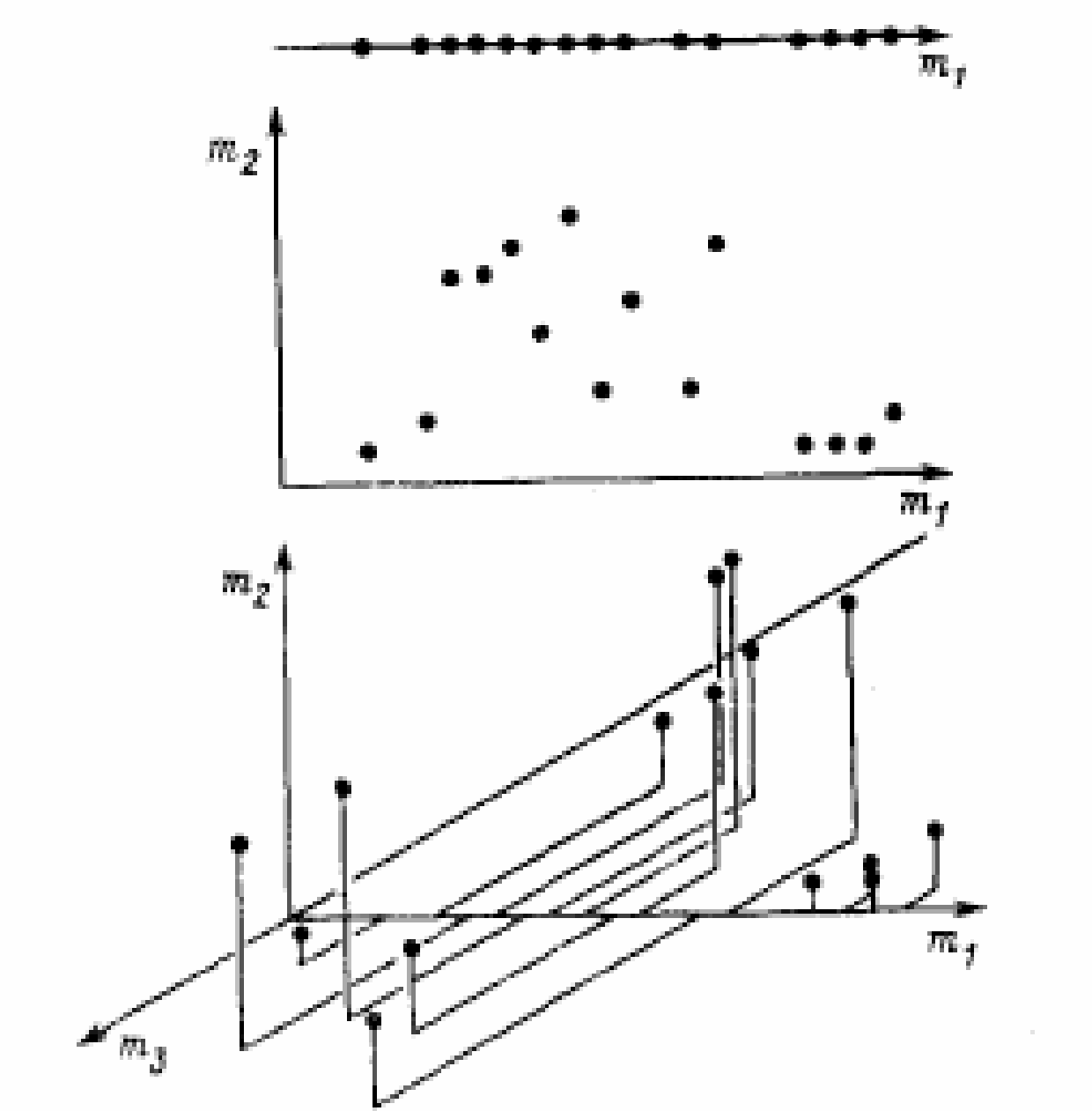
$$m_2 = m_2' - (m_1')^2$$

$$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3$$

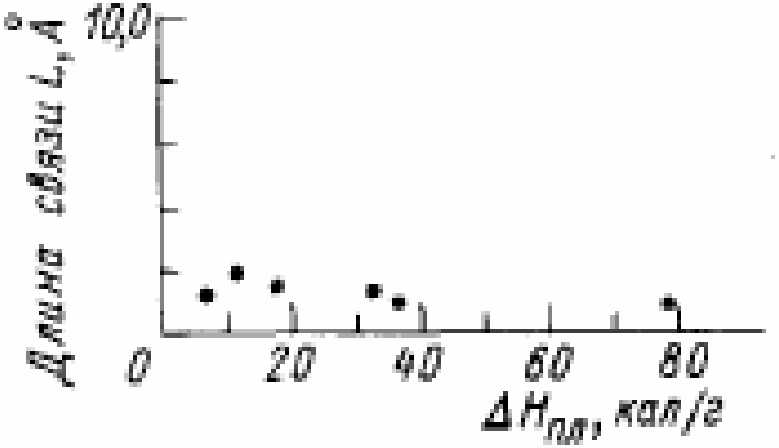
$$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 - 3(m_1')^4$$

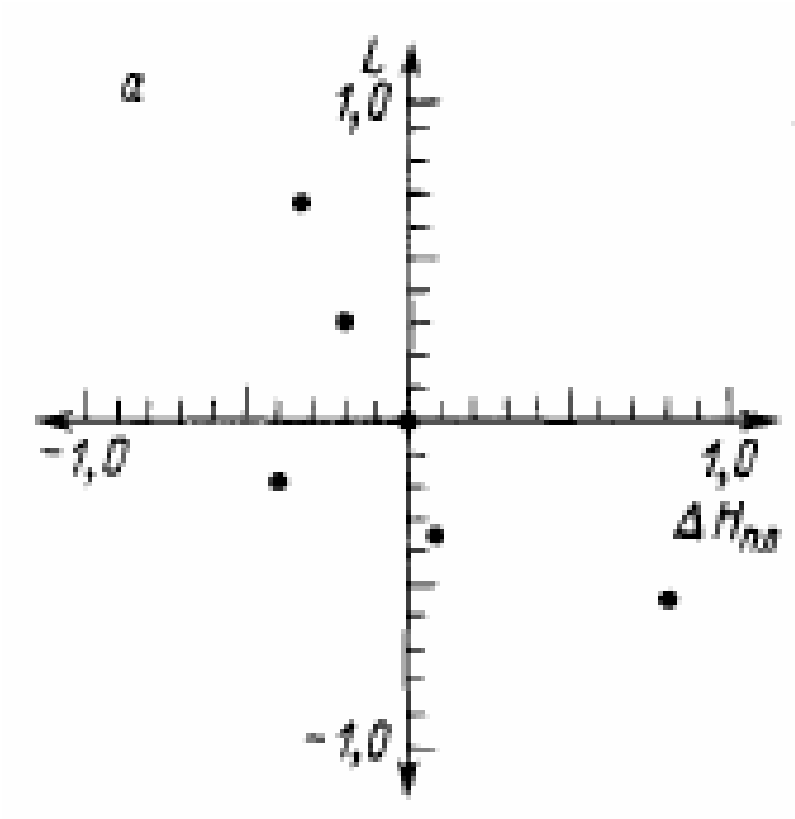
$$\tilde{A} = \frac{m_3}{(m_2)^{3/2}}$$

$$\gamma_2 = \frac{m_4}{(m_2)^2} - 3$$


m_1
m_2
m_1
m_2
m_1
m_3

X =	L, Å	$\Delta H_{пл}$, ккал/г
H_2O	0,96	79,7
SO_2	1,43	32,2
$SiCl_4$	2,03	10,8
AsF_3	1,71	18,9
N_2O	1,13	35,5
BF_3	1,29	7,0





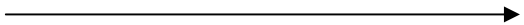
Автомасштабное преобразование

Ковариационная матрица

номер наблюдения

номер признака

x_{11}	x_{12}	…	x_{1M}
x_{21}	x_{22}	…	x_{2M}
…	…	…	…
x_{N1}	x_{N2}	…	x_{NM}



$$\bar{x}_1$$
$$\bar{x}_2$$
$$\dots$$
$$\bar{x}_N$$

$$\operatorname{cov}(x_i, x_j) = \frac{1}{N}\sum_{k=1}^{N}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$D = \begin{pmatrix} s_1^2 & \operatorname{cov}(x_1, x_2) & \dots & \operatorname{cov}(x_1, x_N) \\ \operatorname{cov}(x_2, x_1) & s_2^2 & \dots & \operatorname{cov}(x_2, x_N) \\ \dots & \dots & \dots & \dots \\ \operatorname{cov}(x_N, x_1) & \operatorname{cov}(x_N, x_2) & \dots & s_N^2 \end{pmatrix}$$

Из ковариаций определяем корреляции

$$-1 \le r_{ij} \equiv r(x_i, x_j) = \frac{\mathrm{cov}(x_i, x_j)}{\sqrt{s^2(x_i)s^2(x_j)}} \le 1$$

Вычислите ковариационную матрицу и коэффициент корреляции для измерений

$$x = \begin{pmatrix} 1 & 1.5 \\ 2 & 2.3 \end{pmatrix}$$

Статистичні розподіли випадкової величини. Біноміальний розподіл. Розподіли неперервних величин: рівномірний, Гауса, Лапласа, Пуассона, χ^2. Центральна гранична теорема.

Функцией распределения $F(x)$ случайной величины X называется вероятность того, что она примет значение меньшее, чем аргумент функции x:

$$F(x) = p\{X < x\}$$

Свойства функции распределения

1. $F(-\infty) = 0$.
2. $F(+\infty) = 1$.
3. $F(x1) \le F(x2)$, при $x1 < x2$.
4. $p(x1 \le X < x2) = F(x1) - F(x2)$.

Плотность вероятности (плотность распределения) – нормирована!

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

Биномиальное распределение и распределение Пуассона

2 результата (успех-неуспех)

θ 1-θ

Функция плотности – вероятность получить n успехов в N испытаниях

$$P(n) = \frac{N!}{(N-n)!n!} \cdot \theta^n (1-\theta)^{(N-n)}$$

Распределение редких событий

$$N \longrightarrow \infty, \theta \longrightarrow 0, N\theta = \lambda$$

$$P(n) = \frac{\lambda^n e^{-\lambda}}{n!}$$

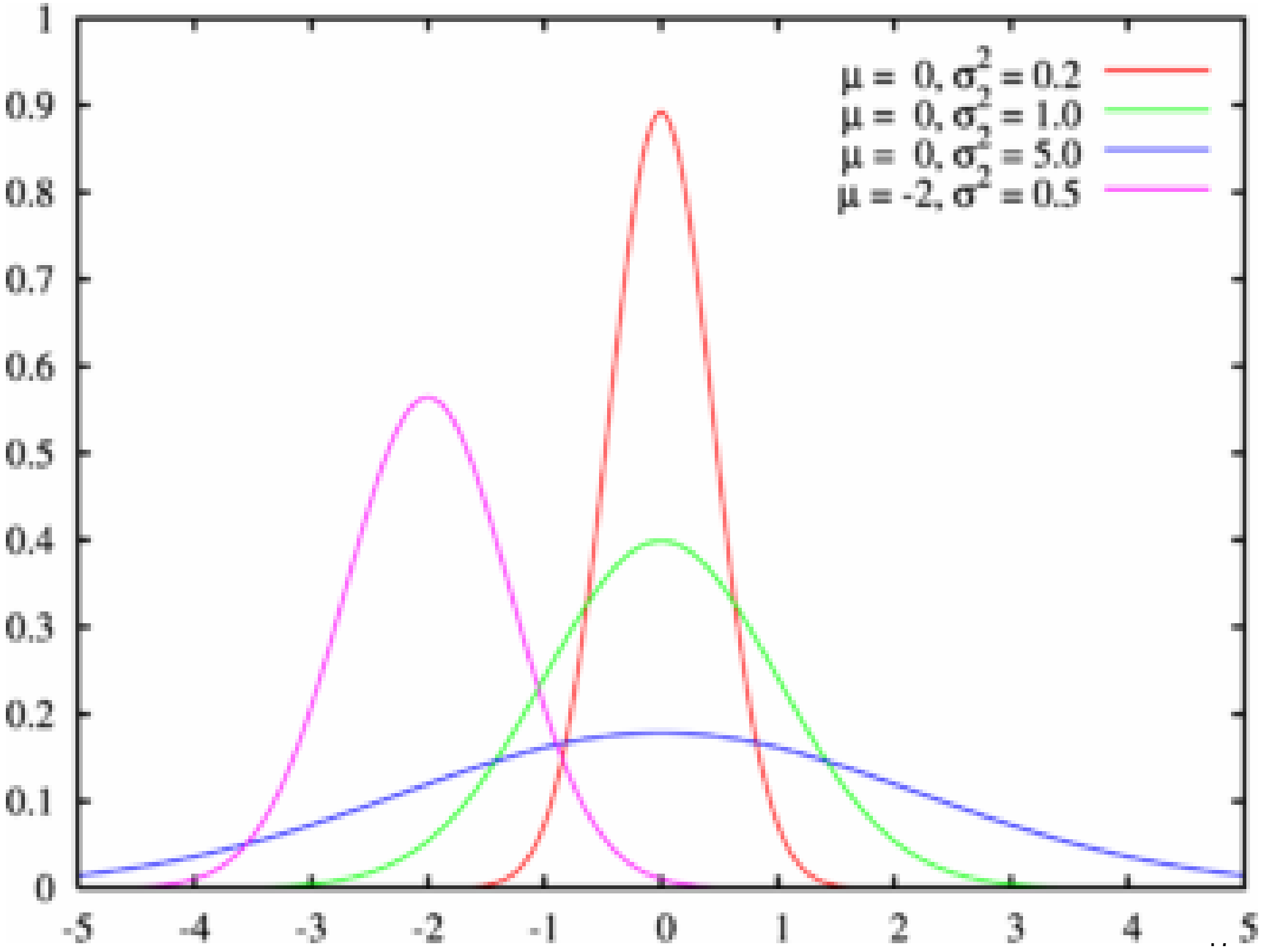
5 орлов при 5 бросаниях монеты
100 орлов при 100 бросаниях монеты

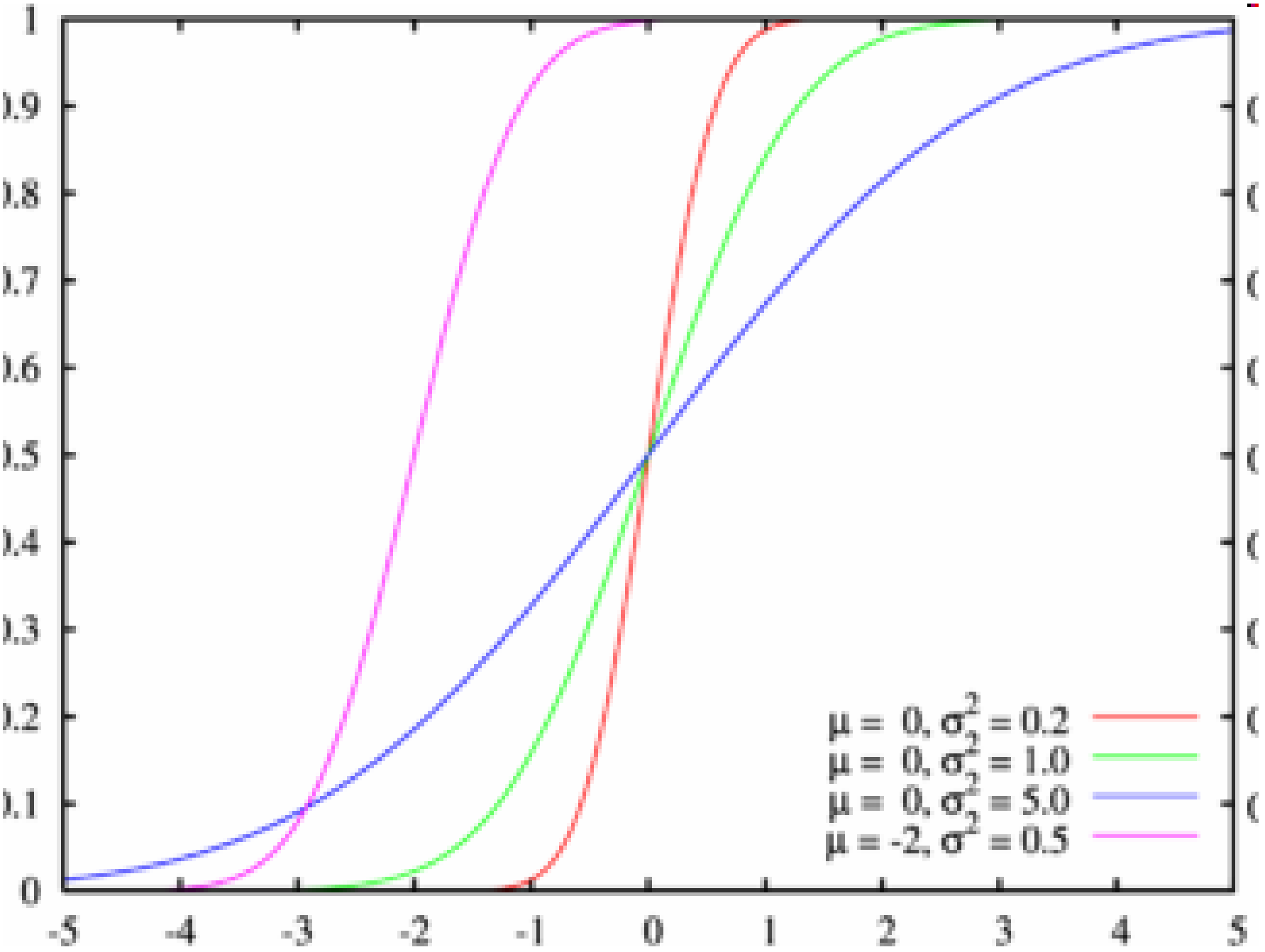
«Все верят в универсальность нормального распределения: физики верят, потому что думают, что математики доказали его логическую необходимость, а математики верят, так как считают, что физики проверили это лабораторными экспериментами»

А. Пуанкаре

Нормальное (Гаусса) распределение

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma} \right]^2 \right)$$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
μ = 0, σ² = 0.2
μ = 0, σ² = 1.0
μ = 0, σ² = 5.0
μ = -2, σ² = 0.5


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
μ = 0, σ² = 0.2
μ = 0, σ² = 1.0
μ = 0, σ² = 5.0
μ = -2, σ² = 0.5

