# Вероятностный анализ задачи восстановления зависимостей по интервальным данным: оправдание допускового множества решений

В.Я. Крейнович, С.П. Шарый

Университет Техаса в Эль-Пасо, Эль-Пасо, США
ФИЦ ИВТ и НГУ, Новосибирск, Россия

29 сентября 2025 г.

# Резюме работы

В предыдущей статье авторов —

**Kreinovich V., Shary S.P.** Interval methods for data fitting under uncertainty: a probabilistic treatment // Reliable Computing. – 2016. – Vol. 23. – P. 105–141

— показано, что упрощённый вероятностный подход к описанию интервальной неопределённости в задаче восстановления зависимостей приводит к объединённому множеству решений.

Сейчас мы показываем, что реалистичный теоретико-вероятностный анализ задачи делает необходимым привлечение другого известного понятия — допускового множества решений интервальных уравнений.

# Общая формулировка задачи

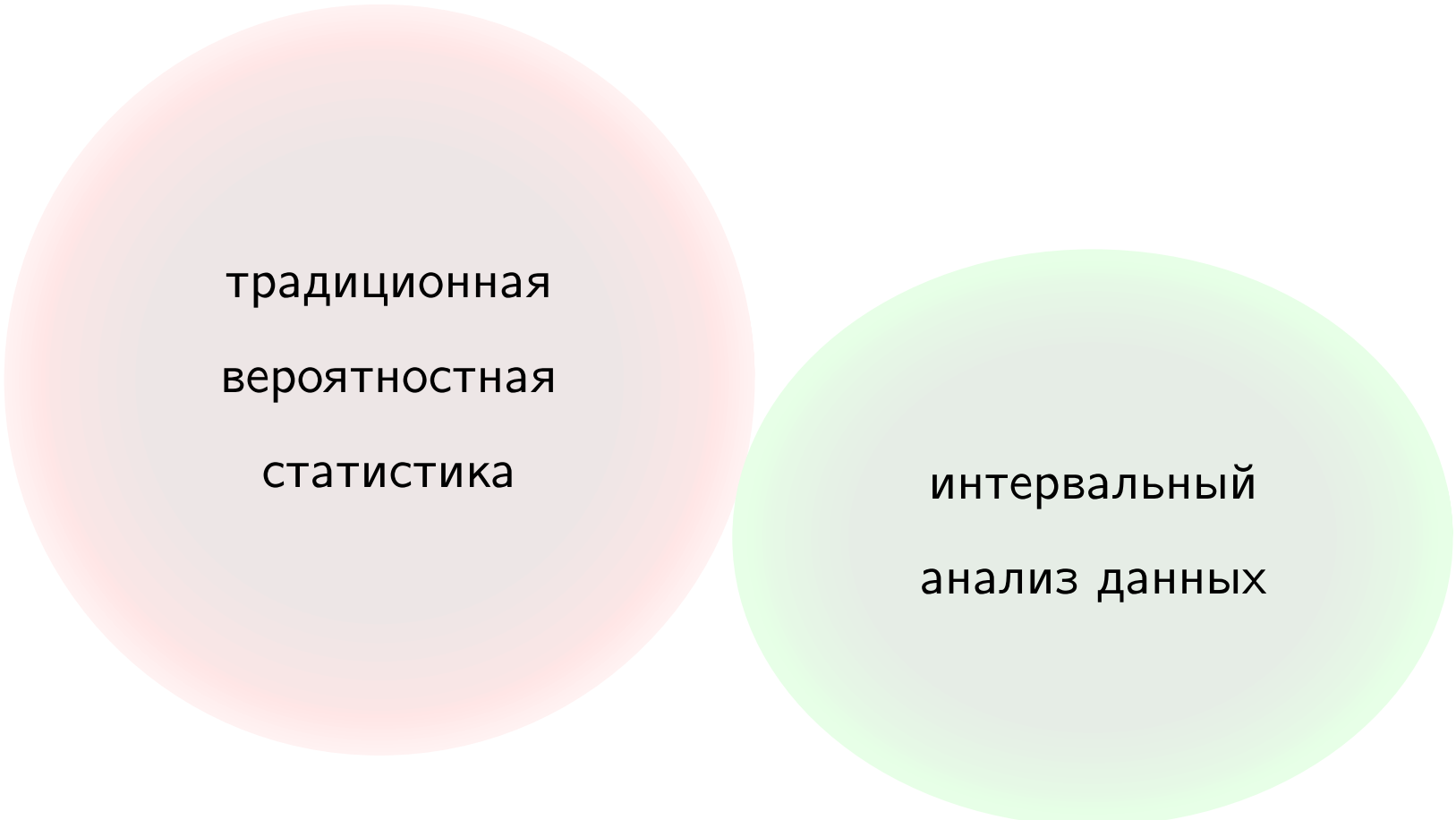
При обработке неточных данных практики часто используют подходы и методы, основанные на вероятностной модели неопределённости.

Как эти методы соотносятся с методами интервального анализа данных в случае интервального описания неопределённости?


