ВЕРОЯТНОСТНЫЕ КОНСТРУКЦИИ В ИНТЕРВАЛЬНОМ АНАЛИЗЕ ДАННЫХ

С.П. Шарый

ФИЦ ИВТ и НГУ, Новосибирск

29 октября 2024 г.

Интервальный анализ данных

= анализ интервальных данных

= анализ данных с интервальными неопределённостями

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 a_{ij} и b_i некоторым интервалам:

$$a_{ij} \in \boldsymbol{a}_{ij} = [\underline{\boldsymbol{a}}_{ij}, \overline{\boldsymbol{a}}_{ij}] \qquad \text{и} \qquad b_i \in \boldsymbol{b}_i = [\underline{\boldsymbol{b}}_i, \overline{\boldsymbol{b}}_i]$$

т. е.

$$\underline{\boldsymbol{a}}_{ij} \leq a_{ij} \leq \overline{\boldsymbol{a}}_{ij} \qquad \text{и} \qquad \overline{\boldsymbol{b}}_i \leq b_i \leq \underline{\boldsymbol{b}}_i$$

Леонид Витальевич Канторович (1912–1986)

пионер нового подхода,

действительный член АН СССР,

основатель кафедры
вычислительной математики НГУ,

лауреат Сталинской премии,

лауреат Ленинской премии,

лауреат Нобелевской премии
по экономике

СИБИРСКИЙ МАТЕМАТИЧЕСКИЙ ЖУРНАЛ

Том III, № 5 Сентябрь — Октябрь 1962 г.

Л. В. КАНТОРОВИЧ

О НЕКОТОРЫХ НОВЫХ ПОДХОДАХ К ВЫЧИСЛИТЕЛЬНЫМ МЕТОДАМ И ОБРАБОТКЕ НАБЛЮДЕНИЙ*.

Введение

Имевшие место сдвиги в развитии математики и вычислительных средств должны иметь следствием коренные изменения в технике, а возможно и теории численных методов и обработки наблюдений. В той или иной форме отдельные высказываемые ниже соображения встречались в литературе, но не разрабатывались систематически. В частности, мы считаем, что существенное значение имеют следующие моменты:

1. Бóльшая ответственность за результаты расчетов, на которых сейчас нередко базируются решения, касающиеся сложных дорогостоящих объектов современной физики и техники, наличие больших не наблюдаемых этапов при машинных вычислениях повышают требования к надежности окончательных и промежуточных данных, получаемых в процессе применения численных методов и при обработке данных наблюдений. Это обуславливает систематический переход от построения приближенных значений и результатов, к получению точных двухсторонних границ для искомых величин или, если говорить о нечисловых величинах, областей расположения искомых и наблюдаемых величин;

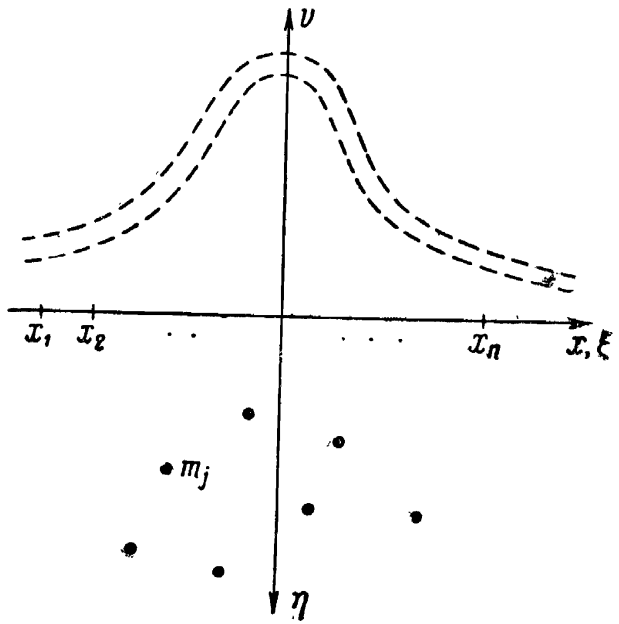
§ 3. Некоторые задачи прикладной математики

1. Задача обработки наблюдений. Обычно полученную в результате измерений избыточную систему уравнений обрабатывают по методу наименьших квадратов Гаусса. При этом происходит значительная потеря информации. По-видимому, в настоящее время более целесообразна другая техника. Уравнения, связывающие искомые величины, выписать с учетом погрешностей в форме неравенств

$$l_i - \delta \leqslant \sum_{k=1}^{n} c_{ik} x_k \leqslant l_i + \delta,$$

$$i = 1, \ldots, m,$$

и разыскивать возможные границы для x_k методами линейного программирования.