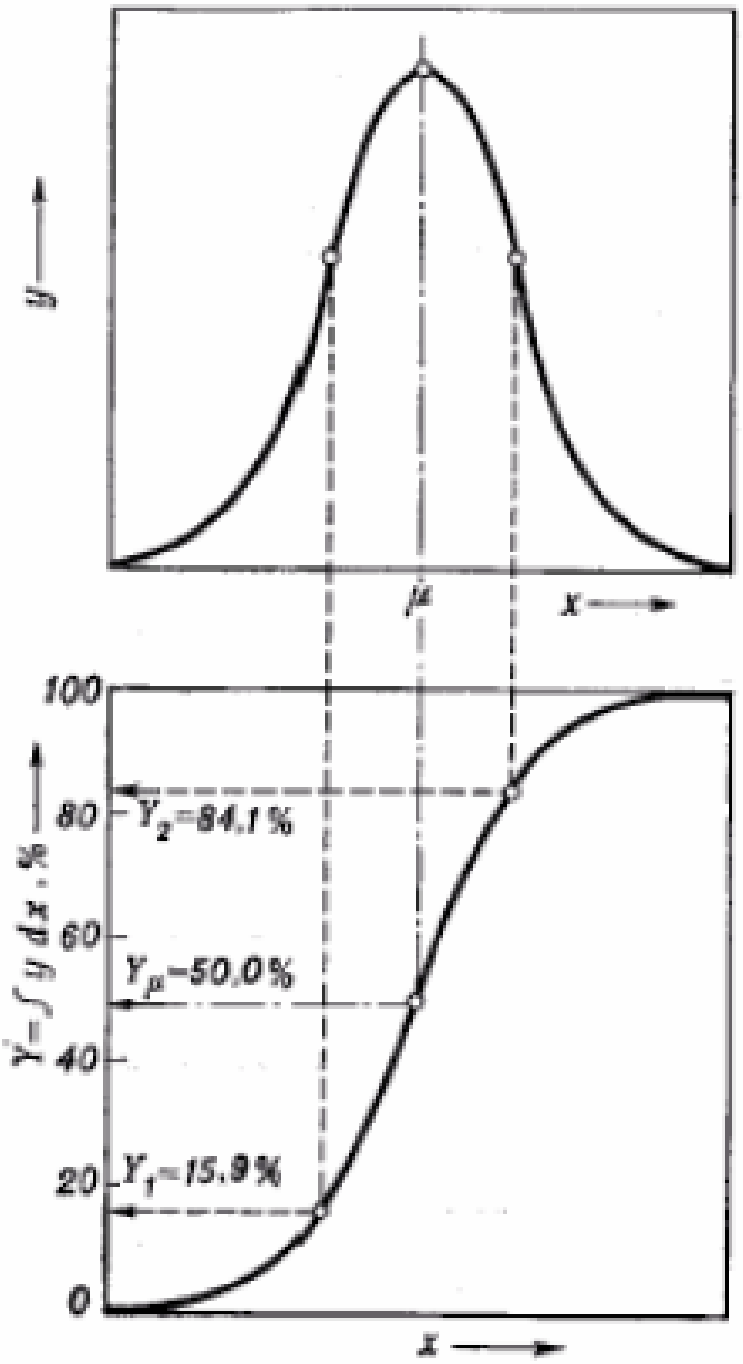
Некоторые свойства нормального распределения

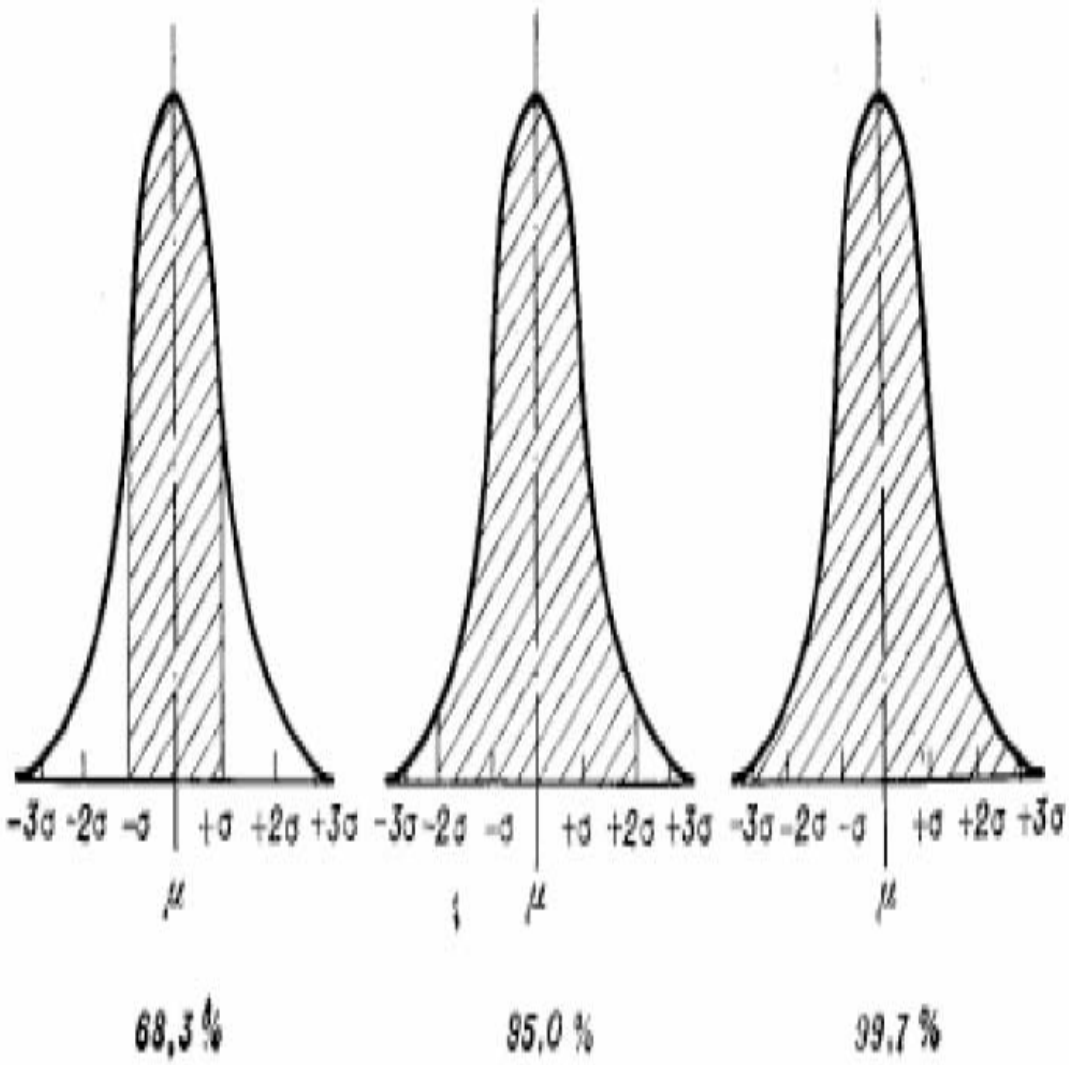
Если случайные величины $X1$ и $X2$ независимы и имеют нормальное распределение с математическими ожиданиями $\mu 1$ и $\mu 2$ и дисперсиями $(\sigma 1)^2$ и $(\sigma 2)^2$ соответственно, то $X1 + X2$ также имеет нормальное распределение с математическим ожиданием $\mu 1 + \mu 2$ и дисперсией $[(\sigma 1)^2 + (\sigma 2)^2]$

Вероятность попадания случайного измерения в интервал [a;b]

$$p(a \le x \le b) = \int_a^b f(x)dx$$

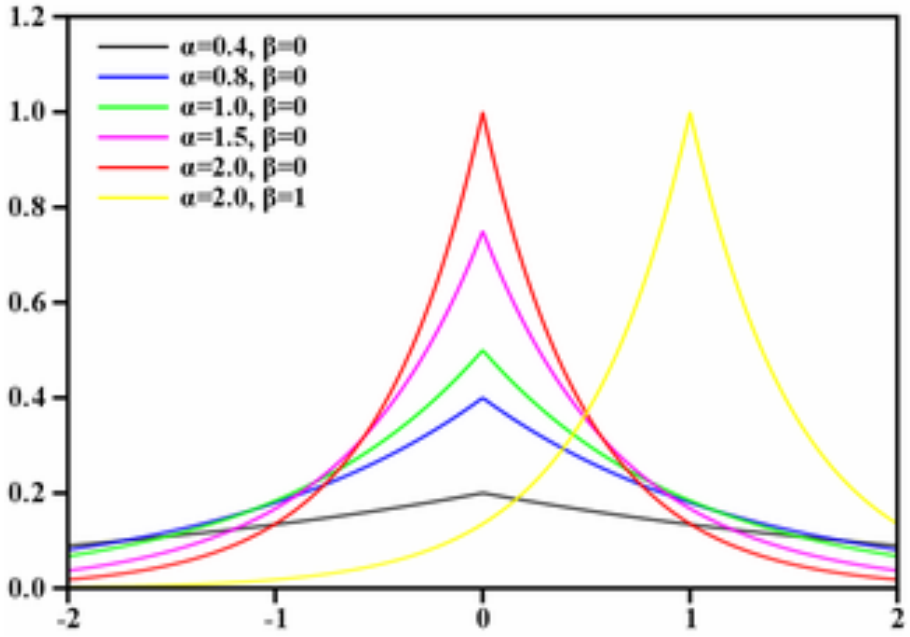
Парадокс нулевой вероятности

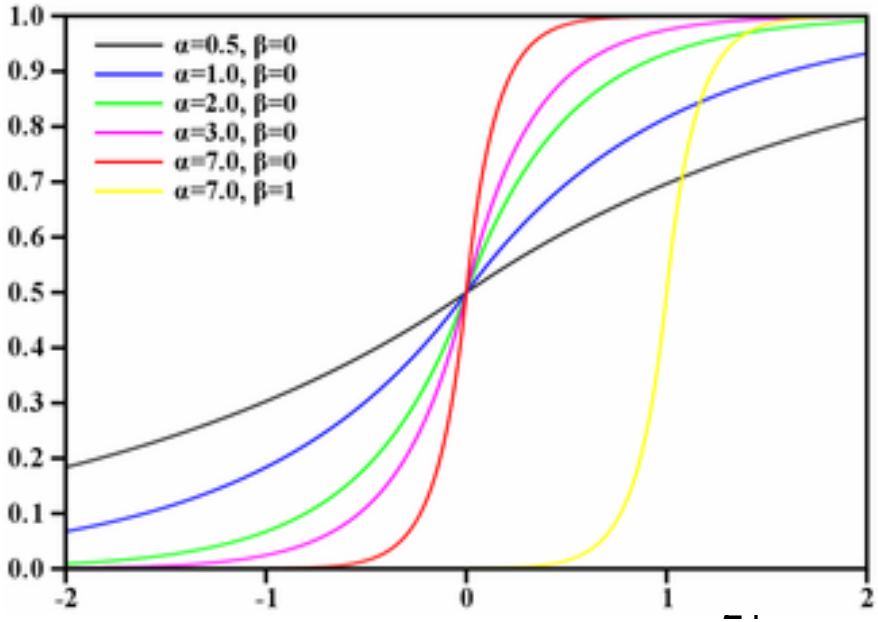

y
μ
x
100
80
60
40
20
0
Y_2=84.1%
Y_μ=50.0%
Y_1=15.9%
$Y=\int y\,dx$, %
x


-3σ -2σ -σ +σ +2σ +3σ
μ
68,3 %
95,0 %
99,7 %

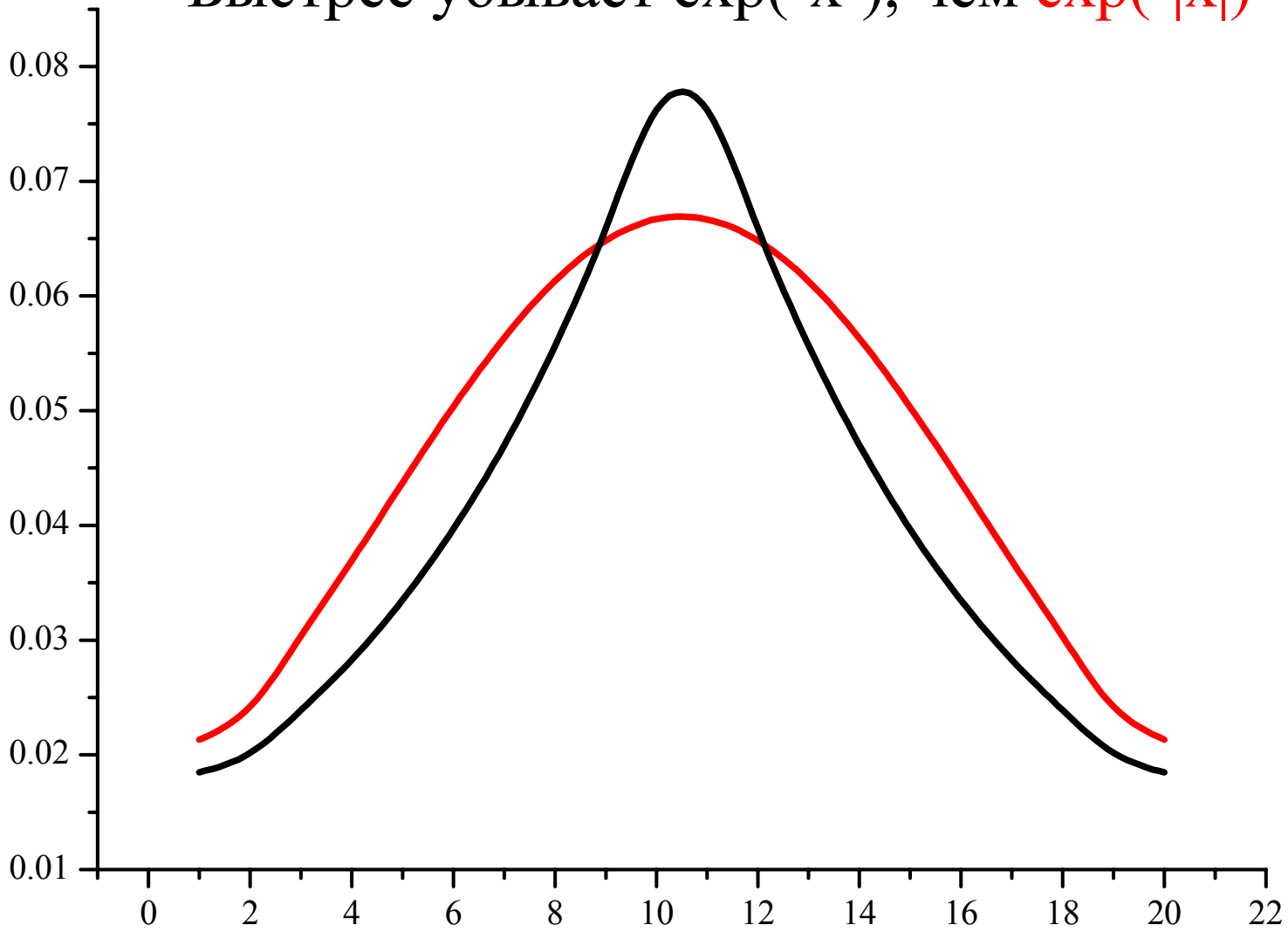
$$f(x) = \frac{\alpha}{2} e^{-\alpha|x-\beta|}$$

Функция Лапласа



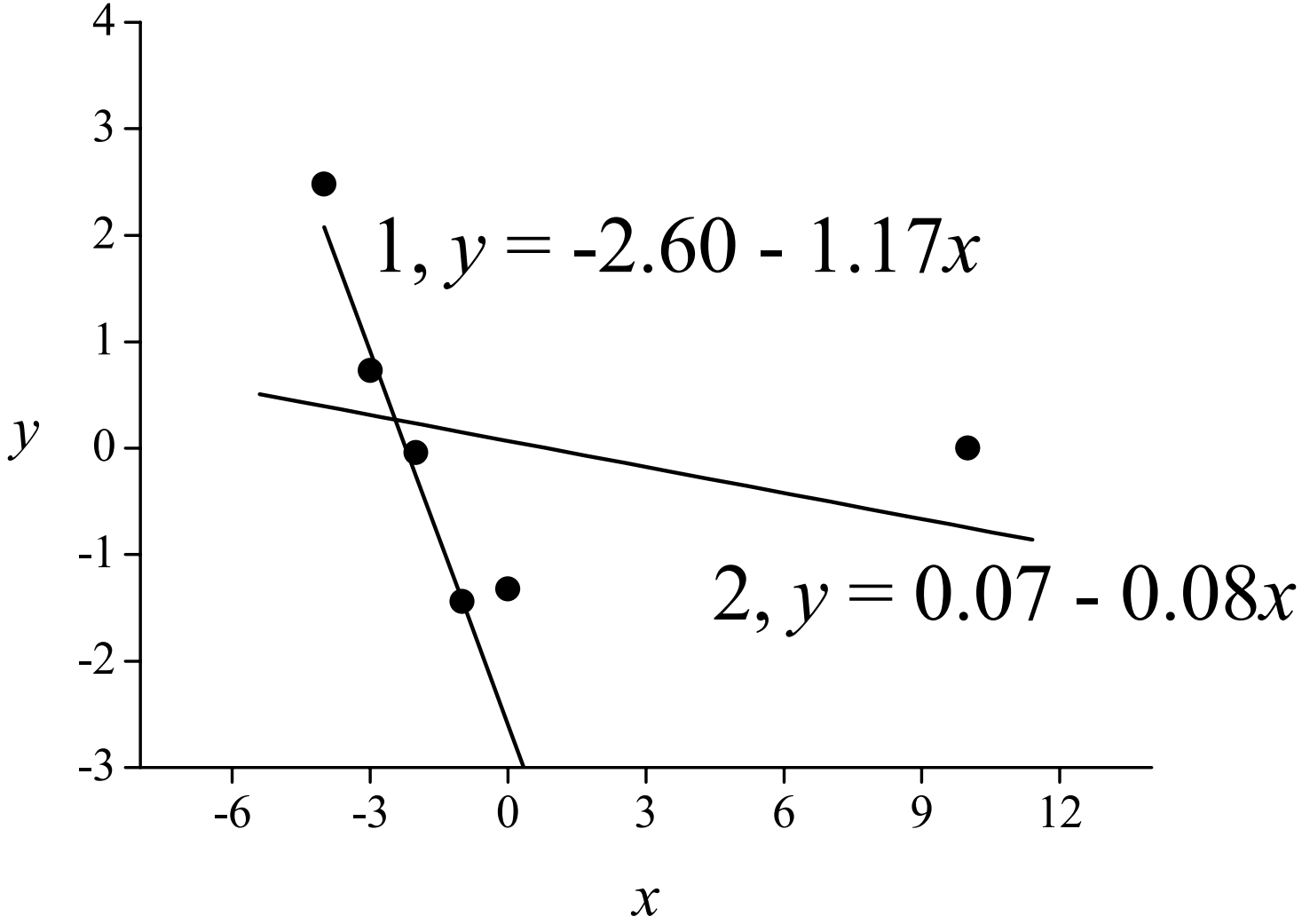


Быстрее убывает $\exp(-x^2)$, чем $\exp(-|x|)$



Робастность

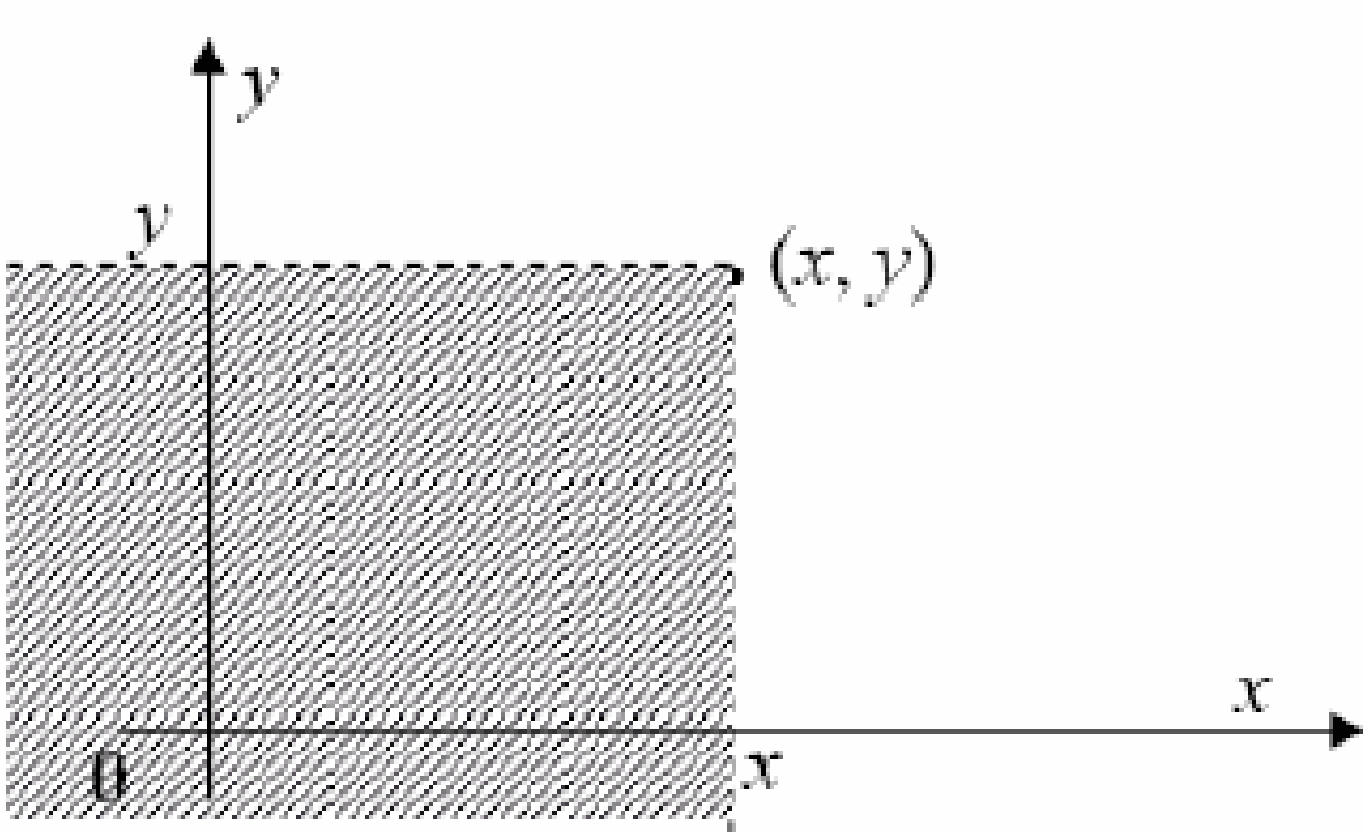
Робастность в статистике - нечувствительность к различным отклонениям и неоднородностям в выборке, связанным с теми или иными, в общем случае неизвестными, причинами. Это могут быть ошибки детектора, регистрирующего наблюдения, чьи-то добросовестные или не очень попытки «подогнать» выборку до того, как она попадёт к статистику, ошибки оформления, вкравшиеся опечатки и многое другое.