- Bounding approaches to system identification / Milanese M., Norton J., Piet-Lahanier H., Walter E. (eds.) – New York: Springer, Plenum Press, 1996.
- **Jaulin L., Kieffer V., Didrit O., Walter E.** Applied interval analysis, with examples in parameter and state estimation, robust control, and robotics. – Springer, London, 2001.
- **Баженов А.Н., Жилин С.И., Кумков С.И., Шарый С.П.** Обработка и анализ интервальных данных. – Издательство «ИКИ»: Ижевск-Москва, 2024.

# Взаимное отношение

. . . интервального анализа данных и традиционной статистики?



# Задача восстановления зависимости

Во многих ситуациях

- мы знаем общий вид $y = f(x, c)$ функциональной зависимости вещественной величины $y$ от величин $(x_1, \ldots, x_n) = x$, но
- мы не знаем точных значений параметров $c = (c_1, \ldots, c_m)$, которые определяют конкретную функцию.

## Примеры

- линейная зависимость

$$y = c_0 + c_1 x_1 + \ldots + c_n x_n.$$

- общая квадратичная зависимость

$$y = \sum_{i,j} c_{ij} x_i x_j.$$

- полином из экспонент (в частности, для радиоактивного распада)

$$y = \sum_{i=1}^{m} c_{2i-1} \exp(-c_{2i}\, t),$$

- и т. д.

# Задача восстановления зависимости

Для нахождения $c_i$ измеряем $x_1$, $x_2$, …, $x_n$ и $y$ несколько раз $(K)$.

На основе результатов измерений

$$\widetilde{x}^{(k)} = (\widetilde{x}_1^{(k)}, \widetilde{x}_2^{(k)}, \ldots, \widetilde{x}_n^{(k)}) \quad \text{и} \quad \tilde{y}^{(k)}, \qquad k = 1, 2, \ldots, K,$$

необходимо оценить значения параметров, которые «согласуются» («совместны» и т. п.) с данными, или даже наилучшим образом.

Подставляя их в выражение для искомой функции, получим

$$\begin{cases} f(\widetilde{x}_1^{(k)}, \widetilde{x}_2^{(k)}, \ldots, \widetilde{x}_n^{(k)}, c_1, c_2, \ldots, c_m) = \tilde{y}^{(k)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad k = 1, 2, \ldots, K, \end{cases}$$

— система уравнений для определения $c_i$.

# Задача восстановления зависимости

Измерения никогда не являются абсолютно точными.

Как следствие, необходимо принимать во внимание, что результаты измерений $\widetilde{v}$, вообще говоря, отличаются от истинных значений $v$, т. е. существует ненулевая измерительная погрешность $\Delta v := \widetilde{v} - v$.

Соответствующие погрешности измерений нужно учитывать при оценивании параметров $c_i$ искомой функции $f$.

# Модель неточностей и погрешностей в данных

Данные почти всегда неточны . . .

*Какую модель неопределённости данных мы принимаем?*

Традиционный выбор — теоретико-вероятностная модель
(К.Ф. Гаусс, П.С. Лаплас и т. д.):

ошибки измерений и наблюдений — это случайные величины теории вероятностей с (более-менее) известными характеристиками

# Интервальная модель погрешностей

*Неопределённости и неточности в данных*

*часто удобнее описывать интервально*

Даны интервальные оценки величин, т. е. принадлежности $x_i^{(k)}$ и $y^{(k)}$ некоторым интервалам:

$$x_i^{(k)} \in \boldsymbol{x}_i^{(k)} = [\underline{\boldsymbol{x}}_i^{(k)}, \overline{\boldsymbol{x}}_i^{(k)}] \qquad \text{и} \qquad y^{(k)} \in \boldsymbol{y}_i = [\underline{\boldsymbol{y}}^{(k)}, \overline{\boldsymbol{y}}^{(k)}]$$

т. е.

$$\underline{\boldsymbol{x}}_i^{(k)} \leq x_i^{(k)} \leq \overline{\boldsymbol{x}}_i^{(k)} \qquad \text{и} \qquad \overline{\boldsymbol{y}}^{(k)} \leq y^{(k)} \leq \underline{\boldsymbol{y}}^{(k)}$$

# Формулировка задачи

Интервальные данные —

$$\begin{array}{cccc} \boldsymbol{x}_1^{(1)}, & \boldsymbol{x}_2^{(1)}, & \dots & \boldsymbol{x}_n^{(1)}, & \boldsymbol{y}^{(1)}, \\ \boldsymbol{x}_1^{(2)}, & \boldsymbol{x}_2^{(2)}, & \dots & \boldsymbol{x}_n^{(2)}, & \boldsymbol{y}^{(2)}, \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{x}_1^{(K)}, & \boldsymbol{x}_2^{(K)}, & \dots & \boldsymbol{x}_n^{(K)}, & \boldsymbol{y}^{(K)}. \end{array}$$

Подставляя их в выражение для искомой функции, получим

$$\begin{cases} f(\boldsymbol{x}_1^{(k)}, \boldsymbol{x}_2^{(k)}, \dots, \boldsymbol{x}_n^{(k)}, c_1, c_2, \dots, c_m) = \boldsymbol{y}^{(k)}, \\ k = 1, 2, \dots, K, \end{cases} \qquad (\blacklozenge)$$

— система уравнений для определения $c_i$.

