Статистическая обработка эксперимента

Восков Алексей Леонидович, с.н.с., к.х.н.
Лаборатория химической термодинамики, кафедра физической химии

Email: alvoskov@gmail.com; комн. Ц-19, Химический факультет МГУ
Сайт: http://td.chem.msu.ru (есть материалы к курсу)

Форма отчётности: недифференцированный зачёт по итогам работы в семестре

Краткое содержание курса

Часть 1. Погрешности, распределения и доверительные интервалы

1. Виды погрешностей. Правила округления. Погрешность косвенных измерений
2. Распределения и доверительные интервалы
3. Проверка статистических гипотез
4. Корреляция

Часть 2. Метод наименьших квадратов

1. Основы работы в GNU Octave
2. Одномерная и многомерная линейная регрессия
3. Нелинейная регрессия. Численные методы

Часть 3. Нестандартные ситуации и их решение

1. Основы двоичной арифметики, стандарт IEEE-754, погрешности округления
2. Глобальная оптимизация: генетические алгоритмы, метод отжига, символьная регрессия
3. Автоматическое дифференцирование

Аппаратное и программное обеспечение, литература

Аппаратное обеспечение

Обязательно: инженерный калькулятор

Рекомендуется x86-совместимый ноутбук (нетбук, планшет и т.п.)

Минимальная конфигурация

- Поддержка MS Office 97 или выше и GNU Octave в текстовом режиме (требует от 64 Мб ОЗУ)

Рекомендуемая конфигурация

- Поддержка MS Office 2007 или выше и GNU Octave или MATLAB с графической оболочкой

Программное обеспечение:

- Microsoft Excel 2007/2010
- GNU Octave (или MATLAB)
- Microsoft PowerPoint 2007/2010
- Просмотровщик PDF
- Текстовый процессор для сдачи домашних заданий (MS Word, LibreOffice, LaTeX)

Рекомендуемая литература:

1. К. Дёрффель, «Статистика в аналитической химии». М., «Мир», 1998.
2. А.В. Гармаш, Н.В. Сорокина «Метрологические основы аналитической химии»
3. «Основы аналитической химии». Отв. ред. Ю. А. Золотов. М., «Высшая школа», 2002.
4. Д. Химмельблау, «Анализ процессов статистическими методами». М., «Мир», 1973.
5. Ю. Н. Тюрин, А. А. Макаров, «Анализ данных на компьютере». М., «Форум», 2008.
6. «Справочник по прикладной статистике». Под ред. Э. Ллойда и У. Ледермана. М., «Финансы и статистика», 1989.
7. Дьяконов В.П. MATLAB. Полный самоучитель. ДМК Пресс, 2012
8. F.A. Graybill, H.K. Iyer. Regression analysis. Concepts and applications. Duxbury Pr., 1994.

Рейтинг и получение зачёта

Подшкалы рейтинга:

1. Присутствие и активность на занятии (30 баллов)
   - Посещение занятие – 2 балла; работа у доски – 1 балл
   - При плохой посещаемости выдаются дополнительные задания
2. Контрольные работы (30 баллов)
   - Погрешности и распределения
   - Метод наименьших квадратов (теория)
   - Глобальная оптимизация и арифметика с плавающей запятой
3. Домашние задания (30 баллов)
   - Погрешности и распределения
   - Основы работы в GNU Octave
   - Метод наименьших квадратов (практика)

Условия получения зачёта без дополнительных заданий:

- Общий балл – не менее 75% от максимума
- Балл по каждой из трёх подшкал – не менее 70% от максимума
- Каждая контрольная и домашняя работа – не менее 60% от максимума

Промежуточный рейтинг:

- октябрь: КР1, Д31, посещаемость
- ноябрь: КР1+КР2, Д31+Д32, посещаемость

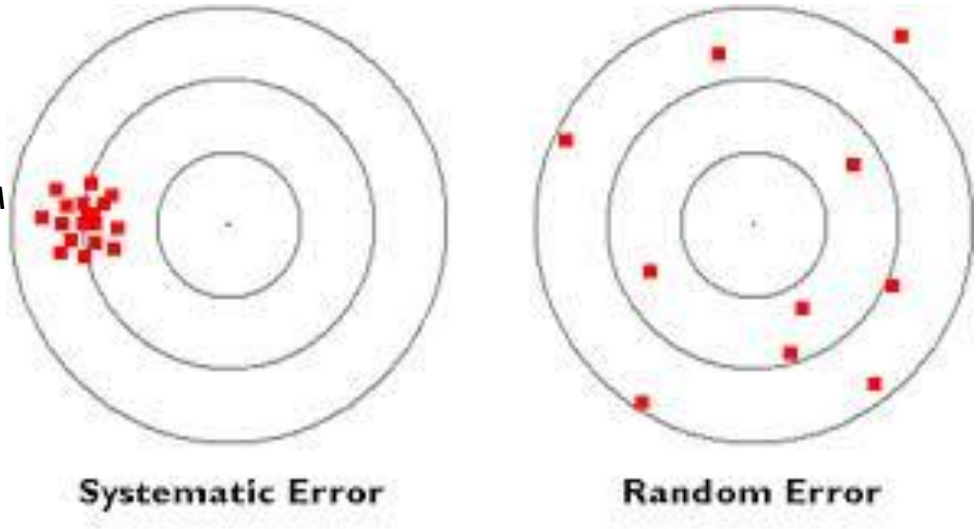
Лекция 1
Погрешности

Виды погрешностей

- **Случайная погрешность** – вызывается большим числом причин в каждом измерении (пример – разброс между результатами титрования)
- **Систематическая погрешность** – обусловлены несовершенством метода измерений (приборы, примеси в реактивах и т.п.)
- **Грубые промахи** – связаны с ошибками экспериментатора (неправильное чтение показаний прибора и т.п.)

Абсолютная погрешность: $\Delta x = |x_{true} - x_{meas}|$- разница между истинным и измеренным значением

Относительная погрешность: $\delta_x = \Delta x / x$



Правила округления

Значащие цифры – все цифры данного числа от первой слева, не равной нулю, до последней справа

Примеры:

- 123 – 3 значащих цифра
- 0.012 – 2 значащих цифры
- $6.022*10^{23}$ – 4 значащих цифры
- $5*10^{3}$ – 1 значащая цифра. **НО: 5000 – 4 значащих цифры!**