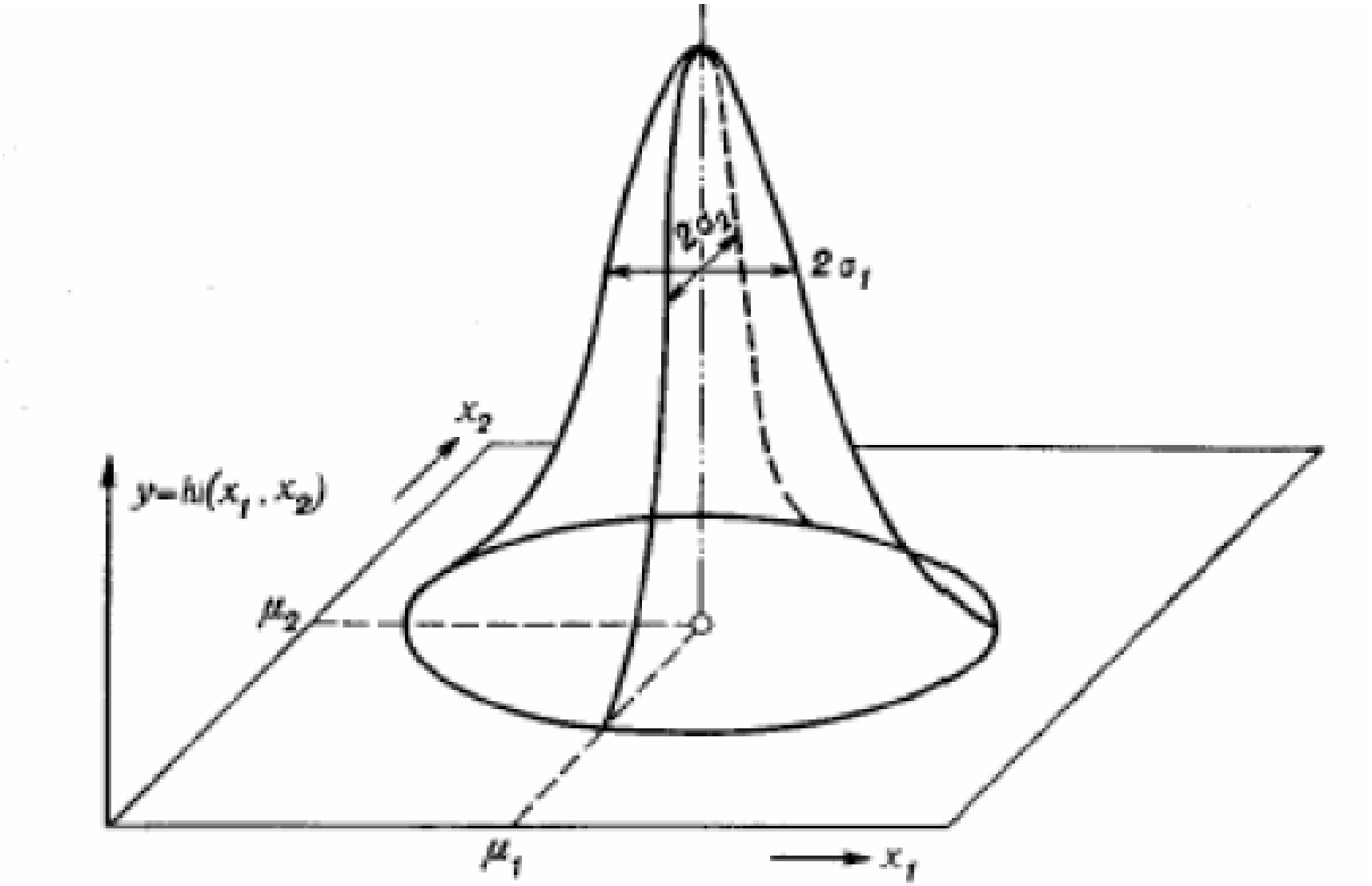

4
3
2
1
0
-1
-2
-3
y
-6
-3
0
3
6
9
12
x
1, y = -2.60 - 1.17x
2, y = 0.07 - 0.08x

Центральная предельная теорема

сумма достаточно большого количества слабо зависимых случайных величин, имеющих примерно одинаковые масштабы (ни одно из слагаемых не доминирует, не вносит в сумму определяющего вклада), имеет распределение, близкое к нормальному.

Двумерная случайная величина (X, Y) – совокупность двух одномерных случайных величин, которые принимают значения в результате проведения одного и того же опыта.




$2\sigma_2$
$2\sigma_1$
x_2
$y = h(x_1, x_2)$
μ_2
μ_1
x_1

Известно, что результаты титрования распределены нормально со средним m=5 ml и стандартным отклонением s=0.5 ml.
Нужно оценить вероятность того, что при выполнении независимого эксперимента по титрованию экспериментатор получит такие объемы титранта:

a) от 4.5 до 5 ml;
b) от 4.5 до 5.5 ml;
c) от 3.5 до 4 ml;
d) от 3 до 6 ml;
e) больше 6 ml;
f) меньше 3.5 ml.

Таблиця Д2. Значення функції стандар

u	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944

С. 32

2.0	0.9772	0.9778	0.9783	0.9788
2.1	0.9821	0.9826	0.9830	0.9834
2.2	0.9861	0.9864	0.9868	0.9871
2.3	0.9893	0.9896	0.9898	0.9901
2.4	0.9918	0.9920	0.9922	0.9925
2.5	0.9938	0.9940	0.9941	0.9943
2.6	0.9953	0.9955	0.9956	0.9957
2.7	0.9965	0.9966	0.9967	0.9968
2.8	0.9974	0.9975	0.9976	0.9977
2.9	0.9981	0.9982	0.9982	0.9983
3.0	0.9987	0.9987	0.9987	0.9988

C. 32

Расчет вероятности того, что при выполнении независимого эксперимента по титрованию экспериментатор получит объемы титранта, заключенные в пределах 4.95-5.06 ml сдайте с указанием даты, группы и фамилии. При этом распределение результатов анализа нормальное со средним 4.97 ml и дисперсией 0.0225 ml^2.

Необходимые и достаточные данные

u	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5832	0.5871	0.5910	0.5948
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389

Метод максимуму правдоподібності. Функція правдоподібності. Правдоподібні оцінки параметрів генеральної сукупності при нормальному та Лапласівському розподілах похибок.

Метод максимального правдоподобия или метод наибольшего правдоподобия (ММП, ML, MLE — Maximum Likelihood Estimation) в математической статистике — это метод оценивания неизвестного параметра путём максимизации функции правдоподобия. Основан на предположении о том, что вся информация о статистической выборке содержится в функции правдоподобия.

Метод максимума правдоподобия

метод оценивания неизвестного параметра путём максимизации функции правдоподобия

1. Измерения – независимые случайные величины
2. ГС характеризуется определенной (одной и той же) функцией распределения

Измерения проводятся N раз, получаем набор (x1, x2, …, xN)

$$P(x1, x2, \ldots, xN) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma} \right]^2 \right) = L \longrightarrow \max$$

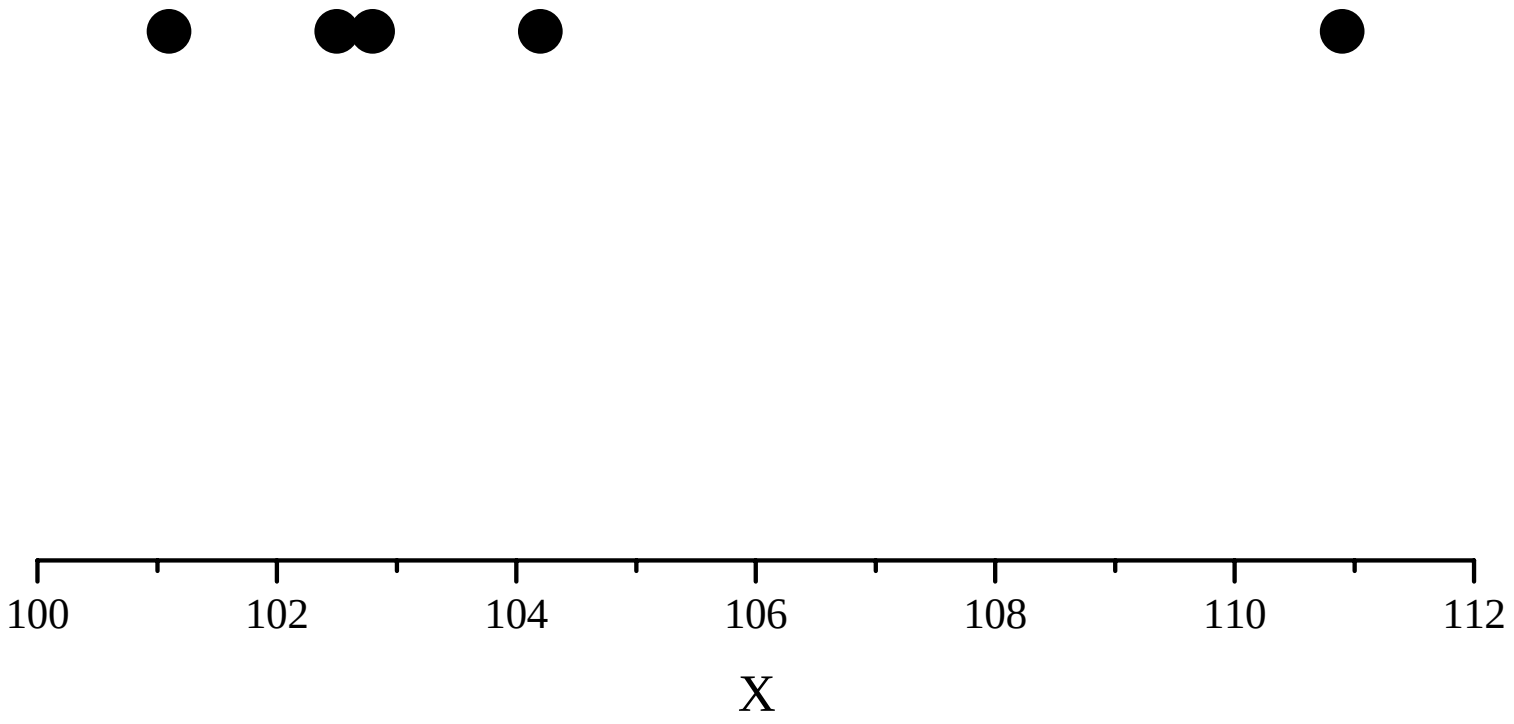
$$L = \prod_i \left(\frac{1}{\sigma^2} \right)^{1/2} \exp\left(-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma} \right]^2 \right) = \left(\frac{1}{\sigma^2} \right)^{N/2} \exp\left(\sum_i -\frac{1}{2}\left[\frac{x_i - \mu}{\sigma} \right]^2 \right)$$

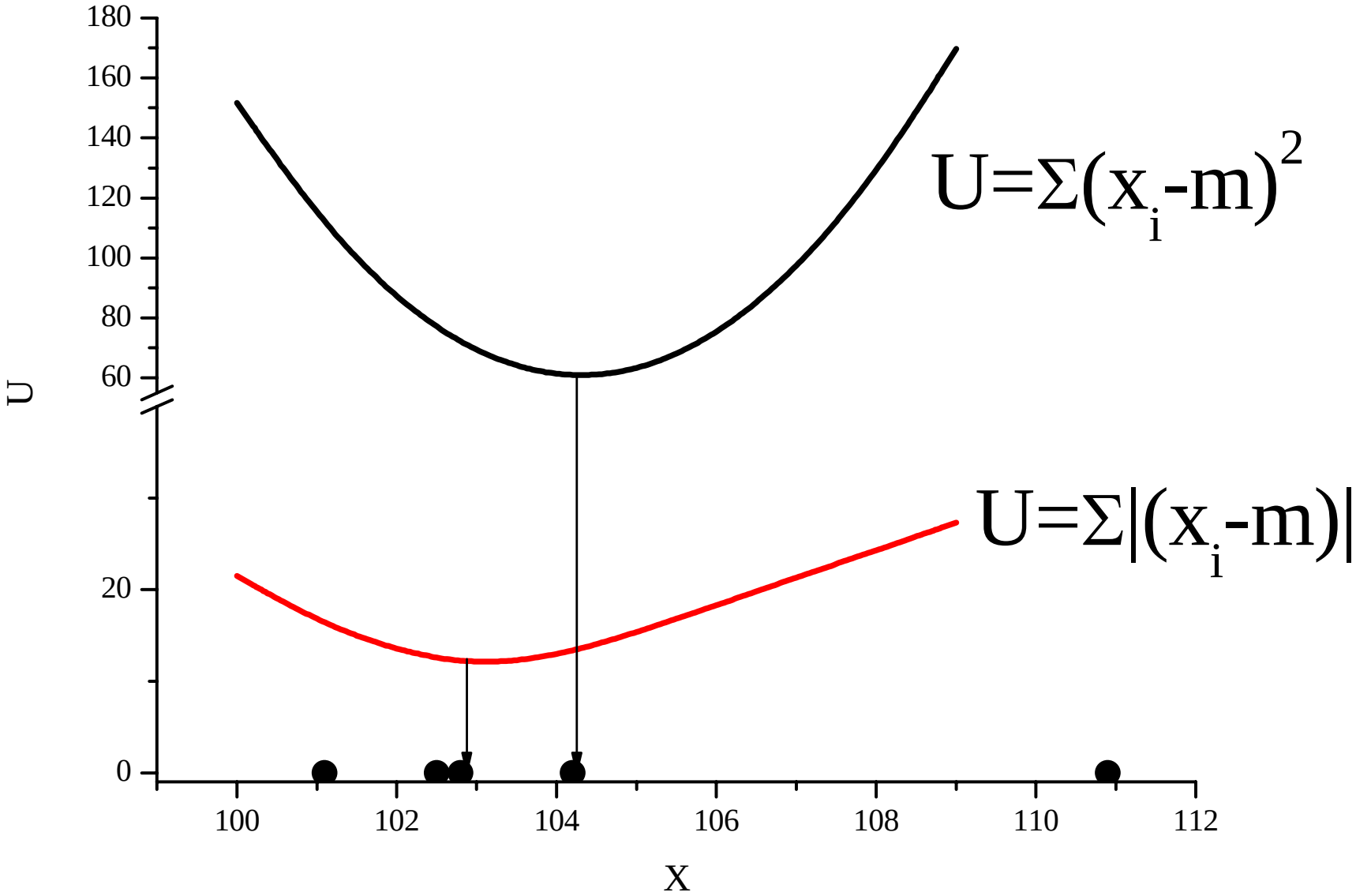
$$\ln L = \ln\left(\frac{1}{\sigma^2} \right)^{N/2} - \sum_i \frac{1}{2}\left[\frac{x_i - \mu}{\sigma} \right]^2 = -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2$$

$$\left(\frac{\partial \ln L}{\partial \mu} \right) = 0$$

$$\left(\frac{\partial \ln L}{\partial \sigma^2} \right) = 0$$

Случайная величина x = 101.1; 102.5; 102.8; 104.2; 110.9




180
160
140
120
100
80
60
20
0
U
100
102
104
106
108
110
112
X
U=Σ(x_i-m)^2
U=Σ|(x_i-m)|

Перевірка статистичних гіпотез. Задача перевірка статистичних гіпотез. Схема перевірки гіпотези. Помилки I та II родів. Потужність критеріїв. Перевірка гіпотез про функції розподілу. Критерій χ^2, графічні способи перевірки гіпотез про функції розподілу.

Статистическая гипотеза – любое предположение относительно функции частот наблюдаемых случайных переменных.

$H_м$ – на Марсе есть жизнь.
$H_Р$ – орбита вновь открытой кометы параболическая.
…
H_1 $p=1/2$
H_2 $1/3<p<2/3$

Если статистическая гипотеза H полностью определяет функцию частот наблюдаемой случайной переменной как однозначную функцию аргумента, то она называется простой гипотезой.

1000 подбрасываний монеты, проверка на симметричность.
511 «орлов» - монета симметрична? А 897 «орлов»? В чем разница?

Общая схема проверки гипотез

1. Выбрать уровень значимости α	α=0.05
2. Описать статистическую модель	Число выпадений герба X подчиняется биномиальному распределению с n=1000 и p
3. Сформулировать нулевую и альтернативную гипотезы	H_0: p=1/2 H_1: p<>1/2

4. Выбрать критериальную статистику, поведение которой известно

$T=(X-np) / (npq)^{0.5}$

5. Определить критическую область. Вероятность попадания значения критерия в КО при условии, что H_0 справедлива, равна α.

$|T|>1.96$

6. Вычислить значение статистического критерия

7. Сделать выводы

Ошибки первого рода (*type I errors, α errors, false positives*) и **ошибки второго рода** (*type II errors, β errors, false negatives*).

Ошибку первого рода часто называют ложной тревогой, ложным срабатыванием или ложноположительным срабатыванием — например, анализ крови показал наличие заболевания, хотя на самом деле человек здоров, или металлодетектор выдал сигнал тревоги, сработав на металлическую пряжку ремня.

Ошибку второго рода иногда называют пропуском события или ложноотрицательным срабатыванием — человек болен, но анализ крови этого не показал, или у пассажира имеется холодное оружие, но рамка металлодетектора его не обнаружила (например, из-за того, что чувствительность рамки отрегулирована на обнаружение только очень массивных металлических предметов).

Возможные ситуации при проверке H_0:

1) H_0 справедлива, и критерий ее допускает
2) H_0 справедлива, и критерий ее отвергает
3) H_1 справедлива, и критерий отвергает H_0
4) H_1 справедлива, и критерий допускает H_0

	Верная гипотеза		
Принятая гипотеза		H_0	H_1
	H_0	H_0	II
	H_1	I	H_1

Функция мощности статистического критерия определяется как вероятность отвергнуть нулевую (основную) гипотезу при заданном распределении наблюдений P. Функция мощности является функцией от распределения P наблюдаемых случайных величин.

В случае, если P соответствует нулевой гипотезе, значение функции мощности называется вероятностью ошибки первого рода. Если P соответствует альтернативной гипотезе, то значение функции мощности называют просто мощностью. Для критерия, основанного на выборке фиксированного объема, мощность равна единице минус вероятность ошибки второго рода.