# Интервальная система уравнений

Но решения и множества решений для интервальных уравнений и систем уравнений могут определены очень многими способами, смысл которых — существенно разный.

Наиболее популярно объединённое множество решений, которое образовано всевозможными решениями точечных систем уравнений, содержащимися в интервальной системе:

$$\Xi_{\text{uni}} := \Big\{ c \in \mathbb{R}^m \mid (\exists x_:^{(k)} \in \boldsymbol{x}_:^{(k)})(\exists y^{(k)} \in \boldsymbol{y}^{(k)})\ f(x_:^{(k)}, c) = y^{(k)}, \quad k = 1, 2, \ldots, K \Big\}$$

# Интервальная система уравнений

Кроме него существуют также другие важные множества решений, в частности, допусковое множество решений:

$$\Xi_{\text{tol}} := \Big\{ c \in \mathbb{R}^m \mid (\forall x_:^{(k)} \in \boldsymbol{x}_:^{(k)})(\exists y^{(k)} \in \boldsymbol{y}^{(k)})\ f(x_:^{(k)}, c) = y^{(k)}, \quad k = 1, 2, \ldots, K \Big\}$$

Оно образовано теми решениями точечных систем уравнений, для которых значение отображения $f$ попадает в интервалы правой части при любых значениях параметров $x_:^{(k)} \in \boldsymbol{x}_:^{(k)}$.

# Допусковое множество решений


- **Nuding E., Wilhelm J.** Über Gleichungen und über Lösungen // ZAMM. – 1972. – Bd. 52. – S. T188–T190.
- **Neumaier A.** Tolerance analysis with interval arithmetic // Freiburger Intervall-Berichte. – 1986. – No. 6. – S. 5–19.
- **Deif A.** Sensitivity analysis in linear systems. – Berlin: Springer, 1986.


Ранняя история исследования этого множества

— в работе А. Ноймайера: «внутренние решения».

Большое внимание уделено ему в книге А. Дейфа, но предложенные в ней методы исследования этого множества носят частный характер.

# Допусковое множество решений

**Rohn J.** Inner solutions of linear interval systems // Interval Mathematics 1985 / Nickel K., ed. – New York: Springer Verlag, 1986. – P. 157–158. – (Lecture Notes in Computer Science; vol. 212).

— представление допускового множества решений через решение системы линейных неравенств.

Можно исследовать пустоту/непустоту допускового множества решений с помощью методов линейного программирования, которые имеют полиномиальную трудоёмкость.

# Допусковое множество решений

**Shary S.P.** Solving interval linear tolerance problem // Mathematics and Computers in Simulation. – 1995. – Vol. 39. – P. 53–85.

— распознающий функционал допускового множества решений интервальной линейной $m \times n$-системы уравнений $\boldsymbol{A}x = \boldsymbol{b}$:

$$\mathrm{Tol}(x, \boldsymbol{A}, \boldsymbol{b}) = \min_{1 \le i \le m} \left\{ \mathrm{rad}\ \boldsymbol{b}_i - \left| \mathrm{mid}\ \boldsymbol{b}_i - \sum_{j=1}^{n} \boldsymbol{a}_{ij} x_j \right| \right\}$$

# Допусковое множество решений

**Шарый С.П.** Конечномерный интервальный анализ. – Новосибирск: XYZ, 2025. Электронная книга, доступная на http://www.nsc.ru/interval/?page=Library/InteBooks

— глава 6 целиком посвящена допусковому множеству решений для интервальных линейных систем уравнений.

Аналогичные результаты — также в других главах.

# Допусковое множество решений

**Sharaya I.A.** On unbounded tolerable solution sets // Reliable Computing. – 2005. – Vol. 11, No. 5. – P. 425–432.

— выведен простой критерий ограниченности

допускового множества решений.

Оказывается, что допусковое множество решений почти всегда ограниченно, в отличие от объединённого множества решений.

В задачах восстановления зависимостей это означает, что множество параметров, совместных с данными, имеет ограниченную вариабельность (изменчивость).