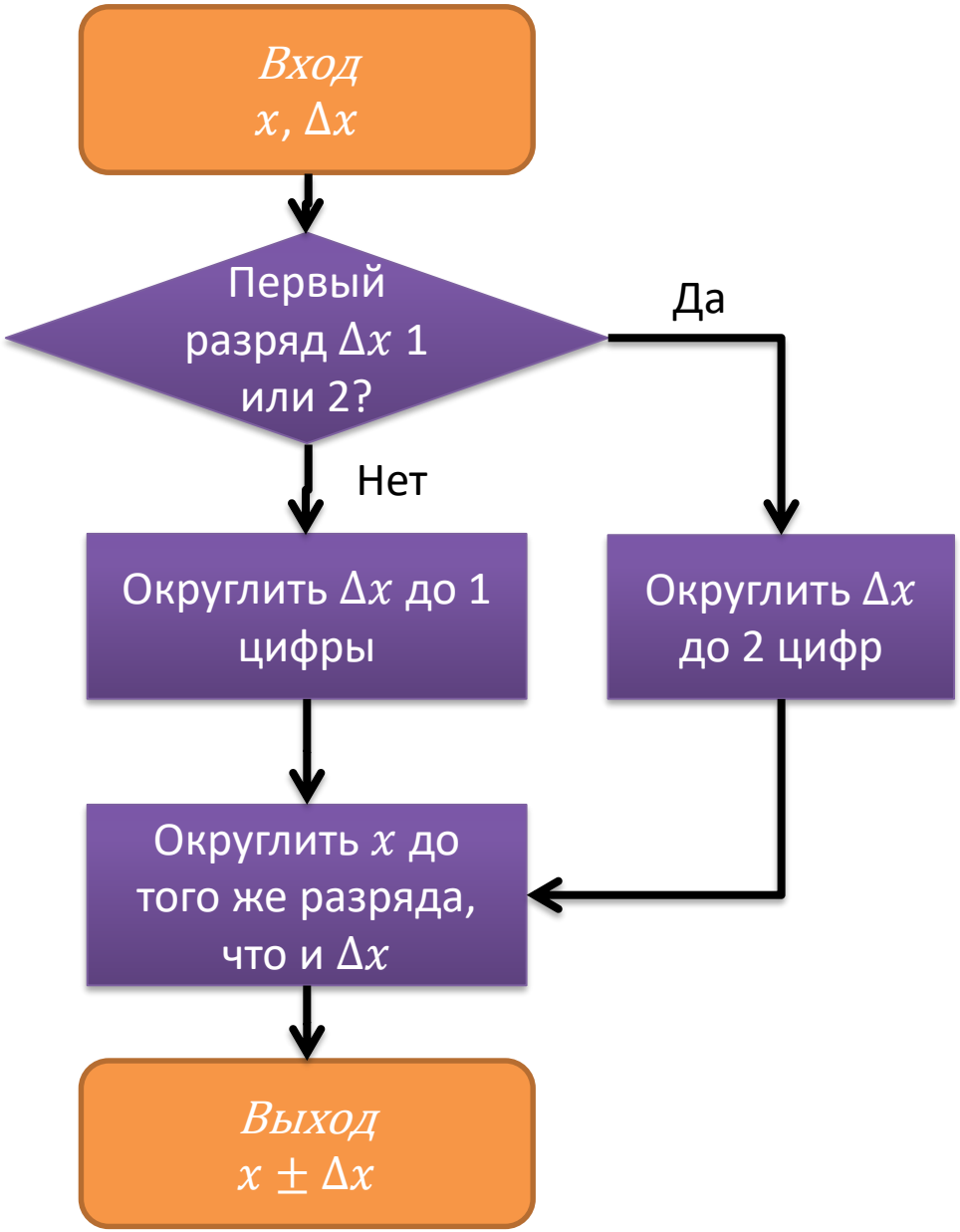
Округление до N-го разряда:

- Если N+1 – ый разряд < 5 – то отбросить все цифры после N-го разряда
- Если N+1 – ый разряд ≥ 5 – то увеличить N-ый разряд на 1 и отбросить все цифры после N-го разряда

Примеры:

- 123 -> 120
- 0.0458 -> 0.05
- 1.95 -> 2.0

Правила округления



Примеры:

- $(53216 \pm 348) \rightarrow (5.32 \pm 0.03) \cdot 10^{4}$
- $(0.0322 \pm 0.012) \rightarrow (3.2 \pm 1.2) \cdot 10^{-2}$
- $(12.482 \pm 0.973) \rightarrow (12.5 \pm 1.0)$

Нельзя округлять:

1. Промежуточные вычисления (потеря точности)
2. Коэффициенты регрессии, полученные МНК (они коррелированы друг с другом)

Сложение погрешностей

Сложение случайных погрешностей при сложении и вычитании:

$$\Delta y = \sqrt{\sum_i (\Delta x_i)^2}$$

Погрешность значения функции:

$$y = f(x_1, \dots, x_n)$$

$$\Delta y = \sqrt{\sum_i \left(\frac{\partial f(\bar{x})}{\partial x_i} \Delta x_i \right)^2}$$

Задачи

1. Вывод формулы для погрешности в случае $y = ab^2$
2. Вывод формулы для погрешности в случае $y = a \ln b$
3. $m_{\text{полн}} = 92.67 \pm 0.05$ г, $m_{\text{тара}} = 52.51 \pm 0.05$ г. Найти массу образца
4. $U = (220 \pm 5)$ В, $I = (4.00 \pm 0.02)$ А. Найти мощность
5. $c[H^+] = (6.3 \pm 0.9) \cdot 10^{-6}$ моль/л. Найти pH