Мощность равна вероятности того, что критерий распознает ложность нулевой гипотезы при истинности альтернативной.

Таблиця Д3

Квантілі розподілу χ^2

f; P	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99
1	0.4549	0.7083	1.0742	1.6424	2.7055	3.8415	5.0239	6.6349
2	1.3863	1.8326	2.4079	3.2189	4.6052	5.9915	7.3778	9.2103
3	2.366	2.9462	3.6649	4.6416	6.2514	7.8147	9.3484	11.3449
4	3.3567	4.0446	4.8784	5.9886	7.7794	9.4877	11.1433	13.2767
5	4.3515	5.1319	6.0644	7.2893	9.2364	11.0705	12.8325	15.0863
6	5.3481	6.2108	7.2311	8.5581	10.6446	12.5916	14.4494	16.8119
7	6.3458	7.2832	8.3834	9.8032	12.017	14.0671	16.0128	18.4753
8	7.3441	8.3505	9.5245	11.0301	13.3616	15.5073	17.5345	20.0902

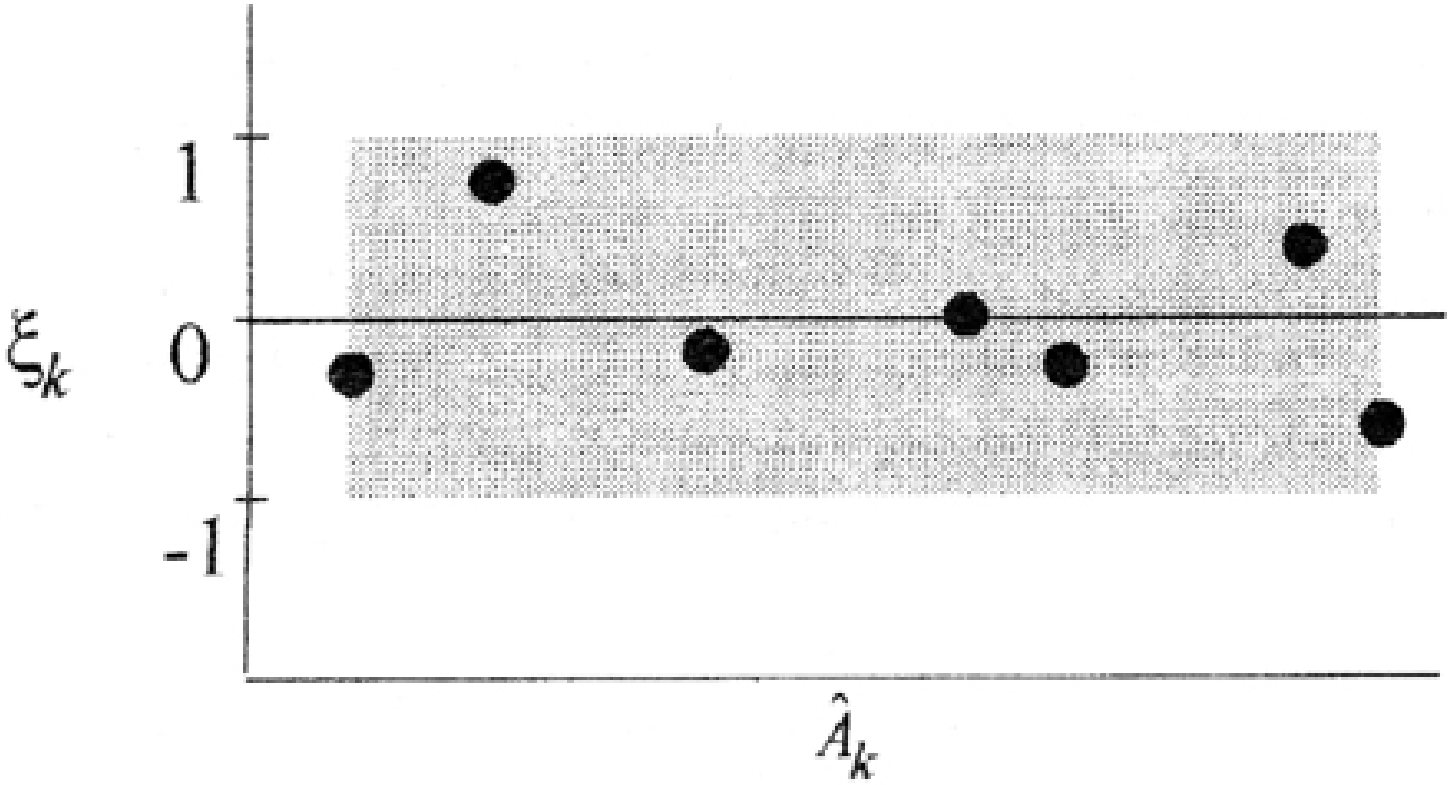
С. 33

1. α=5%.
2. Статистическая модель – ε распределены нормально.
3. H_0: E(ε)=0.
4. Статистика $\chi^2_{\text{эксп}} = \sum_{i=1}^{N} \left(\frac{y_i - \hat{y}_i}{s_{y_i}}. \right)^2$
5-7. $\chi^2_{\text{эксп}} < \chi^2_{f,\alpha} \Rightarrow$ Принимаем H_0

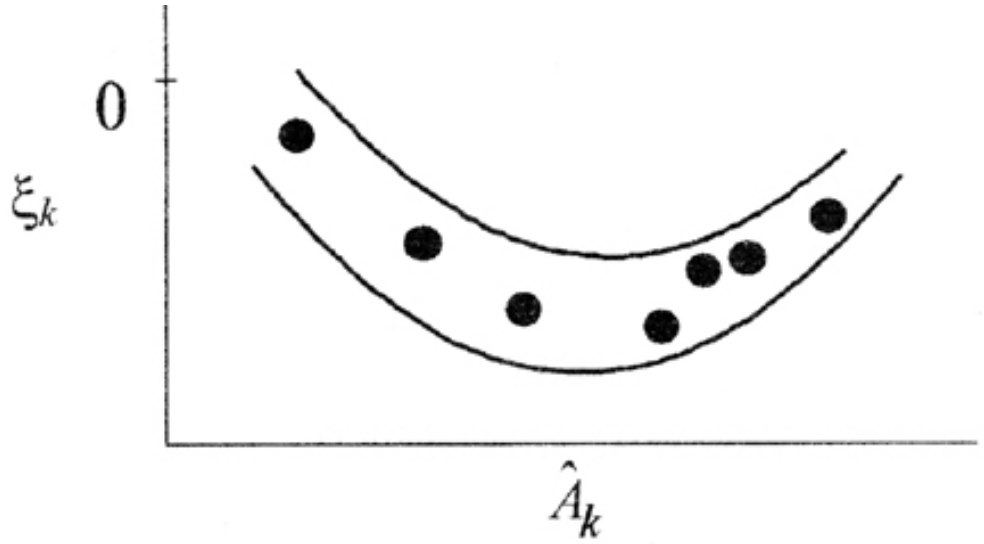
Другие глобальные критерии

$$\sum_i \xi_i = \sum_{i=1}^{N} \frac{y_i - \hat{y}_i}{s_{y_i}} \approx 0 \qquad \sum_{i=1}^{N} \left| \xi_i \right| \approx 0.8N$$

Локальные критерии адекватности. Исследование невязок.



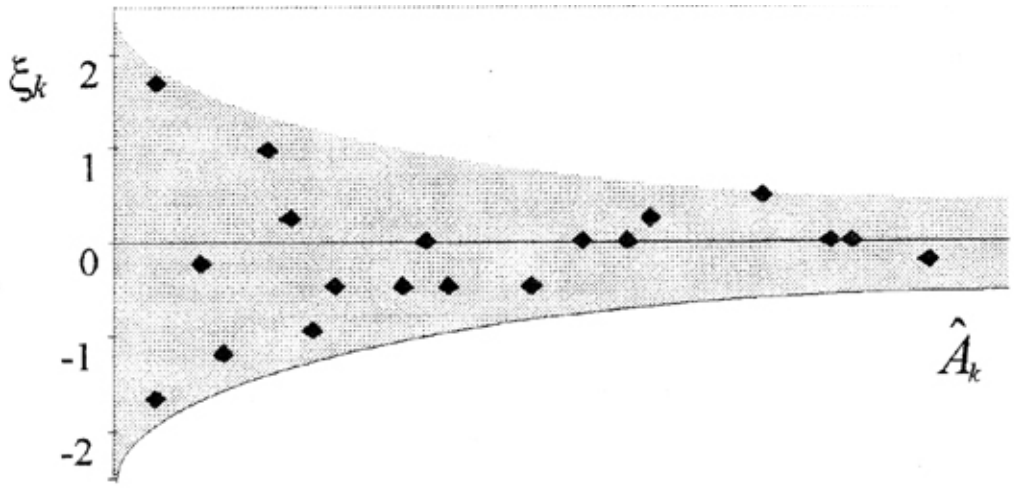
Адекватная модель



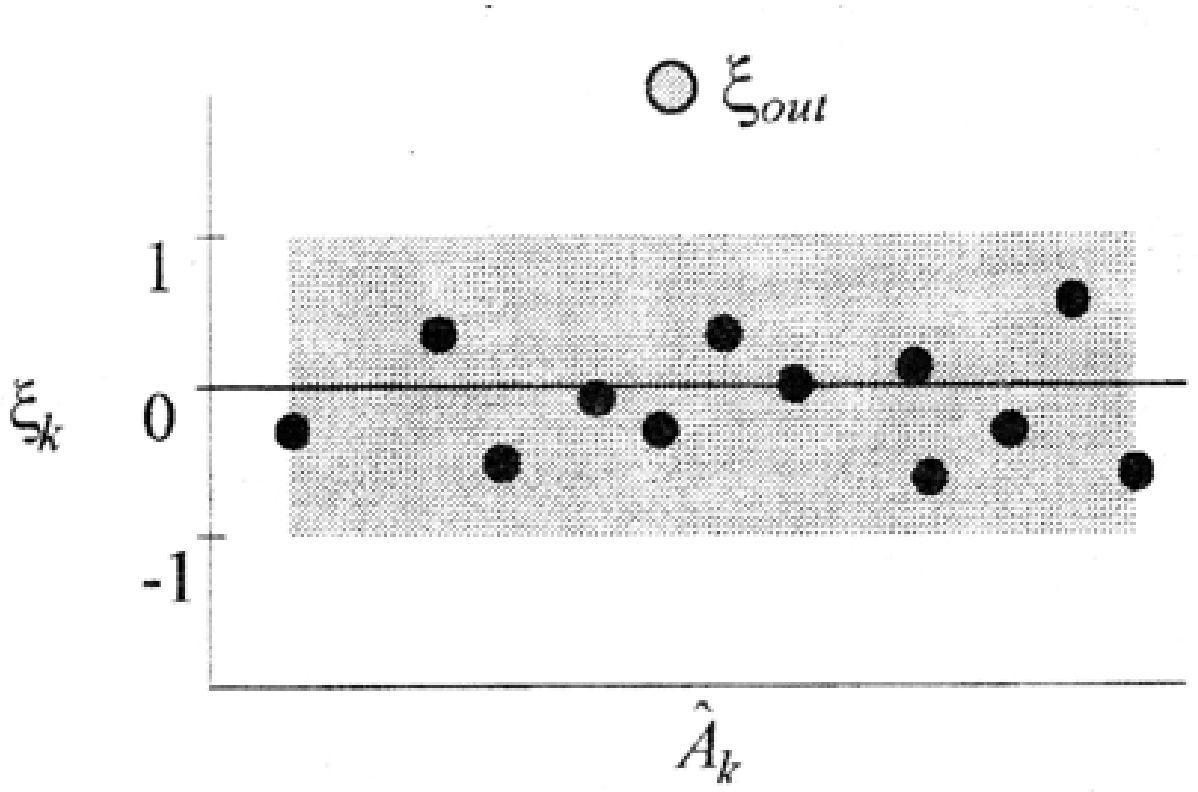
Неполная

по параметрам модель

Модель с неверно назначенными весами

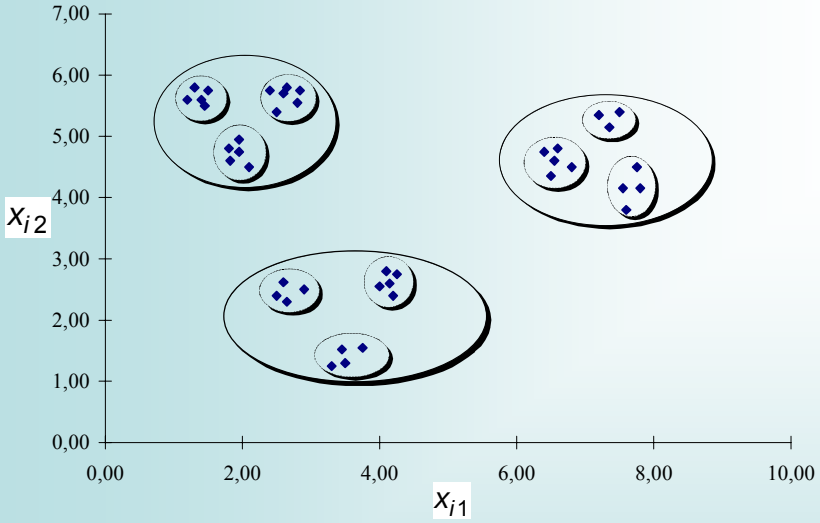


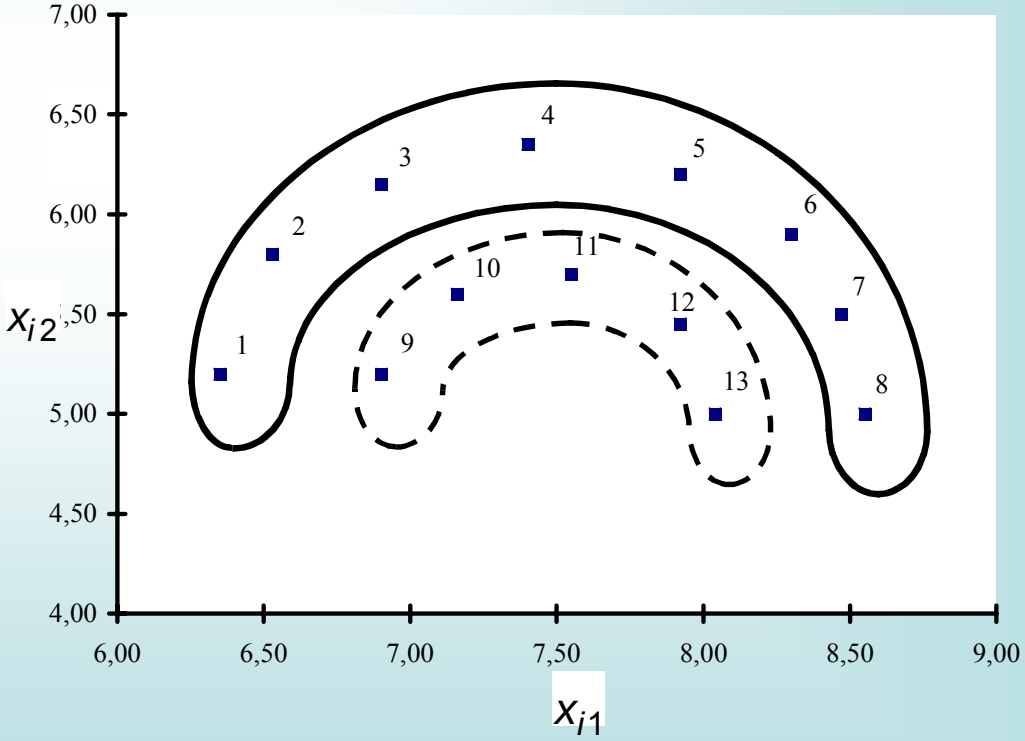
Визуализация резко выпадающего значения



МЕТОДЫ ХЕМОМЕТРИКИ.

1. КЛАССИФИКАЦИЯ И КЛАСТЕРНЫЙ АНАЛИЗ



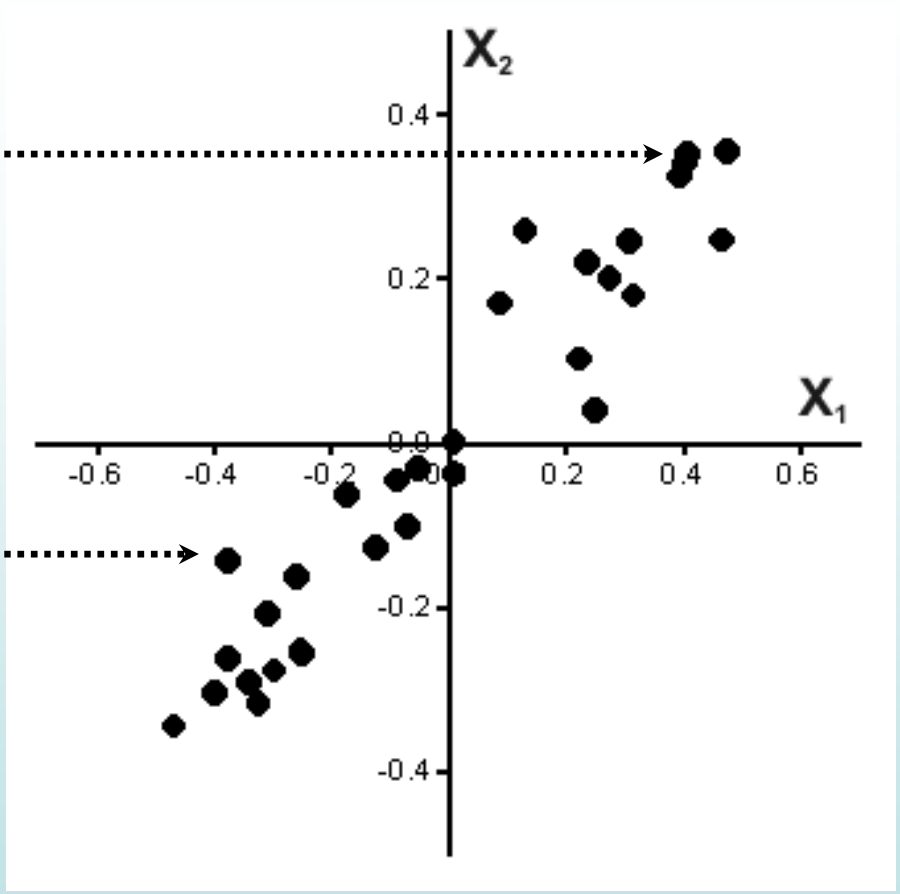


2. ФАКТОРНЫЙ АНАЛИЗ

Персона №	Пол M/F	Регион S/N	Рост см	Вес кг	Волосы S/L	Обувь размер	Возраст лет	Доход K€/год	Пиво л/год	Вино л/год	Сила баллы	IQ баллы
1	M	N	198	92	S	48	48	45	420	115	98	100
2	M	N	184	84	S	44	33	33	350	102	92	130
3	M	N	183	83	S	44	37	34	320	98	91	127
4	M	N	182	80	S	42	35	30	398	65	85	140
5	M	N	180	80	S	43	36	30	388	63	84	129
6	M	N	183	81	S	42	37	35	345	45	90	105
7	M	N	180	82	S	44	43	37	355	82	88	109
8	M	N	180	81	S	44	46	42	362	90	86	113
9	M	S	185	82	S	45	26	16	295	180	92	109
10	M	S	187	84	S	46	27	17	299	178	95	119
11	M	S	177	65	S	41	26	18	209	160	86	120
12	M	S	180	72	S	43	33	19	236	175	85	115
13	M	S	181	75	S	43	42	31	198	161	83	105
14	M	S	176	68	S	42	50	36	195	177	82	96
15	M	S	175	67	L	42	55	38	185	187	80	105
16	M	S	178	75	S	42	30	24	203	208	81	118
17	F	N	166	47	S	36	32	28	270	78	75	112
18	F	N	170	60	L	38	23	20	312	99	81	110
19	F	N	172	64	L	39	24	22	308	91	82	102
20	F	N	169	51	L	36	24	23	250	89	78	98
21	F	N	168	52	L	37	27	24	260	86	78	100
22	F	N	157	47	L	36	32	32	235	92	70	127
23	F	N	164	50	L	38	41	34	255	134	76	101
24	F	N	162	49	L	37	40	34	265	124	75	108
25	F	S	168	50	L	37	49	34	170	162	76	135
26	F	S	166	49	L	36	21	14	150	245	75	123
27	F	S	158	46	L	34	30	18	120	120	70	119
28	F	S	163	50	L	36	18	11	143	136	75	102
29	F	S	162	50	L	36	20	12	133	146	74	132
30	F	S	165	51	L	36	36	26	121	129	76	126
31	F	S	161	48	L	35	41	32	116	196	75	120
32	F	S	160	48	L	35	40	31	118	198	74	129