# Допусковое множество решений

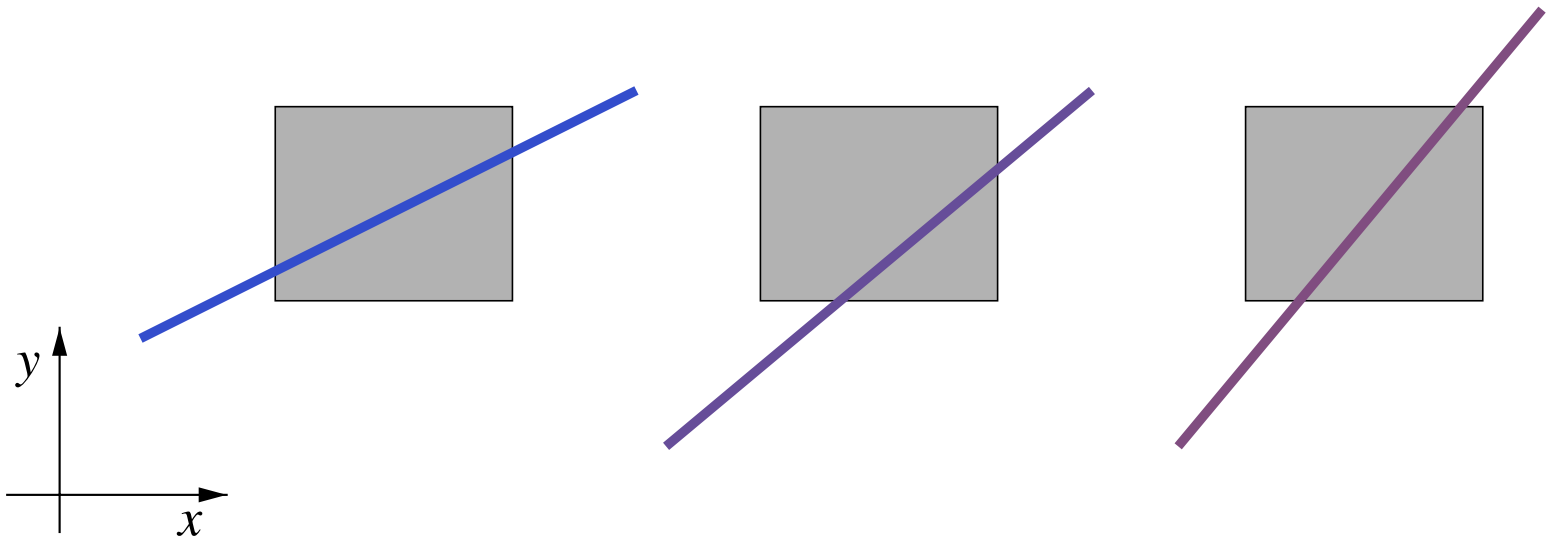
— в задаче восстановления зависимостей по интервальным данным

Шарый С.П. Сильная согласованность в задаче восстановления зависимостей при интервальной неопределённости данных // Вычислительные Технологии. – 2017. – Т. 22, №2. – С. 150–172.

Shary S.P. Weak and strong compatibility in data fitting problems under interval uncertainty // Advances in Data Science and Adaptive Analysis. 2020. Vol. 12, No. 1. Paper 2050002.

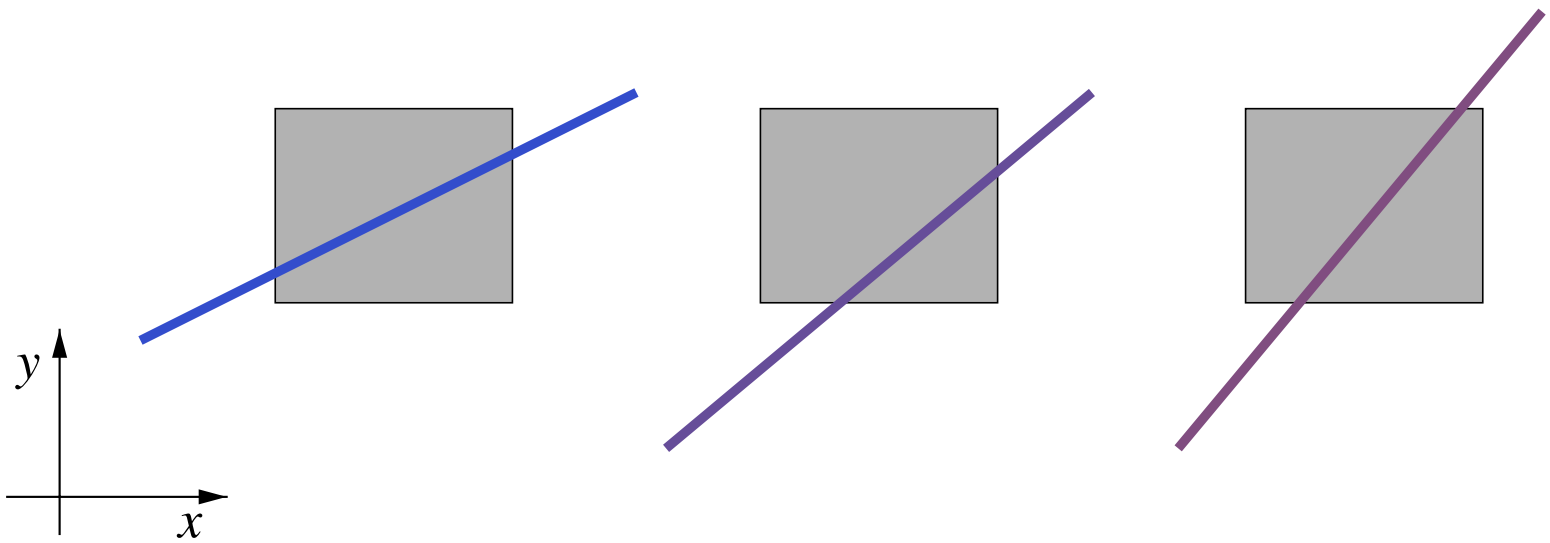
# Допусковое множество решений

— в задаче восстановления зависимостей по интервальным данным



# Допусковое множество решений

— в задаче восстановления зависимостей по интервальным данным




Покажем, что допусковое множество решений

также имеет вероятностный смысл.