Действие	Погрешность
$y = a + b;$ $y = a - b$	$\Delta y = \sqrt{(\Delta a)^2 + (\Delta b)^2}$
$y = ab;$ $y = a/b$	$\delta_y = \sqrt{\delta_a^2 + \delta_b^2}$
$y = \ln a$	$\Delta y = \delta_a$
$y = a^n$	$\delta_y = n\delta_a$
$y = \sqrt[n]{a}$	$\delta_y = \delta_a / n$

$$\delta_y = \Delta y / y$$

Средние значения и стандартное отклонение

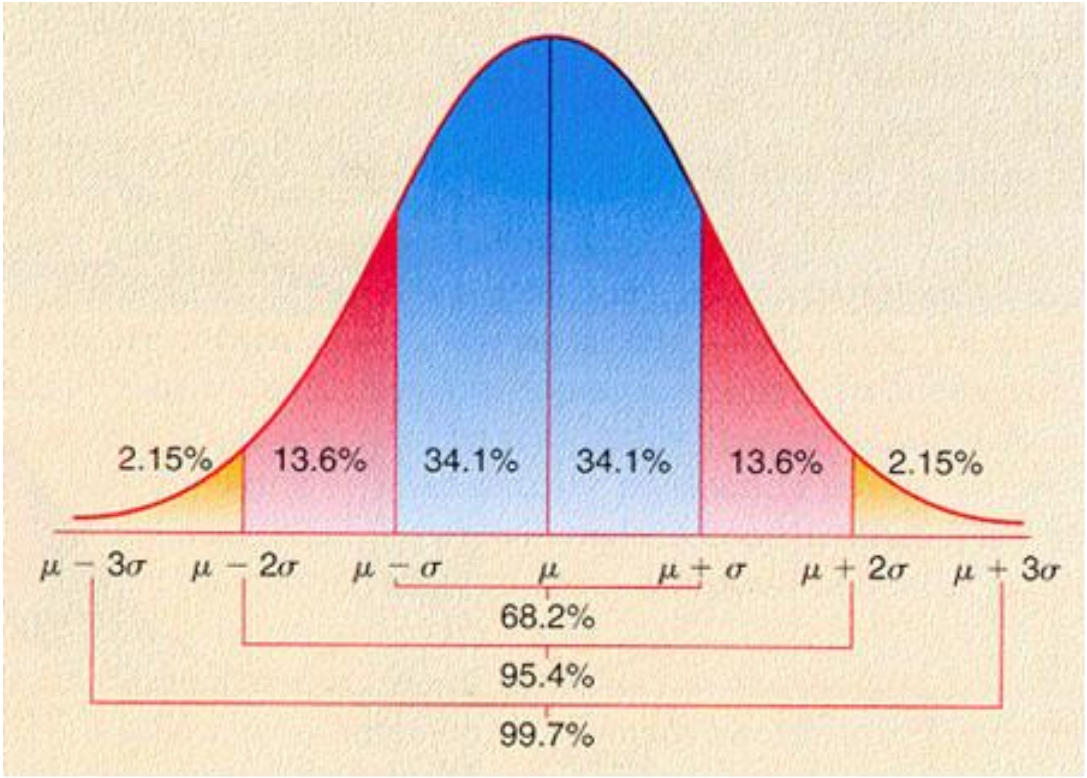
Величина	Формула	Функция MS Excel
Среднее арифметическое	$\bar{x} = \frac{1}{n}\sum_i x_i$	СРЗНАЧ
Среднее геометрическое	$G(x) = \sqrt[n]{\prod_i x_i}$	СРГЕОМ
Медиана	$P(x \le M) = 0.5$	МЕДИАНА
Среднее гармоническое	$A_{-1}(x) = \frac{n}{\sum_i x_i^{-1}}$	СРГАРМ
Стандартное отклонение	$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$	СТАНДОТКЛОН
Среднеквадратичное отклонение	$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$	СТАНДОТКЛОНП
Стандартное отклонение среднего арифметического	$s_x = \frac{s}{\sqrt{n}}$	