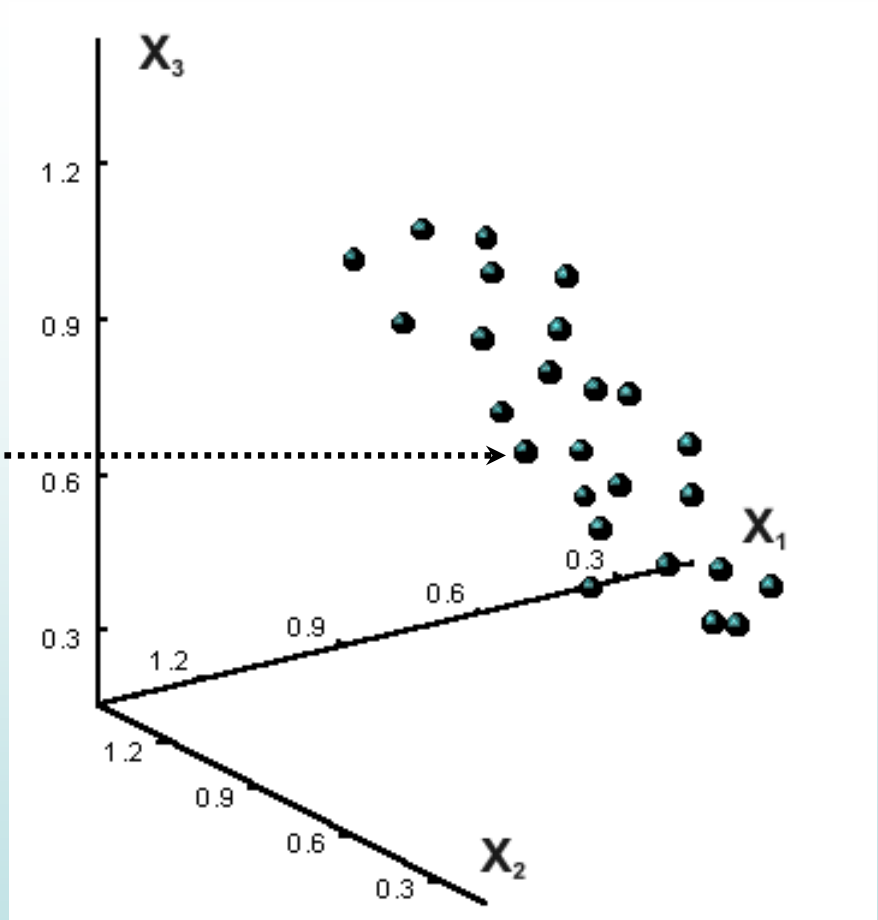
Графическое представление 2D-данных

	X_1	X_2
1	0.407	0.353
2	0.475	0.355
3	-0.088	-0.045
4	0.394	0.325
5	0.274	0.202
6	0.131	0.258
7	-0.053	-0.031
8	-0.124	-0.128
9	-0.469	-0.344
10	0.088	0.171
11	-0.261	-0.162
12	0.401	0.341
13	-0.376	-0.143
14	-0.251	-0.255

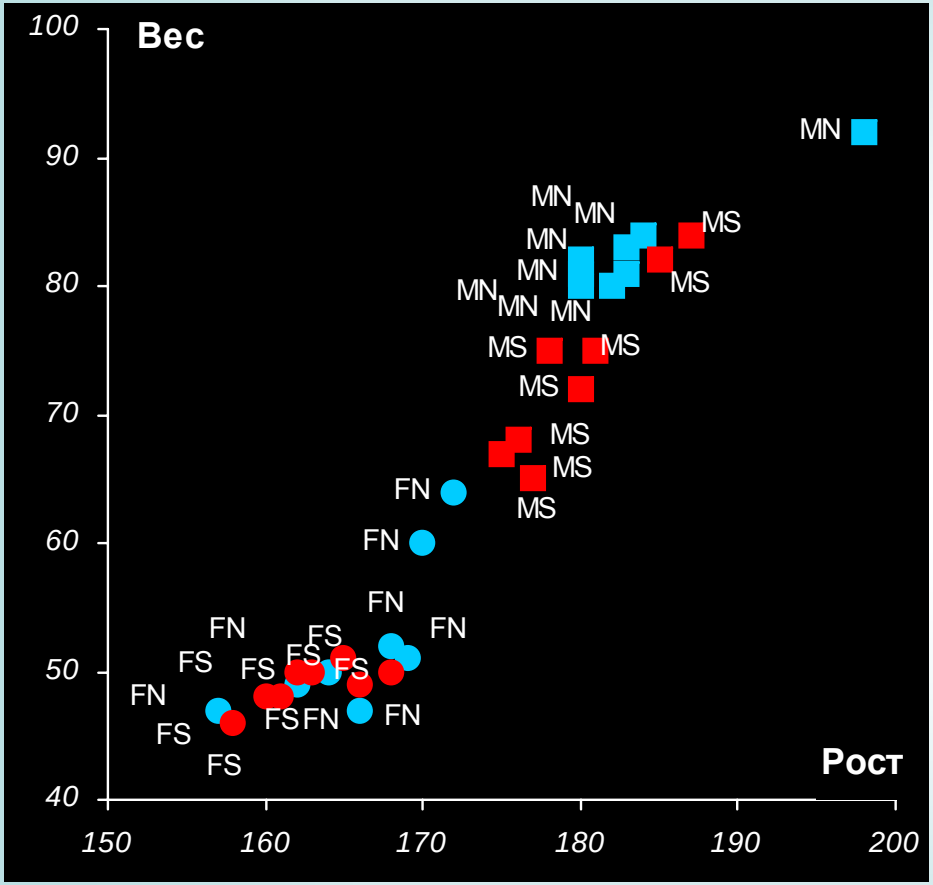


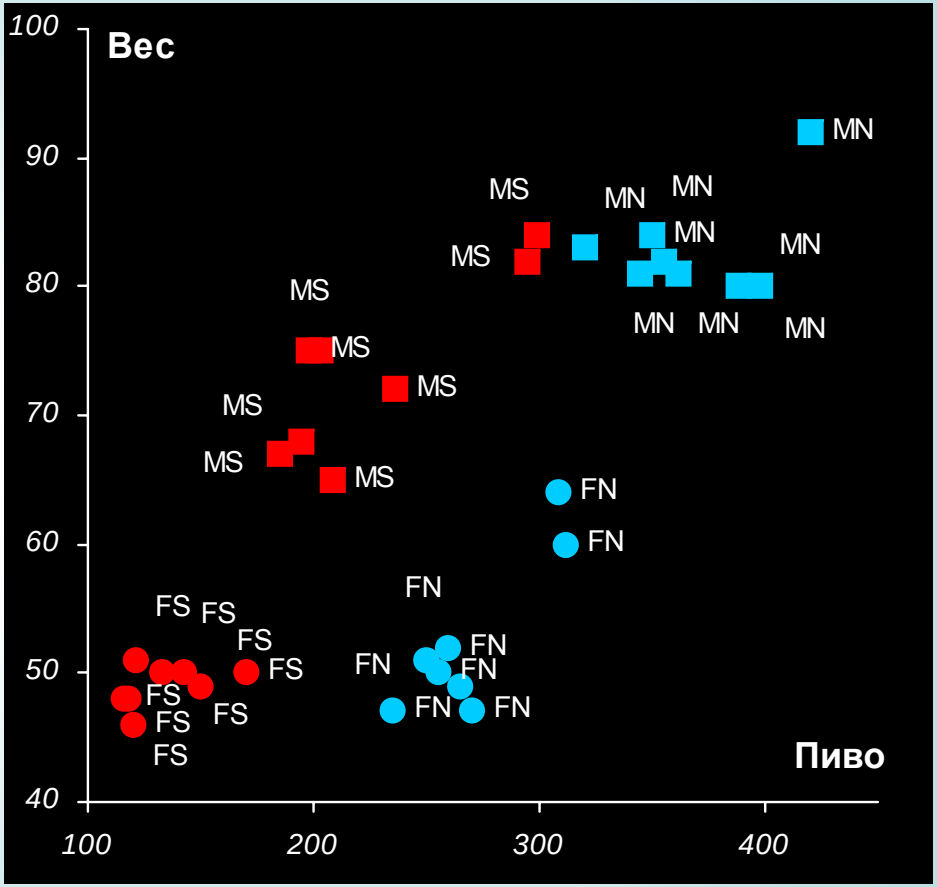
Графическое представление 3D-данных

	X_1	X_2	X_3
1	0.631	0.421	0.504
2	0.663	0.537	0.510
3	0.544	0.825	0.637
4	0.662	0.954	0.736
5	0.581	1.178	0.866
6	0.758	0.338	0.482
7	0.679	0.611	0.634
8	0.644	0.870	0.744
9	0.713	1.030	0.756
10	0.748	1.166	0.914
11	0.787	0.372	0.482
12	0.820	0.635	0.678
13	0.773	0.831	0.676
14	0.735	0.964	0.861

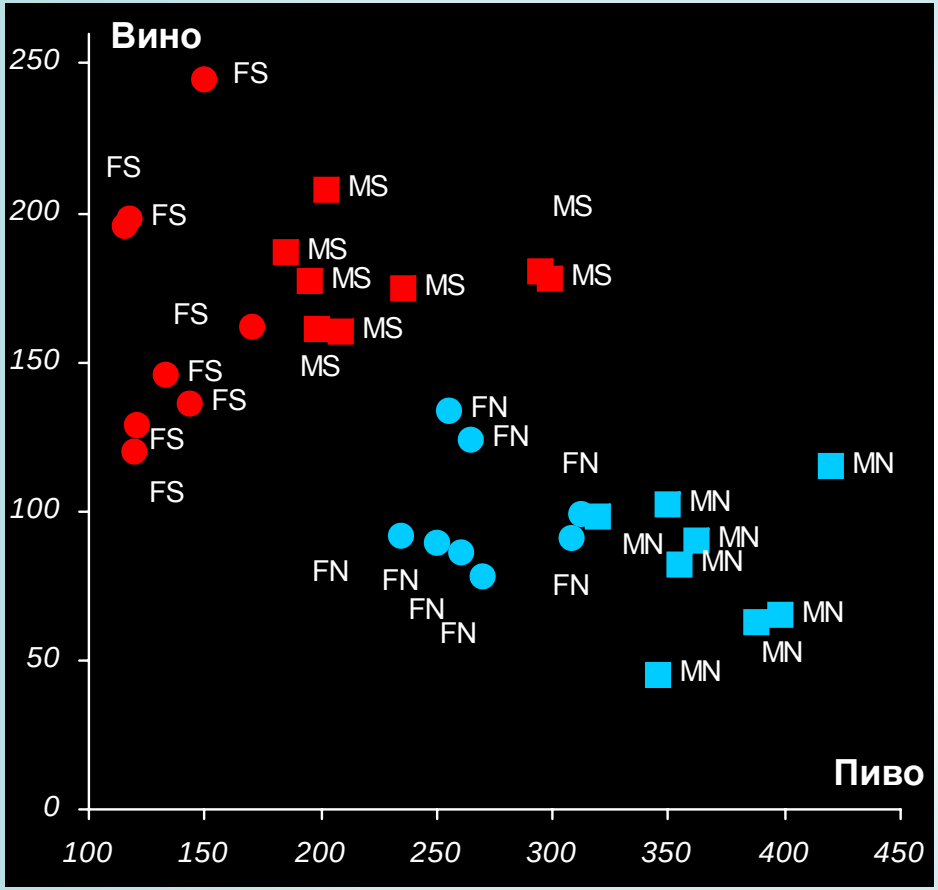


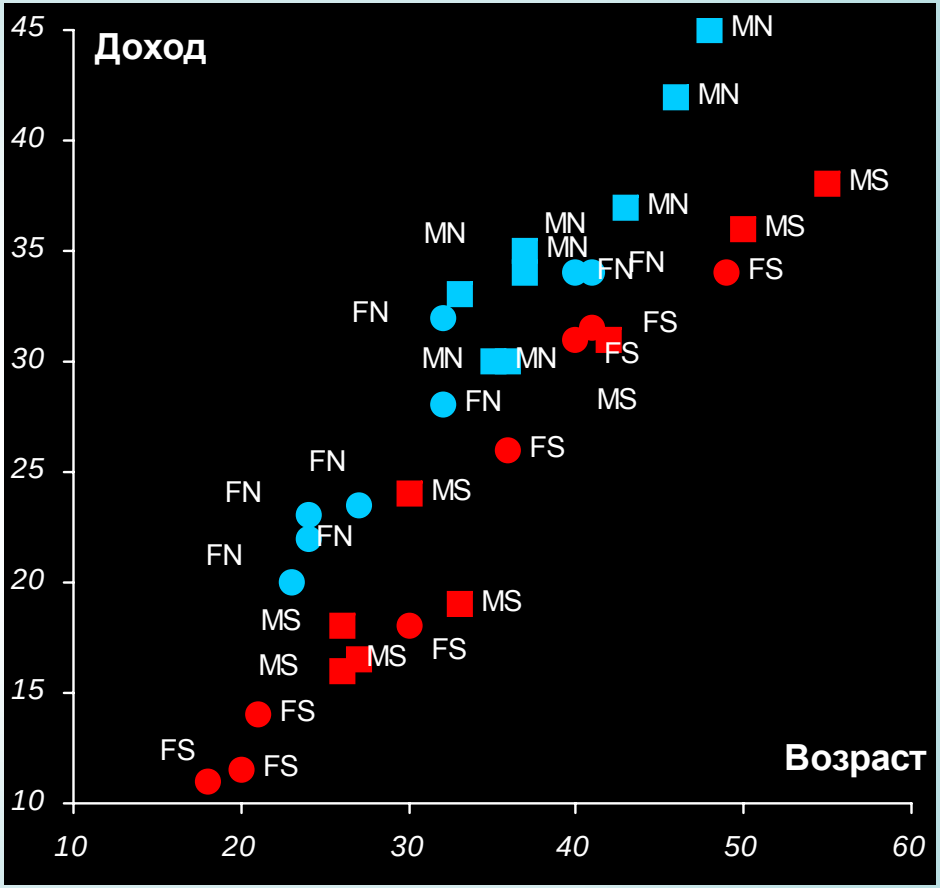
Корреляции 1





Корреляции 2

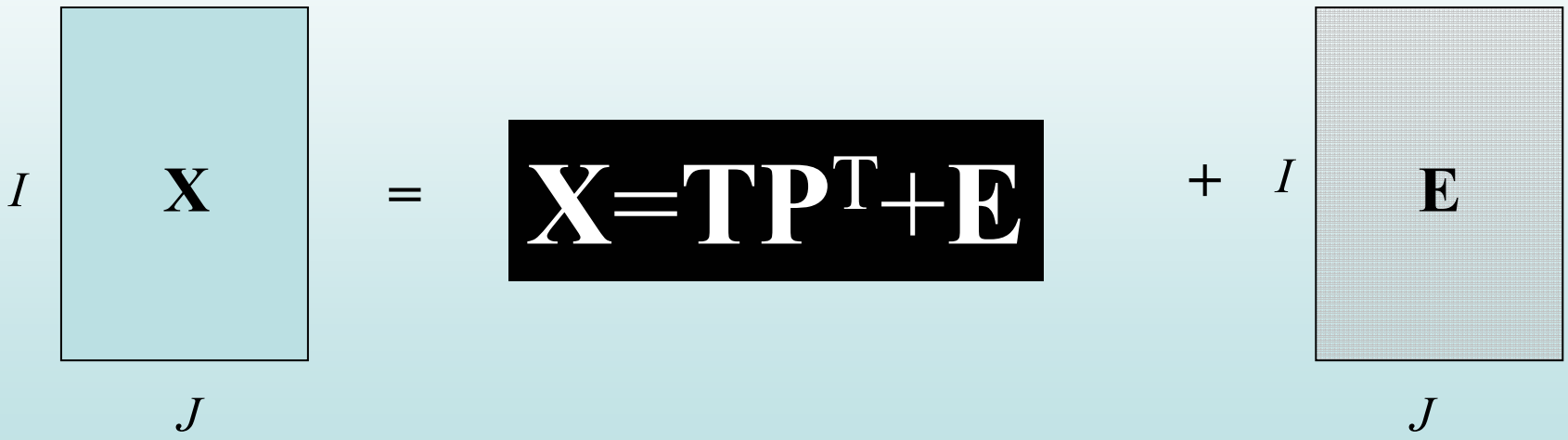




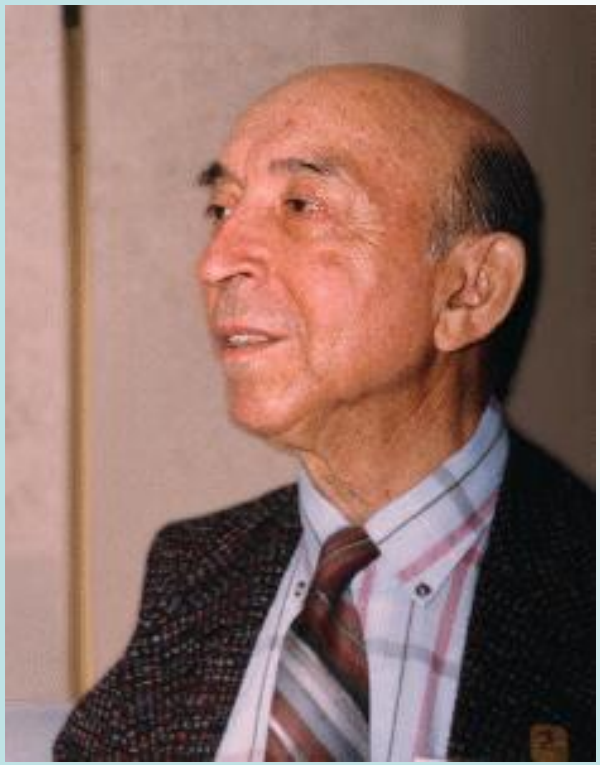
Метод главных компонент (PCA)

Исходные данные

Матрица ошибок



3. ТЕОРИЯ НЕЧЕТКИХ МНОЖЕСТВ (FUZZY SETS THEORY)



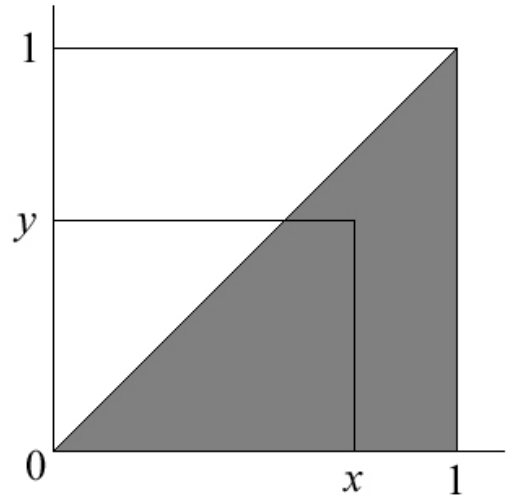
«Отец» нечеткой логики

Профессор Калифорнийского университета Лотфи Заде (р. 1921)