# Обработка вероятностной неопределённости

Во многих ситуациях погрешность в данных может быть адекватно описана как «случайная величина» теории вероятностей, а её распределение вероятностей более-менее известно.

Будем предполагать, что погрешности измерений, соответствующие различным переменным, независимы в смысле теории вероятностей.

Пусть распределения, описывающие случайные погрешности, абсолютно непрерывны и имеют плотности вероятности

- $p_i(\Delta x_i)$ для независимых переменных $x_i$
- $q(\Delta y)$ для зависимой переменной $y$.

# Метод максимума правдоподобия

— один из наиболее популярных методов оценивания параметров:

> оценкой параметров берётся значение, на котором достигается максимум «функции правдоподобия».

Обычно функция правдоподобия равна

- вероятности оценки
  (для дискретных распределений),
- плотности вероятности оценки
  (для абсолютно непрерывных распределений).

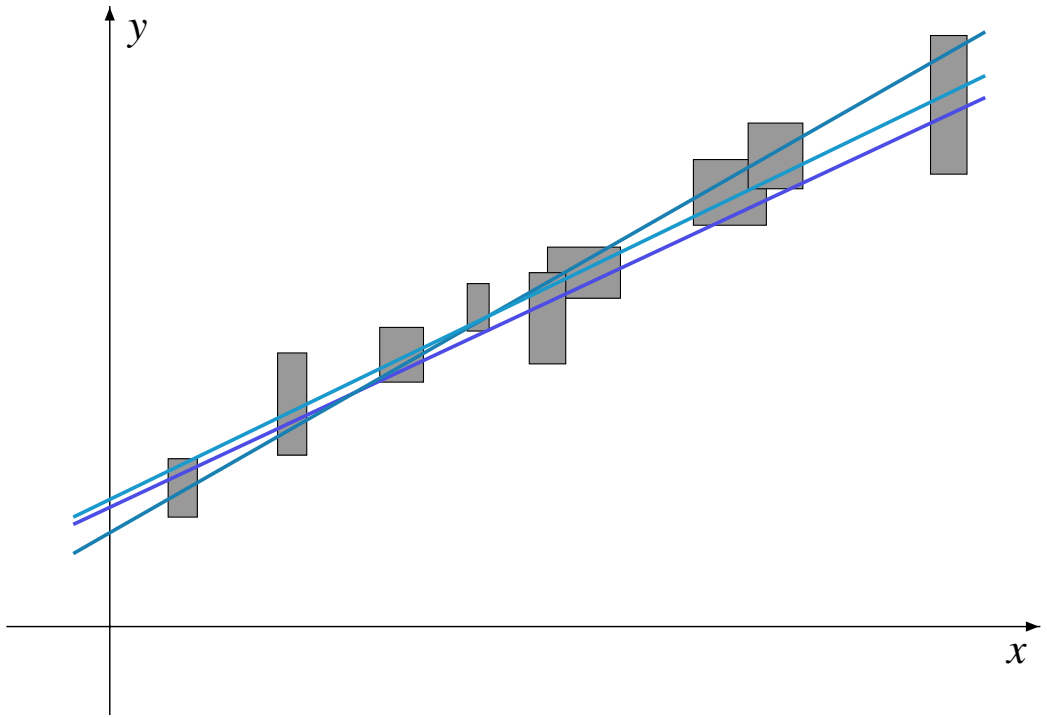
# Предыдущий результат авторов

**Kreinovich V., Shary S.P.** Interval methods for data fitting under uncertainty: a probabilistic treatment // Reliable Computing. – 2016. – Vol. 23. – P. 105–141

Свободно доступна на веб-сайте журнала по ссылке —

`https://interval.louisiana.edu/reliable-computing-journal/volume-23/reliable-computing-23-pp-105-140.pdf`

# Метод максимума совместности




## Результат

Метод максимума совместности в слабой версии является не чем иным как методом максимального правдоподобия при равномерных вероятностных распределениях на интервалах данных.

# Метод максимума правдоподобия

В нашем случае оценкой $c$ берётся вектор значений параметров, для которых плотность вероятности оценки, равная

$$\prod_{k=1}^{K} \left( q\left(\widetilde{y}^{(k)} - f\left(x^{(k)}, c\right)\right) \cdot \prod_{i=1}^{n} p_i\left(\widetilde{x}_i^{(k)} - x_i^{(k)}\right) \right),$$

достигает наибольшего значения.

# Метод максимума правдоподобия

Вместо максимизации правдоподобия обычно решают равносильную задачу максимизации его логарифма:

$$\sum_{k=1}^{K} \left( \log q\big(\widetilde{y}^{(k)} - f(x^{(k)}, c)\big) + \sum_{i=1}^{n} \log p_i\big(\widetilde{x}_i^{(k)} - x_i^{(k)}\big) \right) =$$

$$= \sum_{k=1}^{K} \log q\big(\widetilde{y}^{(k)} - f(x^{(k)}, c)\big) + \sum_{k=1}^{K} \sum_{i=1}^{n} \log p_i\big(\widetilde{x}_i^{(k)} - x_i^{(k)}\big).$$

Этот шаг имеет технические причины:

с суммами гораздо удобнее работать, чем с произведениями.