Средние значения и стандартное отклонение

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

$$\bar{x} = \frac{1}{n}\sum_i x_i$$

Нормальное распределение



Средние значения и стандартное отклонение

Задача 1. Исходные данные (атмосферное давление):
754, 764, 768, 762, 765, 764, 758, 761, 756, 764 мм рт.ст.
Найти среднее значение, медиану, стандартное отклонение, стандартное отклонение среднего арифметического

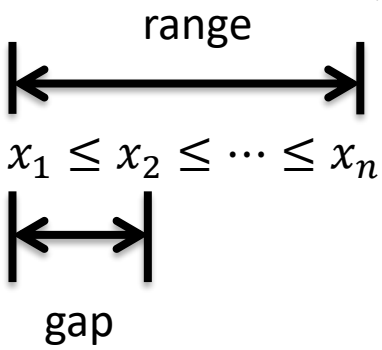
Задача 2. Для выборки 28, 40, 39, 42, 55, 158 найти среднее значение и медиану. Объяснить причину сильного различия между ними.

Задача 3. Установить пакет анализа данных (файл->параметры->настройки->перейти). Сгенерировать массив из 20-50 нормально распределенных случайных чисел. Рассчитать для них среднее значение и стандартное отклонение.

Задача 4*. Сгенерировать массив данных из 500 нормально распределенных случайных чисел. Построить гистограмму с интегральным процентом и без него. Параметры распределения и гистрограммы выбрать самостоятельно

Задача 5*. То же, что и задача 4, но для равномерного распределения

Грубые промахи
Q-критерий (Dixon's q-test)



Особенности:

- Если $Q \geq Q_{tabl}$, то значение – промах
- n = 3-10
- Использовать только один раз для выборки

$$Q = \frac{gap}{range} = \frac{|x_2 - x_1|}{|x_n - x_1|}$$

n	p		
	0.90	0.95	0.99
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

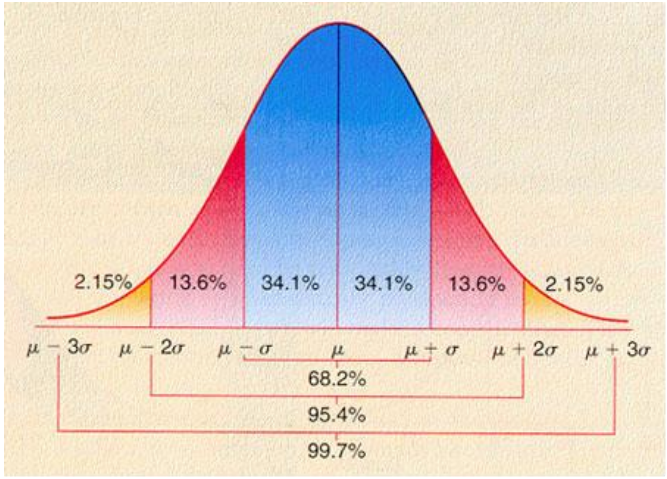
Задача: выявить промах в выборке (p=0.9):

0.189, 0.167, 0.187, 0.183, 0.186, 0.182, 0.181, 0.184, 0.181, 0.177

Грубые промахи
Критерий 3σ

Алгоритм

1. Рассчитать среднее значение
2. Рассчитать стандартное отклонение (исключив предполагаемый промах)
3. Если предполагаемый промах за пределами 3s, то исключить его
4. Применять для *n*=20-100



Задача: найти промах в выборке

8,07	8,06	8,09
8,05	8,04	8,14
8,10	8,11	8,12
8,16	8,09	8,13
8,18	8,14	8,18
8,14	8,11	8,20
8,06	8,15	8,17
8,10	8,16	
8,22	8,50	