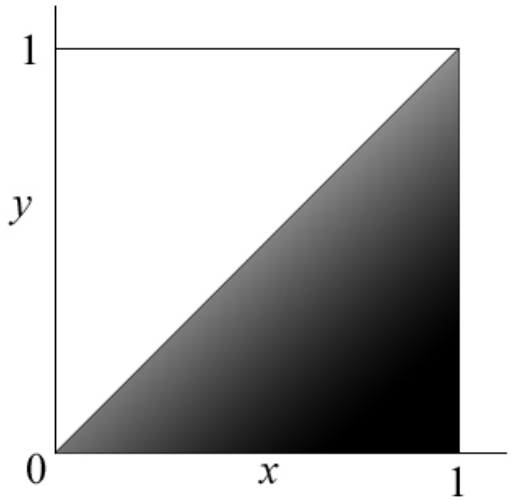
Zadeh L. A., ***Fuzzy sets****. Information and Control, Vol. 8, pp. 338—353. (1965).*

ОСНОВНЫЕ ПОЛОЖЕНИЯ

Классическая теория множеств

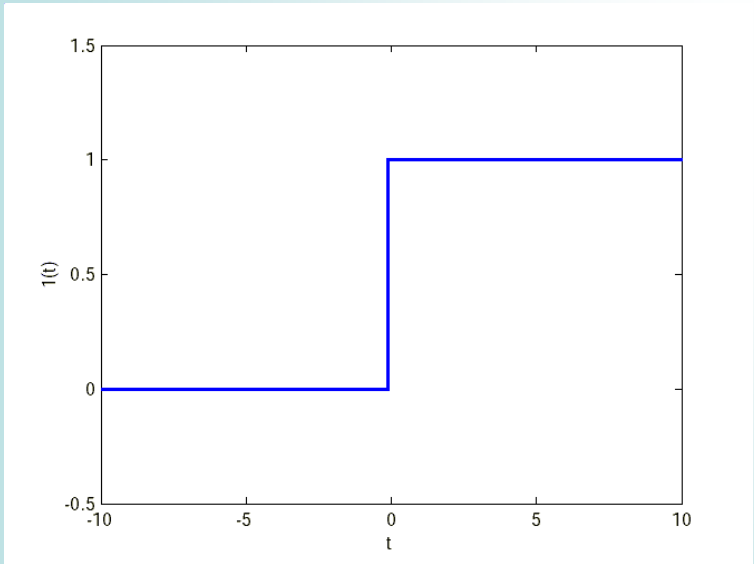


Теория нечетких множеств



Отношение $x>y$

Классическая теория множеств

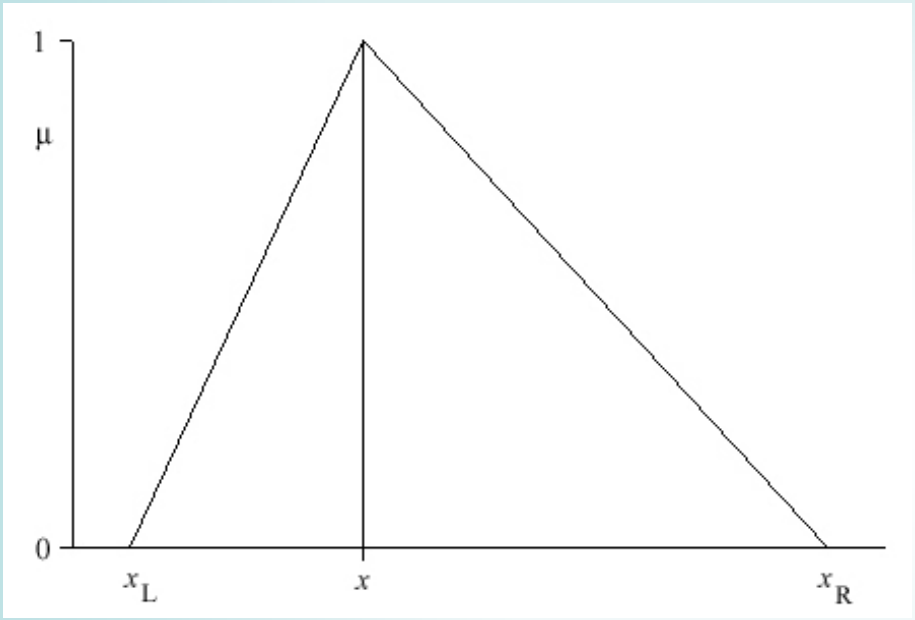


Теория нечетких множеств

Функция принадлежности

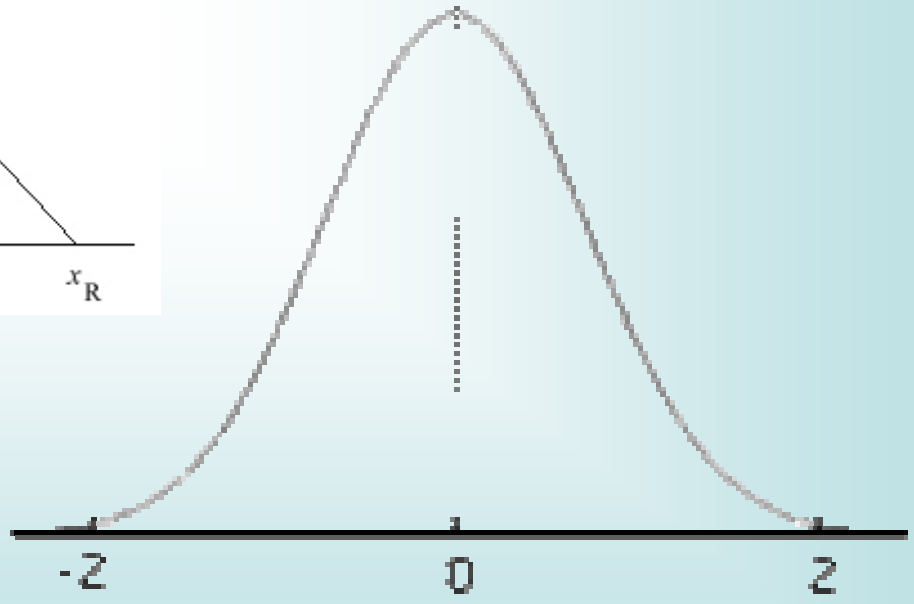
$0<\mu<1$

Представление числа *x* в виде нечеткого (фаззификация)

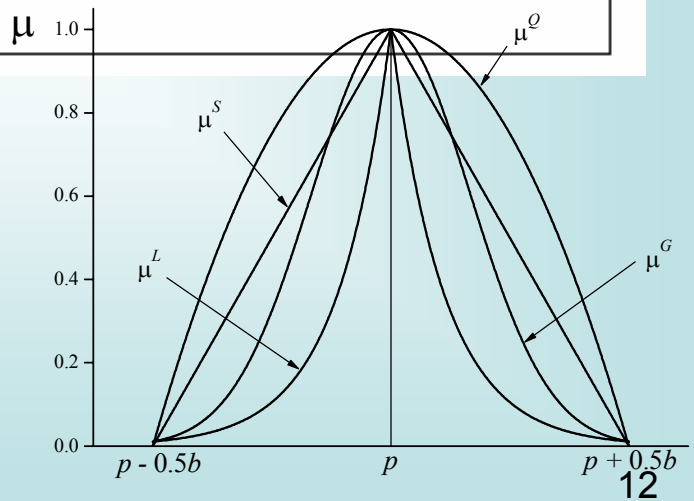


Задать набор

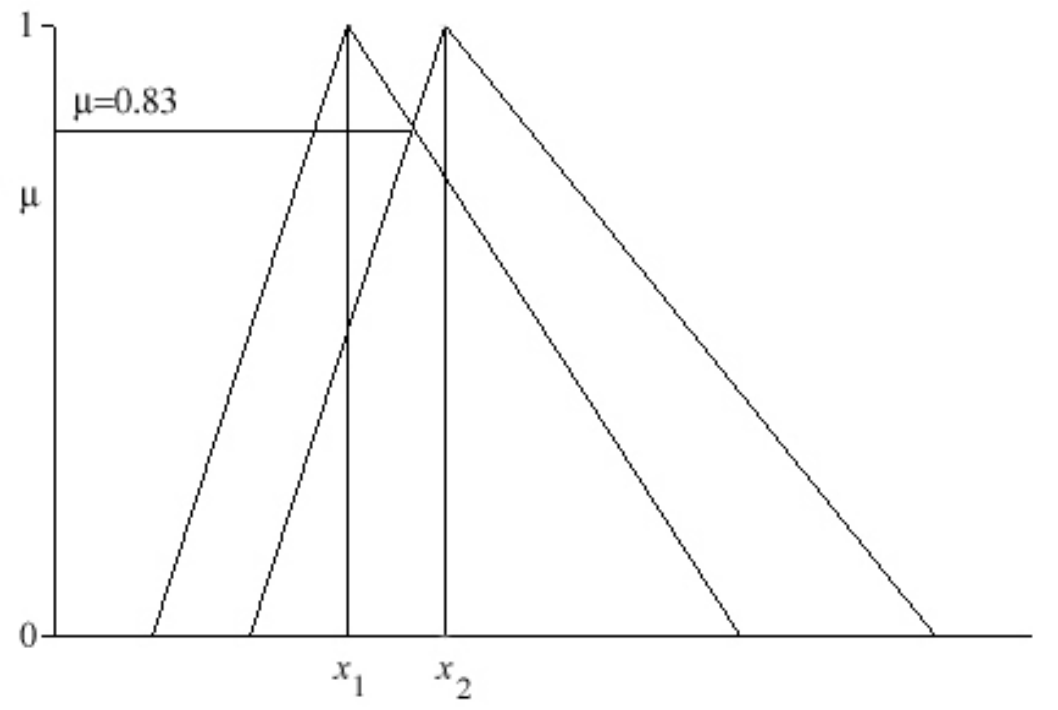
x_i; μ_i



Тип функции принадлежности	Формула для вычисления
Симпсона	$\mu^S(x) = 1 - \frac{2}{b}\lvert p - x\rvert$
Квадратичная	$\mu^Q(x) = a_0(x-p)^2 + a_1\lvert x-p\rvert + a_2$
Гаусса	$\mu^G(x) = \exp\left(-0.5\left[\frac{x-p}{\sigma^G}\right]^2\right)$
Лапласа	$\mu^L(x) = \exp\left(-\left\lvert\frac{x-p}{\sigma^L}\right\rvert\right)$



ДЕЙСТВИЯ НАД НЕЧЕТКИМИ ЧИСЛАМИ

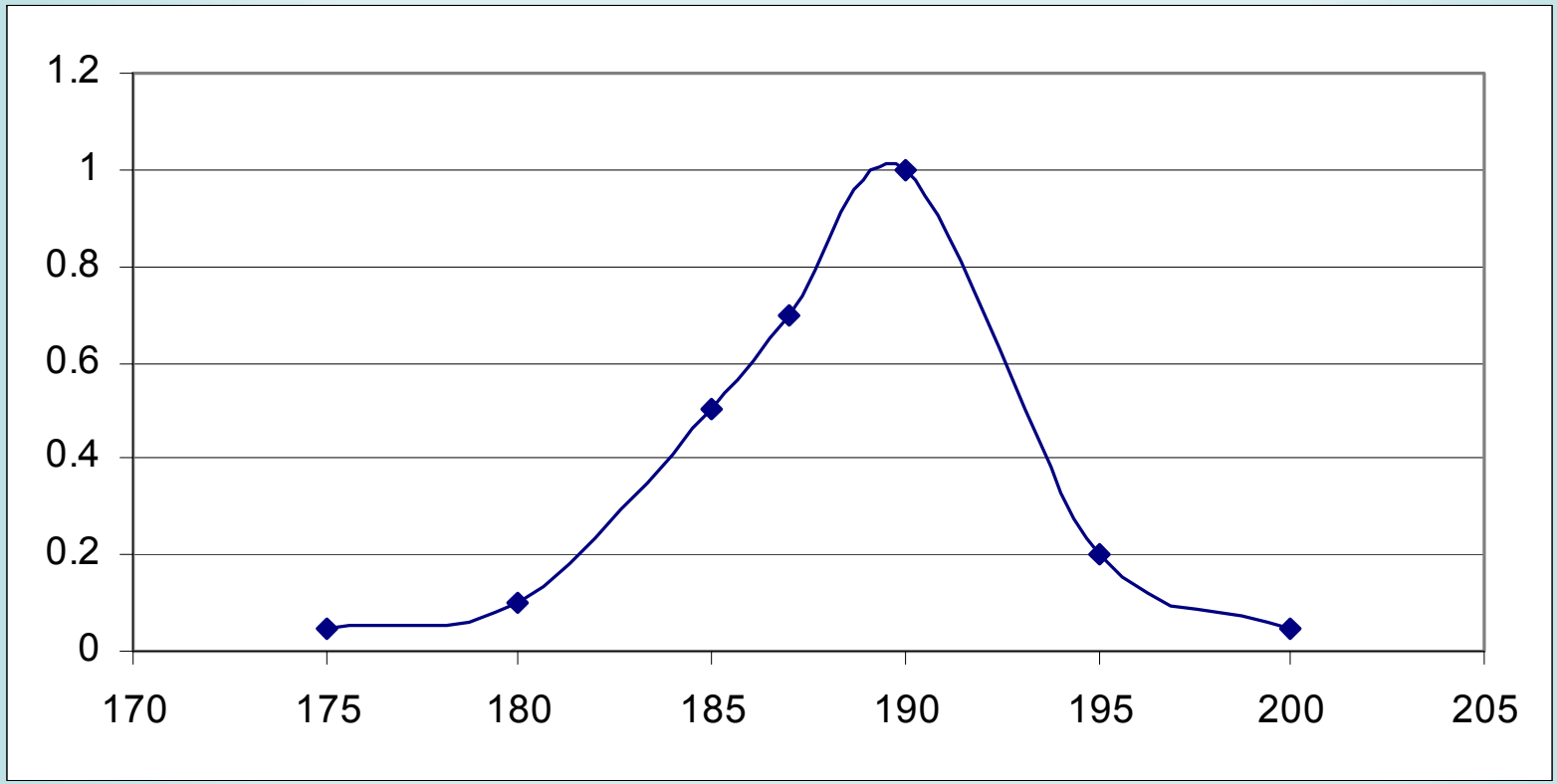


Принадлежность нечетких чисел одному множеству P

Мощность множества $$\mathrm{card}(P) = \frac{1}{N}\sum_{i=1}^{N}\mu_i$$

РОСТ ВЫСОКИХ ЛЮДЕЙ

$$x_i;\ \mu_i = (175;0.05)(180;0.1)(185;0.5)\ (187;0.7)(190;1)(195;0.2)(200;0.05)$$



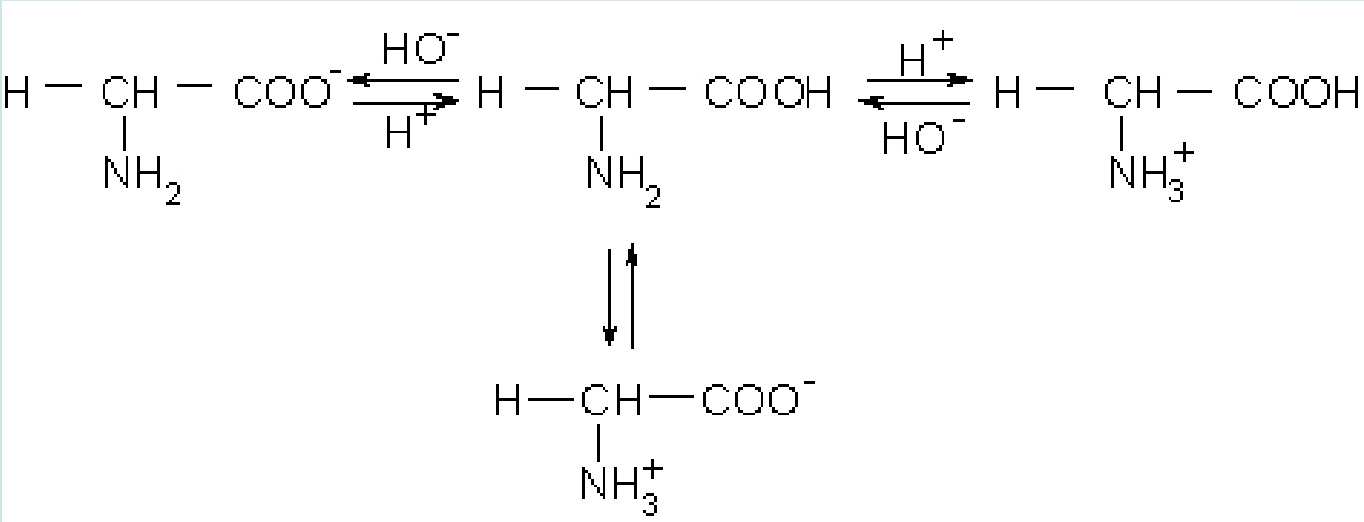
МЕТОДЫ ПЕРЕХОДА К ЧЕТКИМ ЧИСЛАМ (ДЕФАЗЗИФИКАЦИЯ)

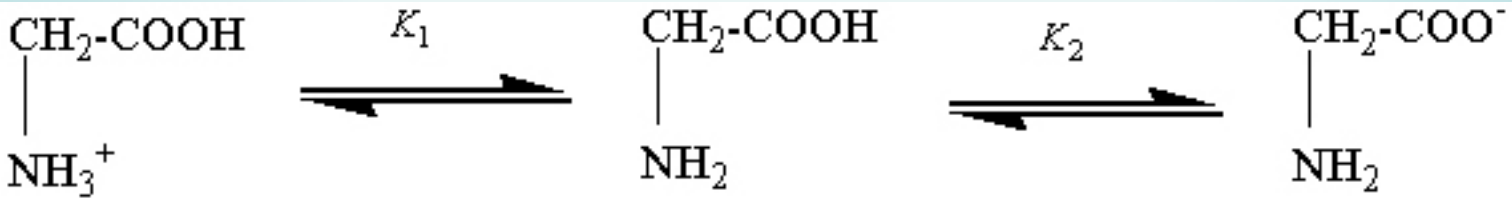
$$\theta^*_{centr} = \frac{\sum_{i=1}^{N} \theta_i \cdot \mu_i}{\sum_{i=1}^{N} \mu_i}$$