Вспомним ЗБЧ, УЗБЧ, ...

# Метод максимума правдоподобия

$$\sum_{k=1}^{K} \log q\big(\widetilde{y}^{(k)} - f(x^{(k)}, c)\big) + \sum_{k=1}^{K} \sum_{i=1}^{n} \log p_i\big(\widetilde{x}_i^{(k)} - x_i^{(k)}\big)$$

Второе слагаемое не зависит от $c$,

так что его можно опустить при поиске максимума по $c$.

В целом необходимо найти наибольшее значение для

$$\sum_{k=1}^{K} \log q\big(\widetilde{y}^{(k)} - f(x^{(k)}, c)\big),$$

а также доставляющий его вектор аргумента $c$.

# Обработка интервальных данных

Часто мы не знаем вероятностное распределение погрешностей, а нередко даже нельзя быть уверенным в том, что погрешности подчиняются закономерностям теории вероятностей.

Предположим, что погрешности измерений $\Delta v$ локализованы на заданном интервале $[-\Delta_v, \Delta_v]$.

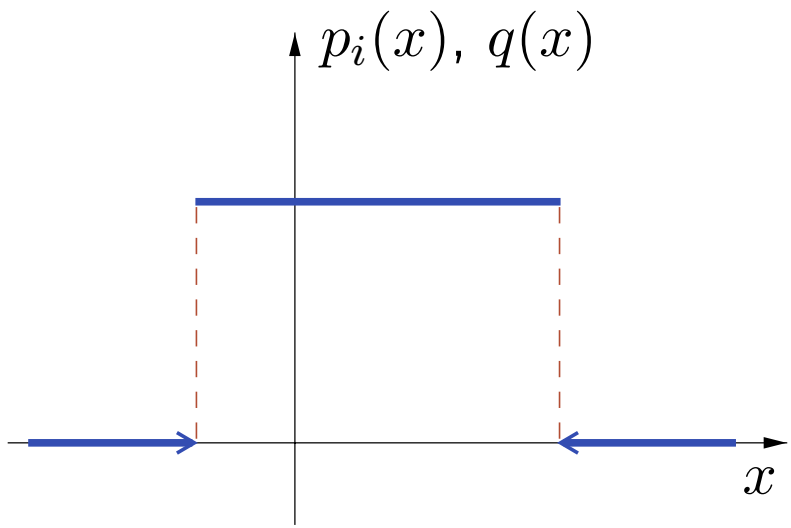
Принимаем также гипотезу, что интервалы измерений и наблюдений являются накрывающими (включающими), т. е. содержат истинные значения измеряемых величин.

# Обработка интервальных данных

Чтобы выполнить вероятностный анализ обработки интервальных данных, нужно задать какое-то распределение на интервалах.

Естественно взять то, что имеет наибольшую энтропию, т. е. даёт наименьшую возможную информацию о величине.

Им оказывается равномерное вероятностное распределение.



# Обработка интервальных данных

Простейший случай — точное измерение значений $x_i$,

и тогда $\widetilde{x}_i^{(k)} = x_i^{(k)}$ для всех $i$ и $k$.

В нашей предыдущей статье показано, что метод максимума правдоподобия выбирает тогда следующее множество:

*Множество всевозможных значений $c$, для которых справедливо $f(x_:^{(k)}, c) \in \boldsymbol{y}^{(k)}$ при любых $k$.*

Вероятностный подход приводит к тому же самому ответу, что и в интервальном анализе данных, где это множество называется *множеством решений*.

# Обработка интервальных данных

Но для интервальных уравнений и систем уравнений решения и множества решений могут быть определены разными способами . . .

Интервальная неопределённость имеет двойственный характер

Существуют объединённое множество решений, допусковое множество решения, множества AE-решений, формальные решения и т. д. . . .

В описанной выше ситуации главные множества решений — объединённое и допусковое — совпадают друг с другом, так что можно говорить о «множестве решений» вообще.

# Метод максимума правдоподобия

В общем случае $x_i^{(k)}$ также имеют интервальную неопределённость.

Тогда при определённых условиях метод максимума правдоподобия выбирает, как было показано,

*множество всех таких* $c$, *для которых* $f(x_:^{(k)}, c) \in \boldsymbol{y}^{(k)}$ *при некоторых* $x_i^{(k)} \in \boldsymbol{x}_i^{(k)}$, $i = 1, 2, \ldots, n$, *и всяком* $k$.

# Метод максимума правдоподобия

В общем случае $x_i^{(k)}$ также имеют интервальную неопределённость.

Тогда при определённых условиях метод максимума правдоподобия выбирает, как было показано,

> *множество всех таких* $c$, *для которых* $f(x_:^{(k)}, c) \in \boldsymbol{y}^{(k)}$ *при некоторых* $x_i^{(k)} \in \boldsymbol{x}_i^{(k)}$, $i = 1, 2, \ldots, n$, *и всяком* $k$.

Это *объединённое множество решений* интервальной системы уравнений

$$\begin{cases} f(\boldsymbol{x}_:^{(k)}, c) = \boldsymbol{y}^{(k)}, \\ \qquad k = 1, 2, \ldots, K, \end{cases} \qquad (\blacklozenge)$$

построенной по данным и виду функции.