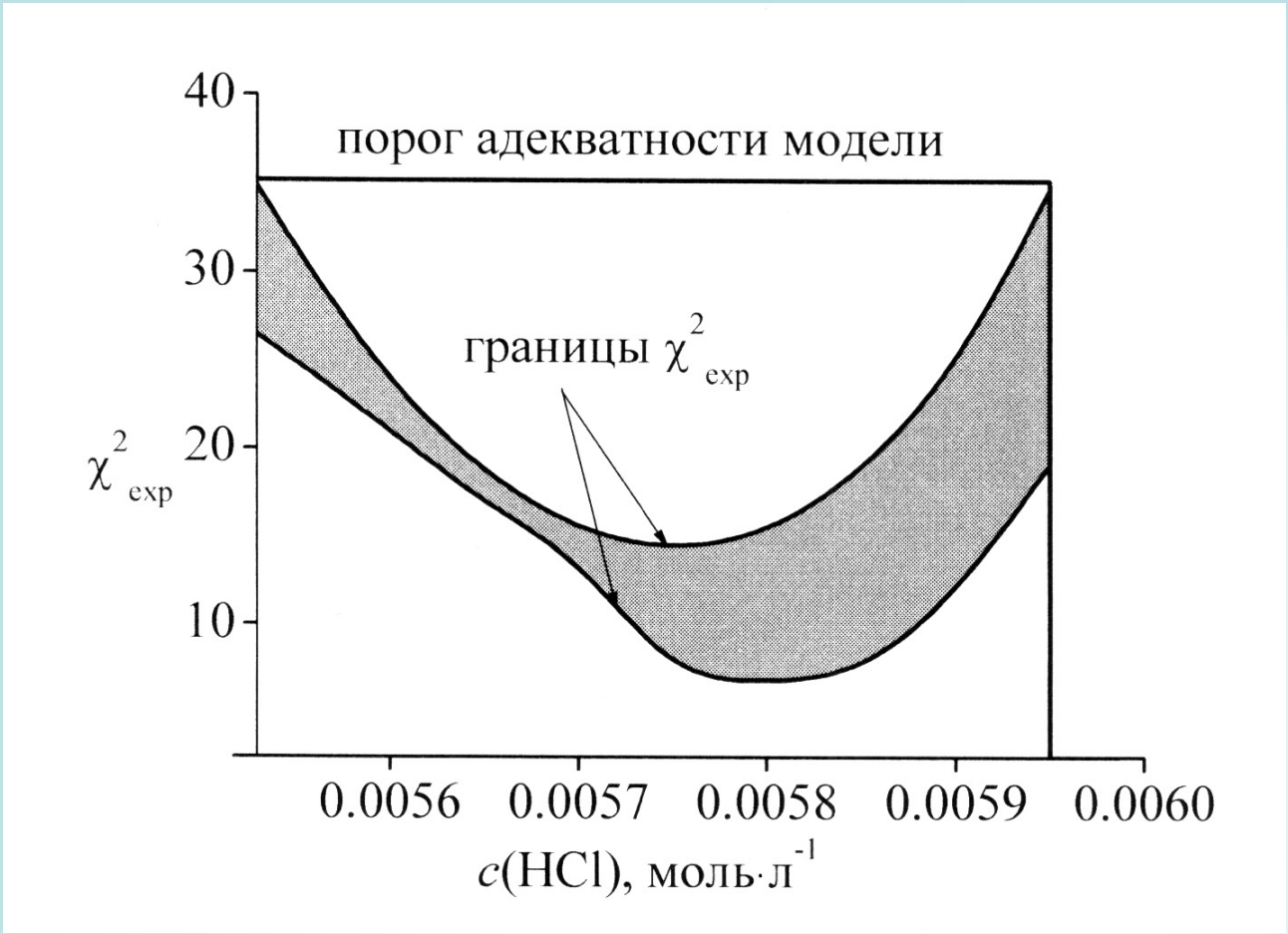
Центроидный метод

Рост высокого человека = 188 см

ПРИМЕР ПОБЛИЖЕ К ХИМИИ

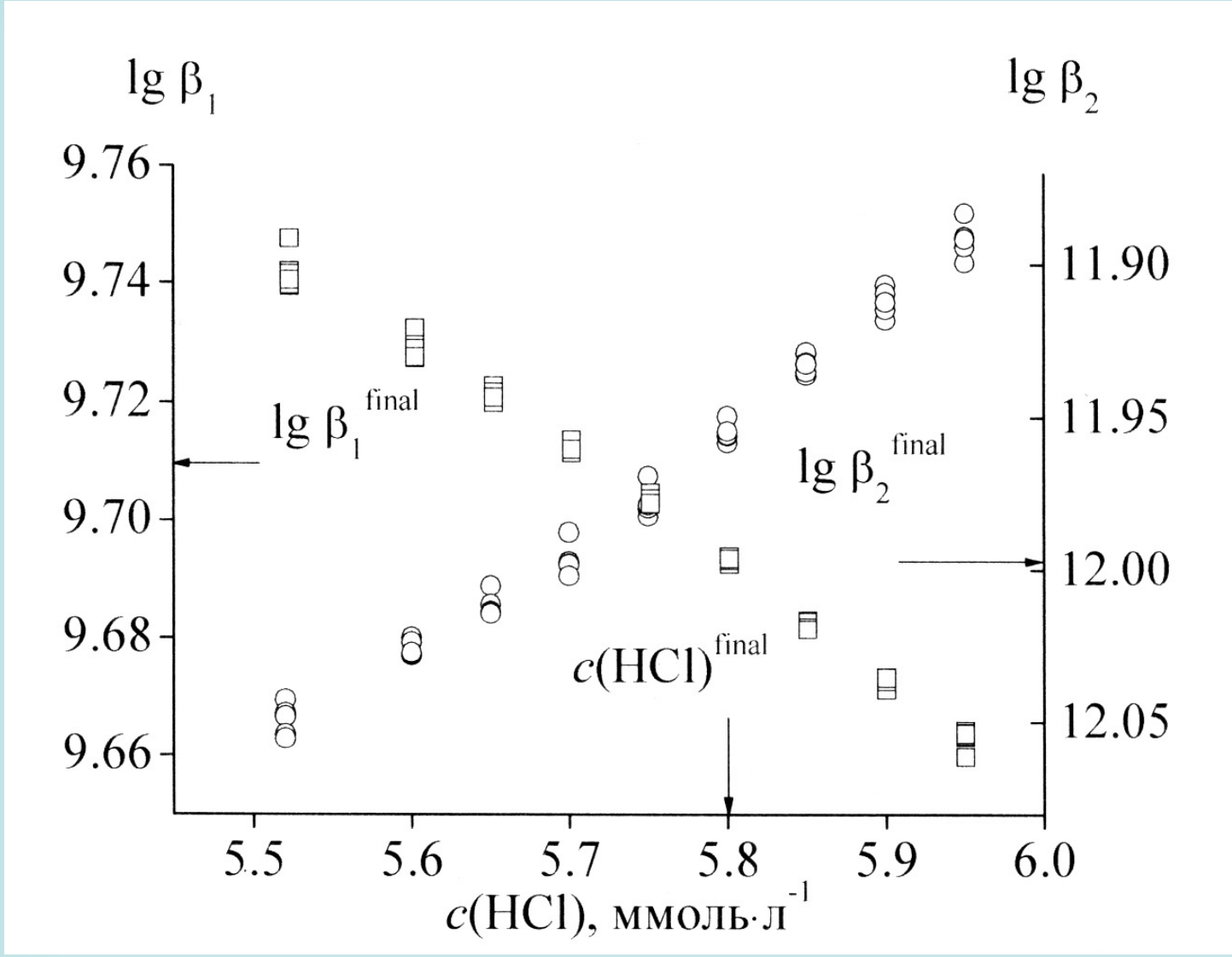




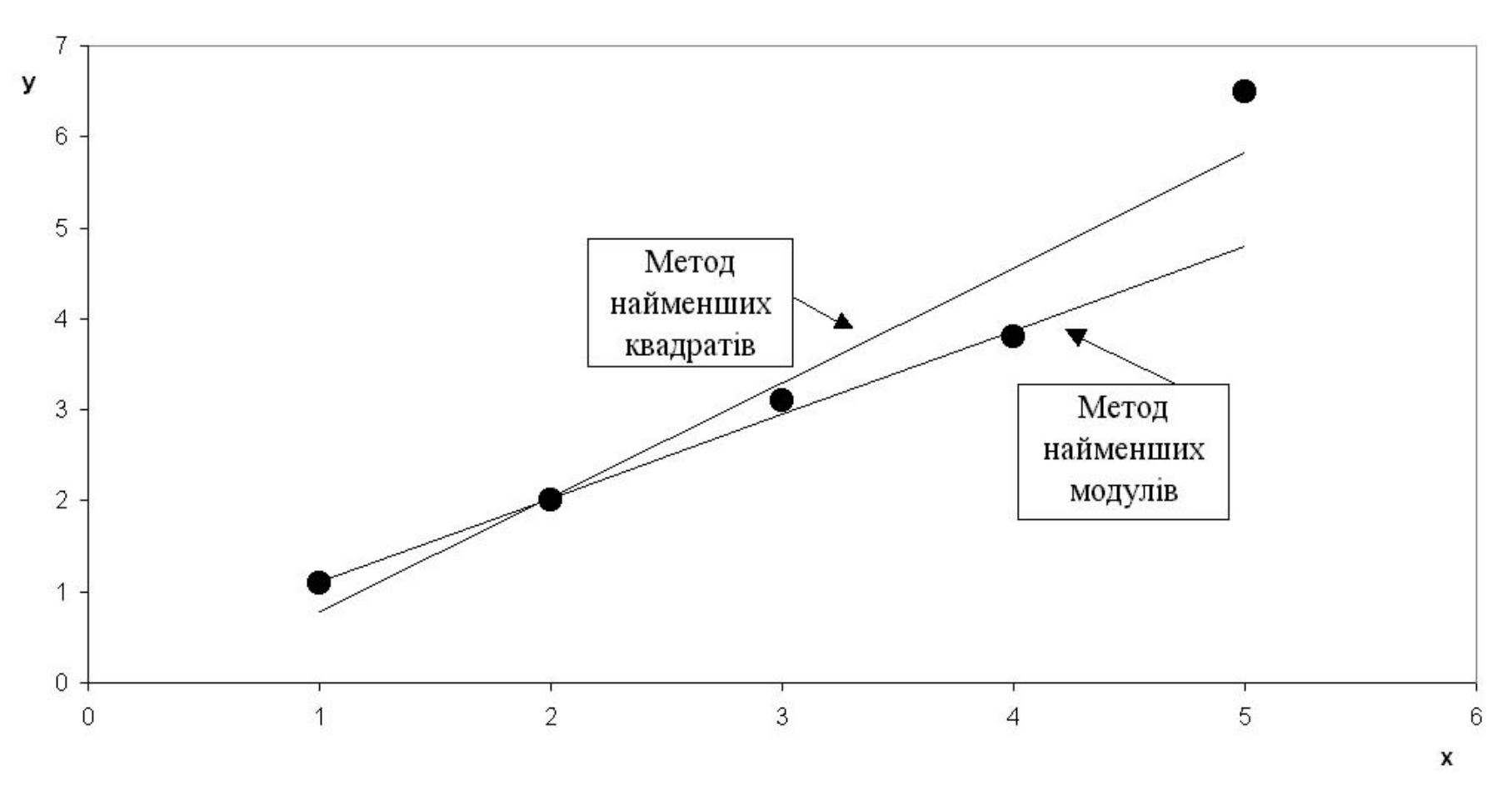


$$\mu_i^{\mathrm{I}} = A \cdot \exp(-\chi_i^2)$$

$$\mu_i^{\mathrm{II}} = \exp\left(-\frac{1}{2}\left(\chi_i^2 - \chi_{\min}^2\right)\right)$$


lg β1
lg β2
9.76
9.74
9.72
9.70
9.68
9.66
11.90
11.95
12.00
12.05
lg β1 final
lg β2 final
c(HCl) final
5.5
5.6
5.7
5.8
5.9
6.0
c(HCl), ммоль·л-1

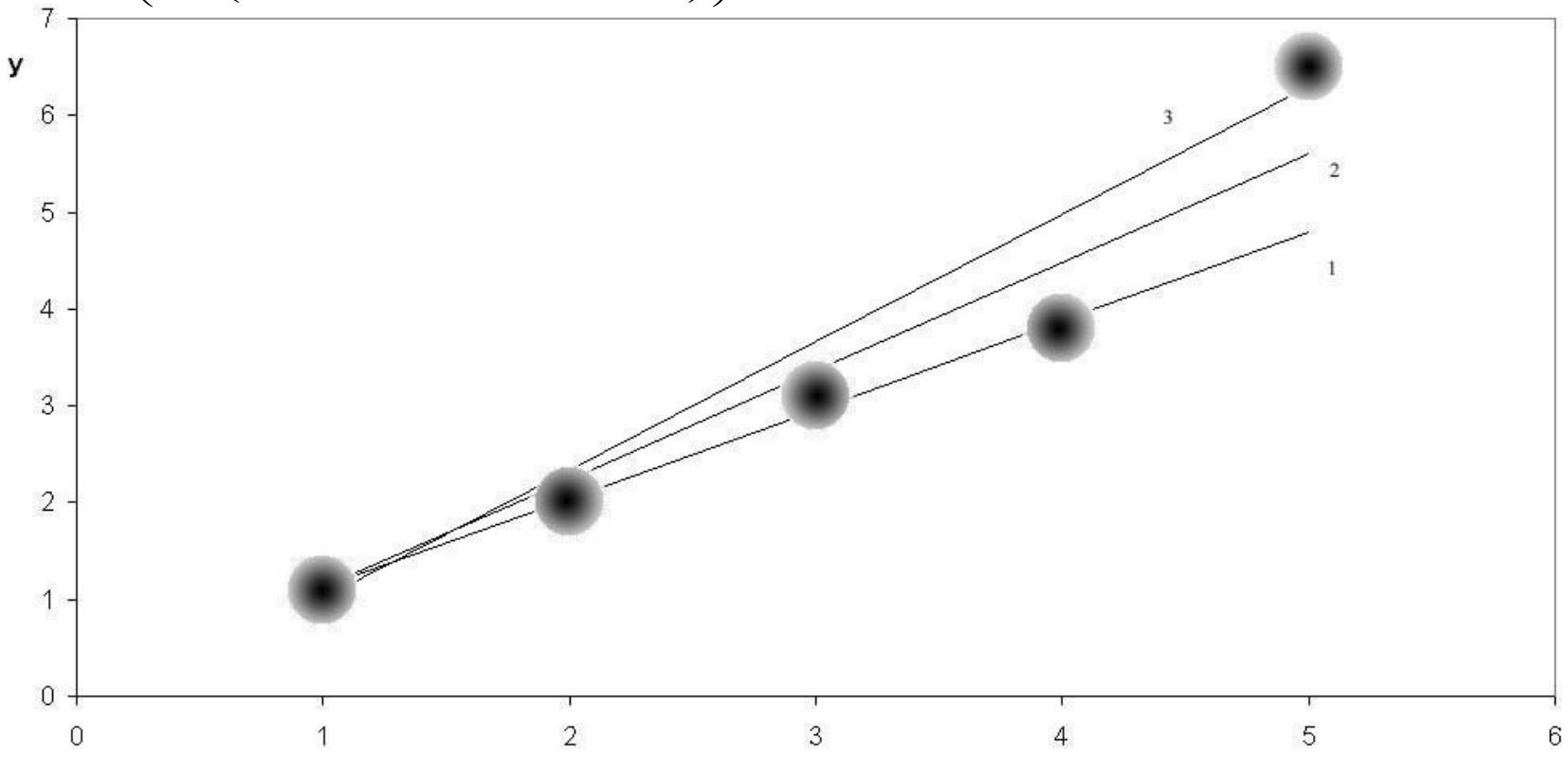
РОБАСТНОСТЬ ФАЗЗИ-ОЦЕНОК



МНК y = 1.26x - 0.48,

МНМ y = 0.92x + 0.19

$$\mu_i = \left(1 - \left(\left(\frac{x - x_i}{u_i}\right)^2 + \left(\frac{y - y_i}{v_i}\right)^2\right)\right)$$



$$1: y = 0.92x + 0.2,\ \mu = 0.77;$$

$$2: y = 1.26x + 0.0,\ \mu = 0.58;$$

$$3: y = 1.32x - 0.3,\ \mu = 0.32$$

СТРАТЕГИЯ НЕЧЕТКОГО ОЦЕНИВАНИЯ

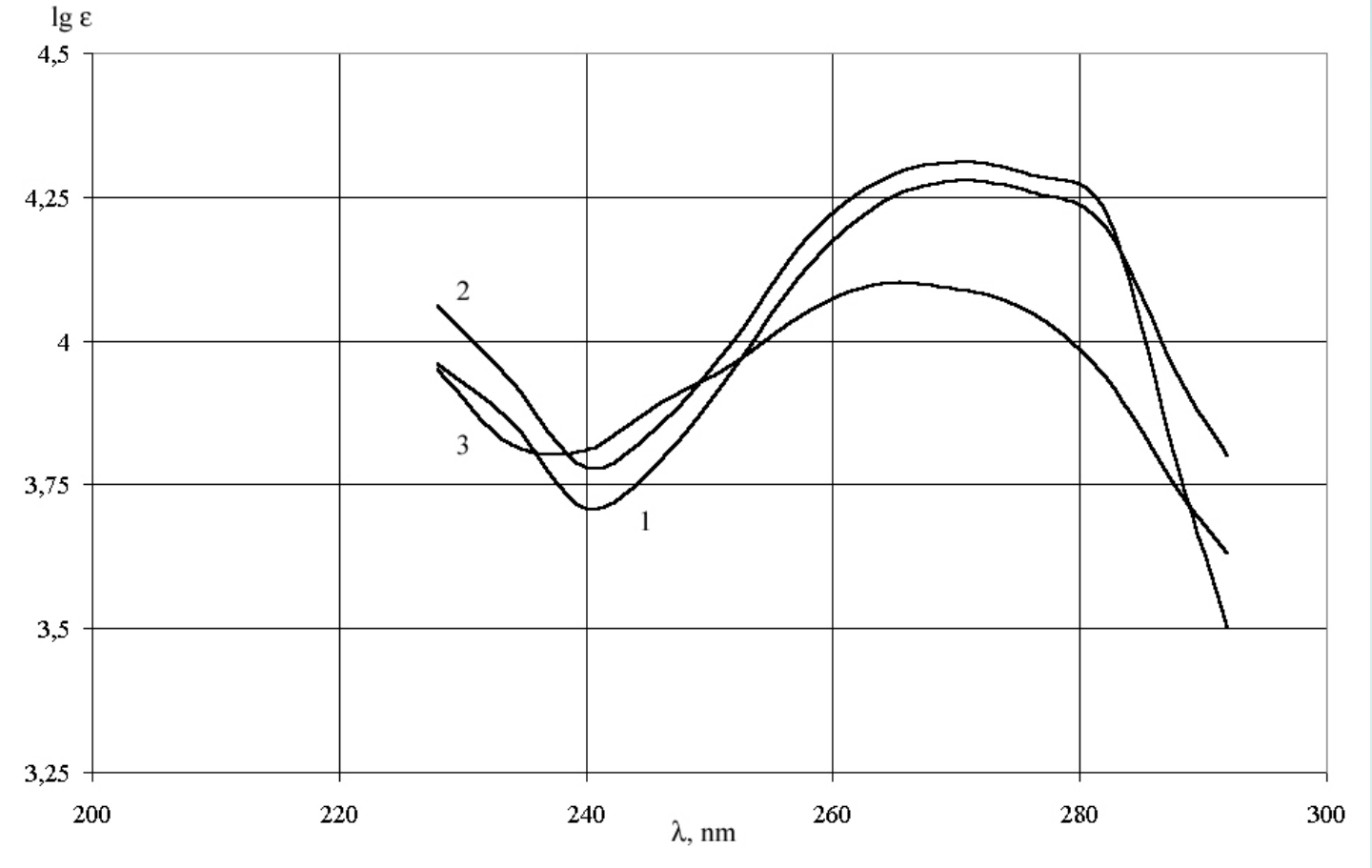
параметров аппроксимирующих функций

$$\left|\theta\right\rangle = \arg\max\left(\frac{1}{M}\sum_{i=1}^{M}\mu_i\right) = \arg\max\left(\mathrm{card}\left(P\right)\right)$$

аналог оценок

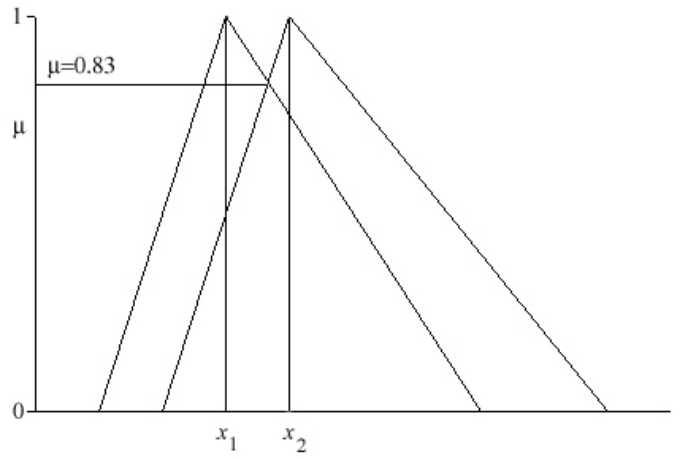
$$\left|\theta\right\rangle = \arg\min\left(\chi^2\right)$$

СПЕКТРАЛЬНЫЕ ДАННЫЕ



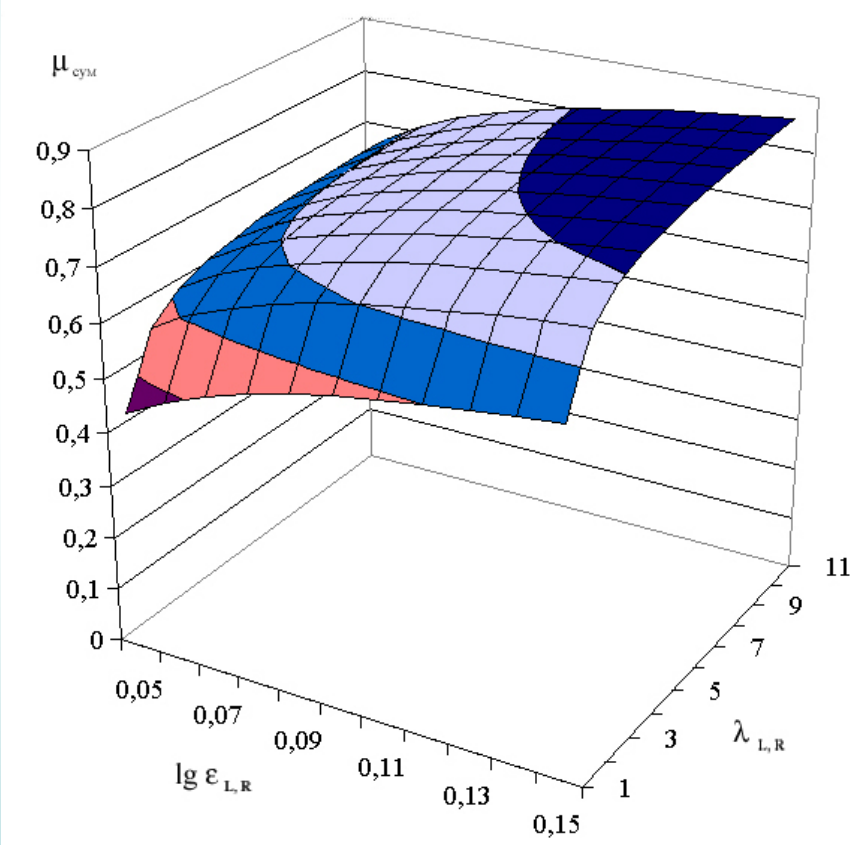
Использовать Евклидову метрику? $d = \sqrt{\sum_i \left(a(x_i) - b(x_i)\right)^2}$

$\mu_{сум}$; d		Номер спектра		
		1	2	3
Номер спектра	1	1; 0		
	2	0.80; 0.54	1; 0	
	3	0.34; 0.85	0.33; 0.87	1; 0



$$\mu_i = \left(1 - \left(\left(\frac{x - x_i}{u_i}\right)^2 + \left(\frac{y - y_i}{v_i}\right)^2\right)\right)$$

Границы нечеткости стабилизируют значение ФП



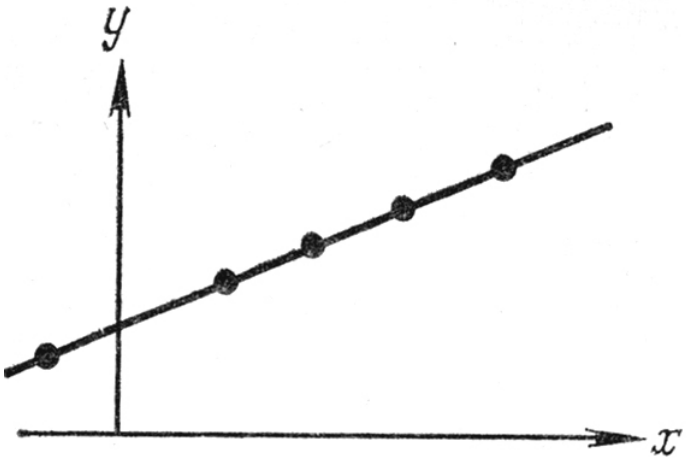
ОБЛАСТИ ПРИМЕНЕНИЯ

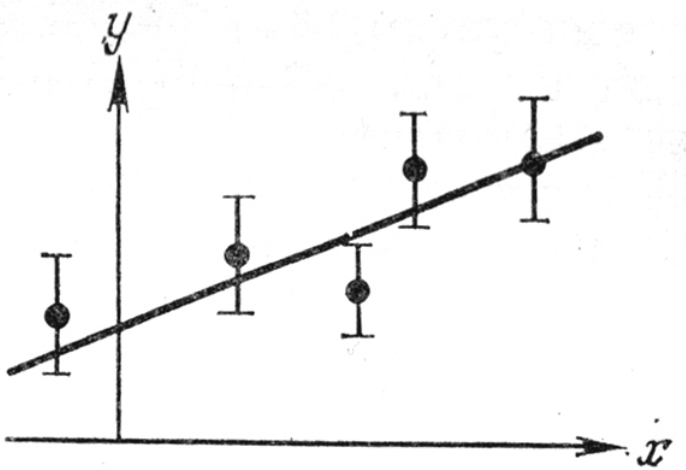
•нелинейный контроль за процессами (производство);
•самообучающиеся системы (или классификаторы), исследование рисковых и критических ситуаций;
•распознавание образов;
•финансовый анализ (рынки ценных бумаг);
•исследование данных (корпоративные хранилища);
•совершенствование стратегий управления и координации действий, например сложное промышленное производство

Регрессионный анализ.

Расчетная схема МНК

2 переменные x и y связаны линейной зависимостью





Идеальный случай

$y=A+Bx$

Реальный случай

$y=A+Bx+\varepsilon$

Предпосылки использования МНК. Условия Гаусса-Маркова.

3 N-мерных вектора: x, y, ε

1. $M(\varepsilon_i)=0$ – переходим от реального случая к идеальному.
2. $D(\varepsilon_i)= D(\varepsilon_j)=\sigma^2$, или, что то же самое $M(\varepsilon_i)= \sigma^2$ – условие гомоскедастичности.
3. $\sigma_{\varepsilon_i \varepsilon_j} = \mathrm{cov}(\varepsilon_i \varepsilon_j) = \begin{cases} 0, i \neq j \\ \sigma^2, i = j \end{cases}$ или, что то же самое $M(\varepsilon_i \varepsilon_j)= =0$ $(i\neq j)$ – отсутствие автокорреляции.
4. $\mathrm{cov}(x_i, \varepsilon_i)=0$.
5. Модель линейна по параметрам.

Матричный МНК

Пример

$x_1 + 10x_2 = 11.0$

$10x_1 + 101x_2 = 111$

$$a = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

$$Y = XA$$

$$X^T Y = X^T XA$$

$$A = (X^T X)^{-1} X^T Y$$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_N \end{pmatrix}$$

ε_i распределены нормально с нулевым средним и дисперсией σ^2

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} \longrightarrow \hat{y} = A + Bx \longrightarrow \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_N \end{pmatrix}$$

Метод максимума правдоподобия

$$P_{A,B}(y_1,\dots,y_N) = \prod_{i=1}^{N} p_{A,B}(y_i) = \prod_{i=1}^{N} \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{y_i - A - Bx_i}{\sigma_y} \right]^2 \right) \sim \frac{1}{\sigma_y^N} \exp\left(-\frac{1}{2}\chi^2 \right)$$

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - A - Bx_i}{\sigma_y} \right)^2 = \sum_{i=1}^{N} \left(\frac{y_i - \hat{y}_i}{\sigma_y} \right)^2$$

Статистические свойства оценок МНК. Анализ адекватности моделей.

Условия Гаусса-Маркова.

1. $M(\varepsilon_i)=0$ – переходим от реального случая к идеальному.
2. $D(\varepsilon_i)= D(\varepsilon_j)=\sigma^2$, или, что то же самое $M(\varepsilon_i)= \sigma^2$ – условие гомоскедастичности.
3. $\sigma_{\varepsilon_i\varepsilon_j} = \mathrm{cov}(\varepsilon_i\varepsilon_j) = \begin{cases} 0, i \neq j \\ \sigma^2, i = j \end{cases}$ или, что то же самое $M(\varepsilon_i \varepsilon_j)= =0$ $(i\neq j)$ – отсутствие автокорреляции.
4. $\mathrm{cov}(x_i, \varepsilon_i)=0$.
5. Модель линейна по параметрам.

Вычислите $\left(\frac{\partial \chi^2}{\partial A}\right) = 0$

$\left(\frac{\partial \chi^2}{\partial B}\right) = 0$

$$AN + B\sum x_i = \sum y_i$$

$$A\sum x_i + B\sum x_i^2 = \sum x_i y_i$$

Выразите коэффициенты!

Окончательные выражения для вычисления коэффициентов

$$A = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{\Delta}$$

$$B = \frac{N(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\Delta}$$

$$\Delta = N(\sum x_i^2) - (\sum x_i)^2$$

Не всегда $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2 = \text{const}$.

Если $\operatorname{cov}(\varepsilon) = \begin{pmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \sigma_N^2 \end{pmatrix}$, то вводится коэффициент

$w_i = \dfrac{1}{\sigma_i^2}$ - статистический вес, и минимизируемый

в задаче МНК функционал принимает вид

$$U = \sum_i w_i \left(y_i - \hat{y}_i \right)^2$$

Правило переноса погрешностей

Если $y = f(x_i)$, то

$$s_y^2 = \sum_i \left(\frac{\partial f}{\partial x_i} \right)^2 \cdot s_{x_i}^2$$

Задача

Оценивают параметры равновесий в растворе с pH=2÷12. Для этого измеряют pH. Известно, что s^2(pH)=0.01. Определите $s^2([H^+])$ и сформируйте матрицу весовых множителей.

1. Оценки МНК несмещенные, т.е. M(A)=A, M(B)=B.
2. Оценки МНК состоятельные, т.е. при

$$N \longrightarrow \infty \; D(A) \longrightarrow 0, D(B) \longrightarrow 0$$

3. Оценки МНК эффективные, т.е. D(A) и D(B) минимальны.

Остаточная дисперсия $$s_0^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$

Откуда берется 2?

Погрешности в коэффициентах регрессии

$$s_A^2 = s_0^2 \sum_i x_i^2 \; / \; \Delta$$

$$s_B^2 = N s_0^2 \; / \; \Delta$$

Наилучшие оценки результатов измерений

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} \longrightarrow \arg\min \sum_{i=1}^{N} \Delta_i^2$$

$$\tilde{x} = \text{чему}\,? \longrightarrow \arg\min \sum_{i=1}^{N} \left|\Delta_i\right|$$

Задача

Случайная величина x=[101.1; 102.5; 102.8; 104.2; 110.9].

Вычислите наилучшие оценки x для двух видов распределений.

Измерение r и его результат x

$$r + \Delta_i = x$$

$$p(\Delta, s^2) = \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\Delta}{s} \right)^2 \right)$$

МНК

$$p(\Delta, \bar{s}^2) = \frac{1}{2s} \exp\left(-\frac{|\Delta|}{s} \right)$$

МНМ

Расчетная схема МНМ.

$$U = \sum_i w_i \left(y_i - \hat{y}_i\right)^2$$ - взвешенный МНК

Веса меняются по итерационной процедуре

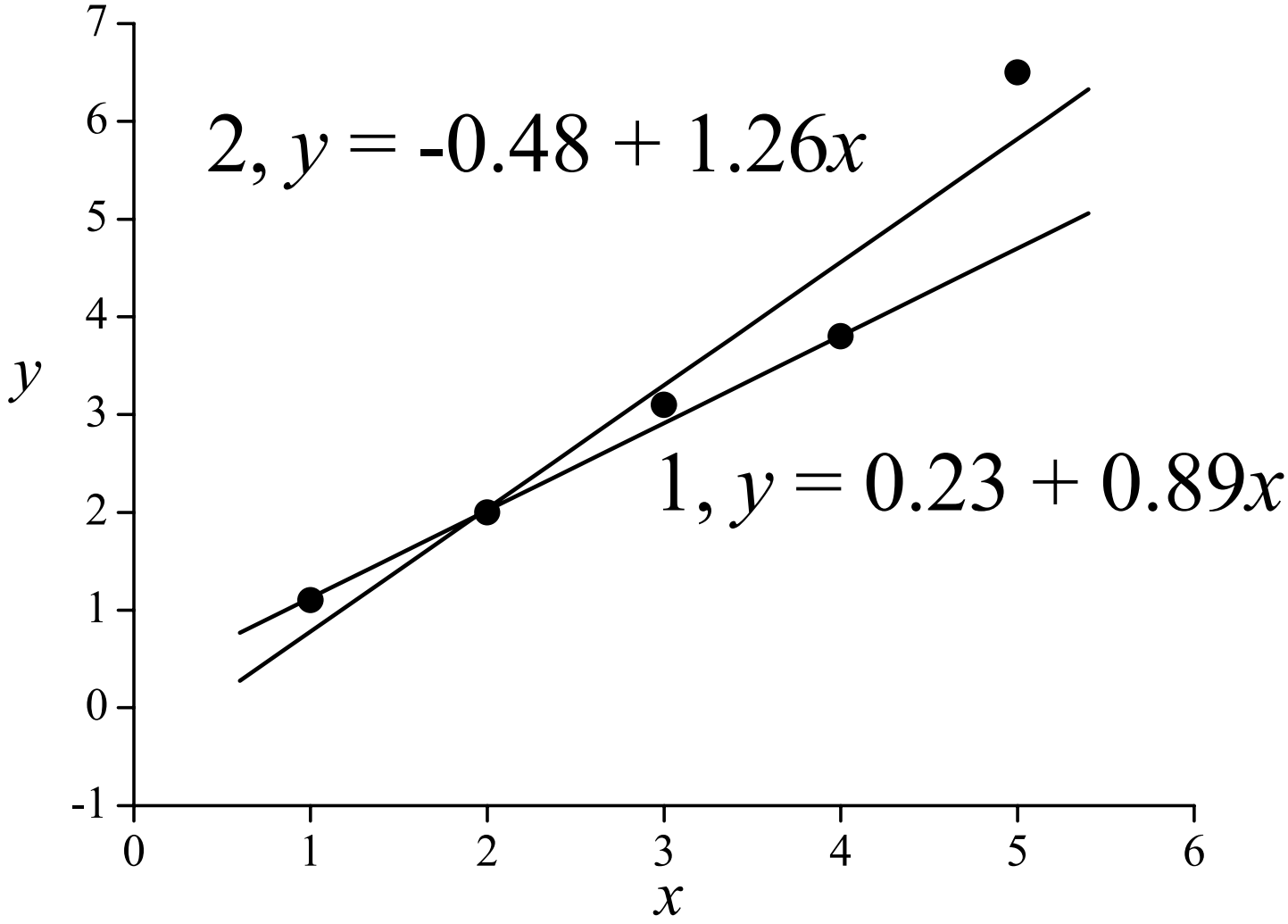
$$w_i = \frac{1}{\left|y_i - \hat{y}_i\right|} \longrightarrow P(a_0, a_1) = \sum \frac{1}{\left|y_i - \hat{y}_i\right|}\left(y_i - \hat{y}_i\right)^2 = \sum \left|y_i - \hat{y}_i\right|$$

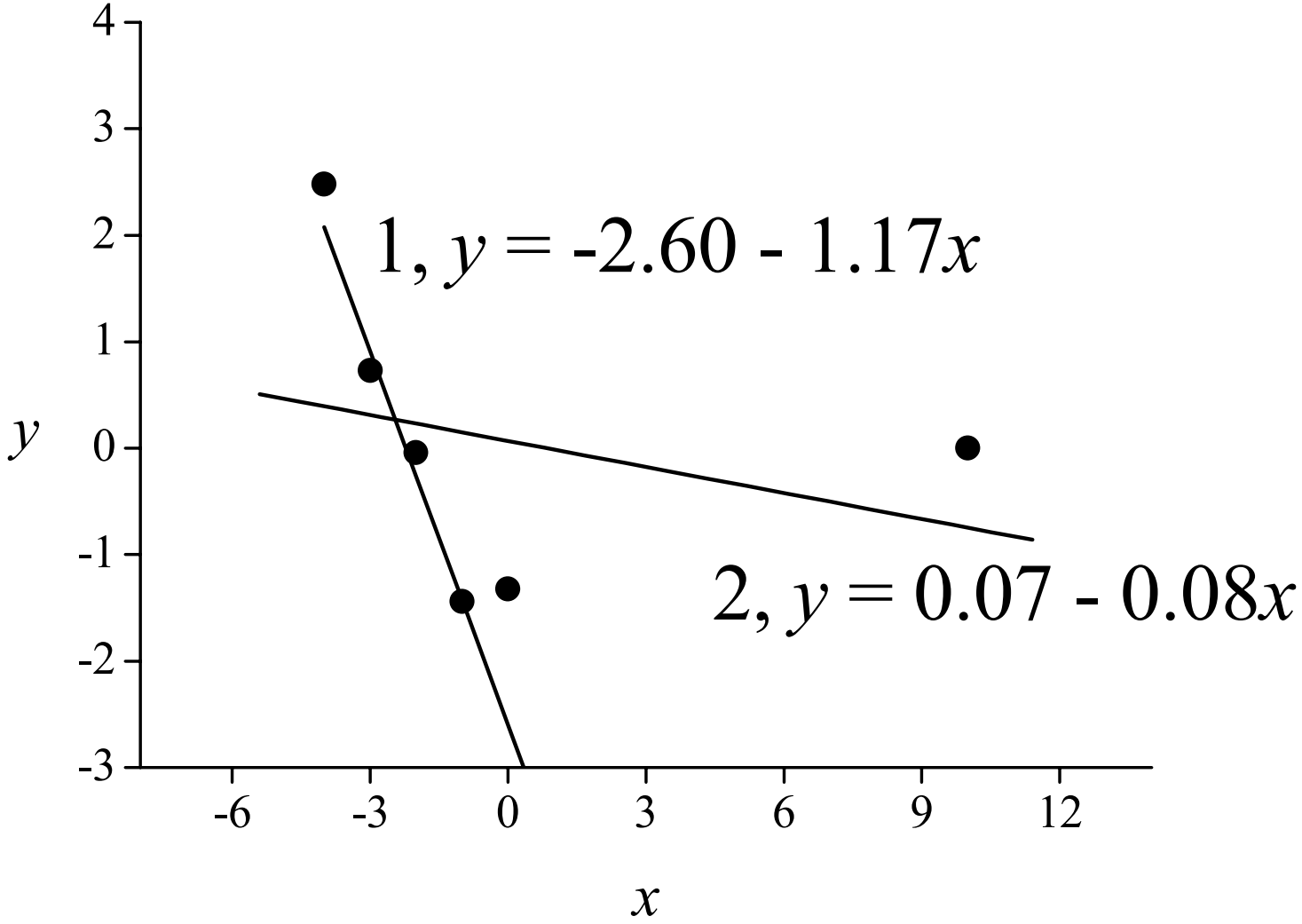
$$\hat{y}_i = a_0 + a_1 x$$

Минимаксная стратегия выбора метода

Исследование асимптотической эффективности оценок

Принятый метод обработки	Закон распределения ошибок измерения	
	Закон Гаусса	Закон Лапласа
МНК	1	0,5
МНМ	$\sqrt{\frac{2}{\pi}}=0,637$	1


2, y = -0.48 + 1.26x
1, y = 0.23 + 0.89x
y
x
7
6
5
4
3
2
1
0
-1
0
1
2
3
4
5
6


4
3
2
1
0
-1
-2
-3
y
-6
-3
0
3
6
9
12
x
1, y = -2.60 - 1.17x
2, y = 0.07 - 0.08x

Мультиколлинеарность и SVD

$$GA = Y$$

$$G^T GA = G^T Y$$

$$(G^T G)^{-1} G^T GA = (G^T G)^{-1} G^T Y$$

$$A = (G^T G)^{-1} G^T Y$$

Вырождение матрицы G

Singular value decomposition

$$G = U\Sigma V^T$$

$$(G^T G)^{-1} = ?$$