Объединённое множество решений имеет вероятностное толкование!

# Более реалистичное описание практической задачи

Нередко при измерениях тех или иных величин

практикуют *повторные измерения*, т. е. выполняемые неоднократно.

Для измерения величин должны быть проведены несколько измерений, по результатам которых формируется итоговое значение величины.

Это часто практикуема процедура, которую выполняют

для повышения общей надёжности процесса измерения.

# Более реалистичное описание практической задачи

## Примеры

- При измерении давления крови тонометром рекомендуется выполнить его 3 (три) раза, и затем результаты усредняются, если они «не сильно разнятся».
- Сверхточные атомные часы обычно состоят из нескольких независимых часов, чьи показатели выдаются пользователю, если большинство их результатов совпадают.
- Именно с помощью повторных измерений вводятся новые значения, будь это уточнённое расстояние до Луны или новое значение атомного веса элемента.

# Более реалистичное описание практической задачи

**Примеры**

- При измерении давления крови тонометром рекомендуется выполнить его 3 (три) раза, и затем результаты усредняются, если они «не сильно разнятся».
- Сверхточные атомные часы обычно состоят из нескольких независимых часов, чьи показатели выдаются пользователю, если большинство их результатов совпадают.
- Именно с помощью повторных измерений вводятся новые значения, будь это уточнённое расстояние до Луны или новое значение атомного веса элемента.

Результат единичного измерения не вполне надёжен, так что для повышения достоверности всегда выполняют несколько измерений, которые дают финальный результат, если «достаточно близки».

# Использование повторных измерений

- Например, можно считать, что окончательный интервал результата многократного измерения сформирован операцией агрегирования индивидуальных результатов измерений:

$$\boldsymbol{x}_i^{(k)} = \left[ \min_{\ell} x_i^{(k\ell)}, \max_{\ell} x_i^{(k\ell)} \right], \qquad \boldsymbol{y}^{(k)} = \left[ \min_{\ell} y^{(k\ell)}, \max_{\ell} y^{(k\ell)} \right].$$

- Можно считать, что общий интервал результата берётся как среднее отдельных измерений, раздутое на несколько стандартных отклонений.

- И т. д.

Для любого измерения $k$ вместо единственного набора $x_{:}^{(k)}$

имеем несколько $x_{:}^{(k\ell)} = (\, x_1^{(k\ell)}, x_2^{(k\ell)}, \ldots, x_n^{(k\ell)})$ для различных $\ell$.

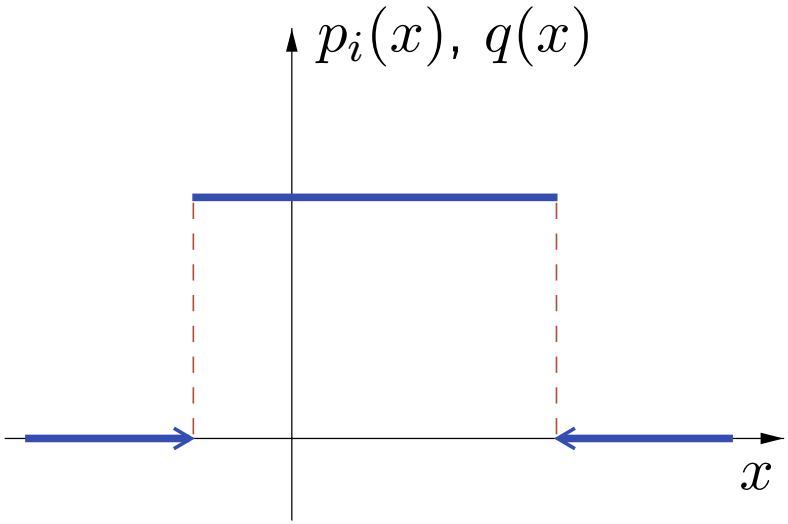
Для каждой комбинации значений $x_i^{(k\ell)} \in \boldsymbol{x}_i^{(k)}$

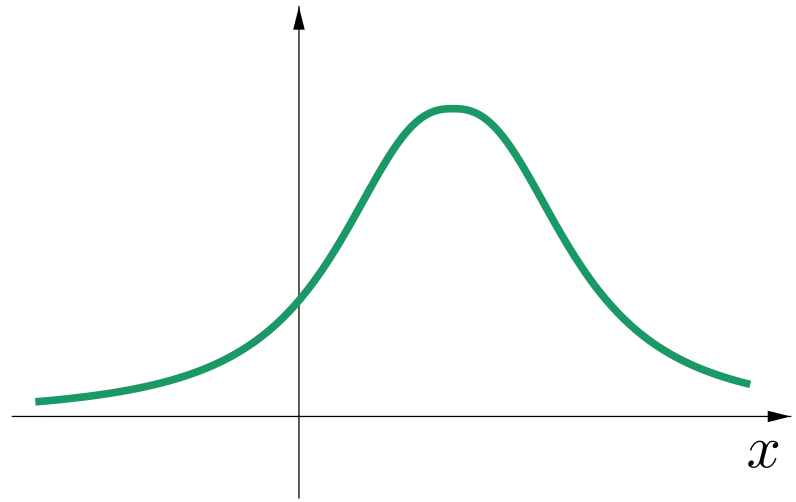
можем сформировать логарифмическое правдоподобие

$$\sum_{k=1}^{K} \sum_{\ell} \log q\big(\widetilde{y}^{(k)} - f(x_{:}^{(k\ell)}, c)\big).$$

Мы не знаем точно $x_i^{(k\ell)}$,

но они лежат в интервалах $\boldsymbol{x}_i^{(k)}$ и равномерно распределены.




Равномерные распределения имеют ограниченный носитель, хотя у типичных вероятностных распределений они неограниченны.

# Какова итоговая оценка?

**Предложение**

Всевозможные оценки параметров, полученные методом максимума правдоподобия, образуют

*Множество всех таких* $c$, *что* $f(x_{:}^{(k\ell)}, c) \in \boldsymbol{y}^{(k)}$ *для любых* $x_i^{(k\ell)} \in \boldsymbol{x}_i^{(k)}$, $i = 1, 2, \ldots, n$, *и всяком* $k$.

## Доказательство

Если для некоторого $c$ условие

$$f\big(x_{:}^{(k\ell)}, c\big) \in \boldsymbol{y}^{(k)}$$

не удовлетворено при каких-то $x_{:}^{(k\ell)} \in \boldsymbol{x}_{:}^{(k)}$, то

$$q\big(\widetilde{y}^{(k)} - f\big(x_{:}^{(k\ell)}, c\big)\big) = 0,$$

и тогда $\log 0 = -\infty$.

Поэтому в сумме

$$\sum_{k=1}^{K} \sum_{\ell} \log q\big(\widetilde{y}^{(k)} - f\big(x_{:}^{(k\ell)}, c\big)\big).$$

появляются слагаемые $(-\infty)$, тогда как остальные слагаемые конечны.

## Доказательство

Если для некоторого $c$ условие

$$f\big(x_{:}^{(k\ell)}, c\big) \in \boldsymbol{y}^{(k)}$$

не удовлетворено при каких-то $x_{:}^{(k\ell)} \in \boldsymbol{x}_{:}^{(k)}$, то

$$q\big(\widetilde{y}^{(k)} - f\big(x_{:}^{(k\ell)}, c\big)\big) = 0,$$

и тогда $\log 0 = -\infty$.

Поэтому в сумме

$$\sum_{k=1}^{K} \sum_{\ell} \log q\big(\widetilde{y}^{(k)} - f\big(x_{:}^{(k\ell)}, c\big)\big).$$

появляются слагаемые $(-\infty)$, тогда как остальные слагаемые конечны.

$\Rightarrow$ Вся сумма $= -\infty$, так что её конечный максимум не достигается.

## Доказательство (продолжение)

Для всякого $c$, выбираемого методом максимума правдоподобия,

в самом деле должны иметь $f(x_{:}^{(k\ell)}, c) \in \boldsymbol{y}_{:}^{(k)}$ для всех $x_{:}^{(k\ell)} \in \boldsymbol{x}_{:}^{(k)}$.

Так как рассматриваем равномерные распределения на $\boldsymbol{y}^{(k)}$, $k = 1, 2, \ldots, K$, то всем значениям в пределах этих интервалов соответствуют одни и те же значения плотности.

Тогда для всех таких кортежей $c$ будем иметь

одинаковые значения слагаемых $\log q\big(\widetilde{y}^{(k)} - f(x_{:}^{(k\ell)}, c)\big)$.

Поэтому все такие кортежи $c$

будут выделены методом максимума правдоподобия.

# Итоги

В целом формула

*Множество всех таких* $c$, *что* $f(x_{:}^{(k\ell)}, c) \in \boldsymbol{y}^{(k)}$ *при любых* $x_i^{(k\ell)} \in \boldsymbol{x}_i^{(k)}$, $i = 1, 2, \ldots, n$, *и всяком* $k$

определяет *допусковое множество решений* для интервальной системы уравнений (♦), построенной по функции $f$ и интервальным данным для отыскания оценок параметров $c$.

Допусковое множество решений

также получает вероятностный смысл.

# Итоги

Дополнительные бонусы —

меньшая трудоёмкость и конечная вариабельность:

- в отличие от объединённого множества решений, для которого распознавание и оценивание является NP-трудным даже если $f(x, c)$ — линейная функция от $c$,
- вычисление допускового множества решений для линейной функции $f(x, c)$ сводится к решению системы линейных неравенств и потому имеет полиномиальную трудоёмкость.

# Итоги

Дополнительные бонусы —

меньшая трудоёмкость и конечная вариабельность:

- в отличие от объединённого множества решений, для которого распознавание и оценивание является NP-трудным даже если $f(x, c)$ — линейная функция от $c$,
- вычисление допускового множества решений для линейной функции $f(x, c)$ сводится к решению системы линейных неравенств и потому имеет полиномиальную трудоёмкость.
- оценки параметров с помощью допускового множества решений почти всегда имеют конечную вариабельность в силу критерия ограниченности И.А. Шарой.

# Спасибо за внимание!