Черт. 1

2. Обратная задача теории потенциала*. По измеренным значениям гравитационного, магнитного или иного потенциала в ряде точек (рассматриваем для простоты плоский случай) нужно дать заключение о рудном теле, нарушающем поле.

Пусть m_j — сосредоточенная неизвестная масса в точке

$$(\xi_j, \eta_j), \quad j = 1, \ldots, n.$$

Тогда теоретически вычисленное значение потенциала в точке x_i будет

$$v_i^0 = \sum_{j=1}^{n} \frac{\gamma \eta_j}{(x_i - \xi_j)^2 + \eta_j^2} m_j = \sum_{j=1}^{n} a_{ij} m_j,$$

$$i = 1, \ldots, m,$$

Анализ данных с интервальной неопределнностью

Л.В. Канторович

О некоторых новых подходах к вычислительным методам и обработке наблюдений

// Сибирский математический журнал. – 1962.

– Том 3, №5. – С. 701–709.

Работа была продолжена

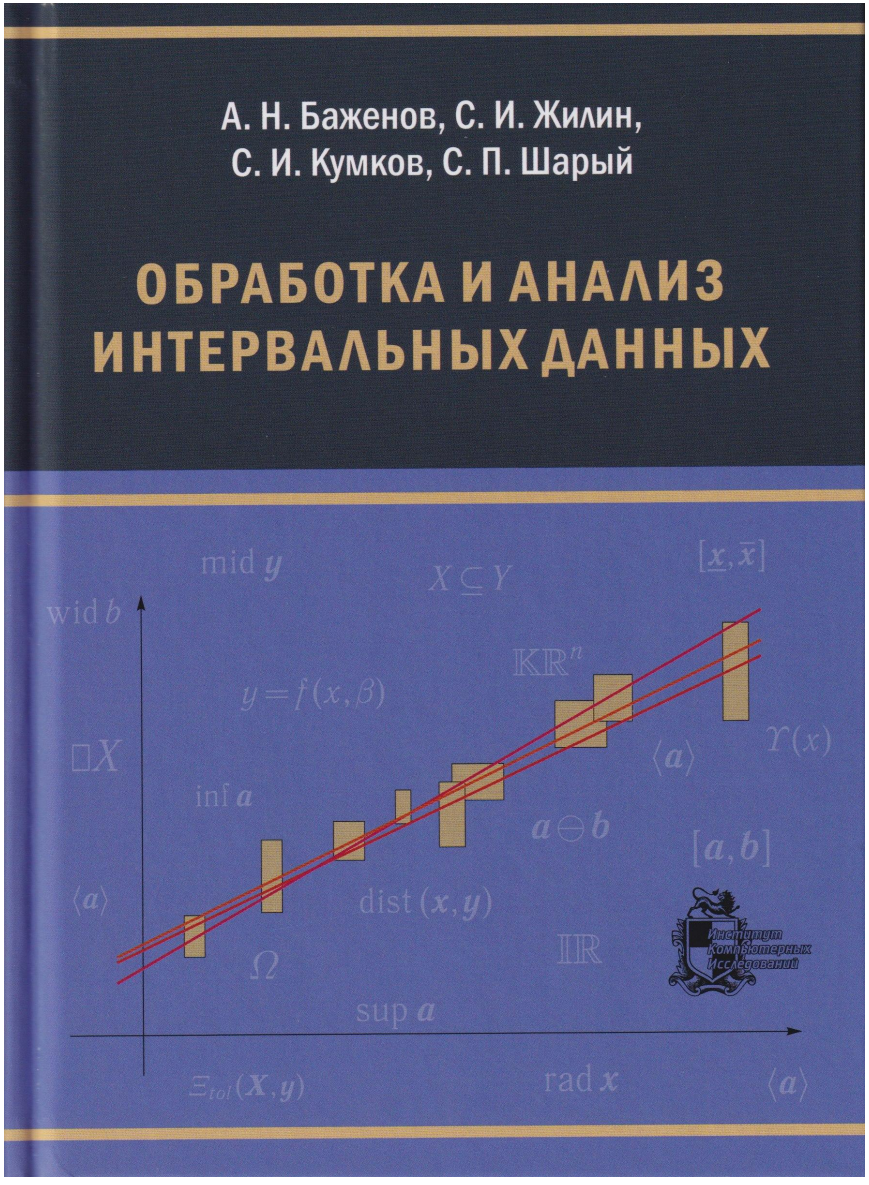
С.И. Спивак, А.П. Вощинин, А.И. Орлов, Н.М. Оскорбин,

С.И. Жилин, С.И. Носков, Б.Т. Поляк, С.И. Кумков, С.П. Шарый, ...

F.S. Schweppe, R.E. Moore, J.P. Norton, M. Milanese,

G. Belforte, L. Pronzato, P. Combettes, E. Walter, L. Jaulin, ...

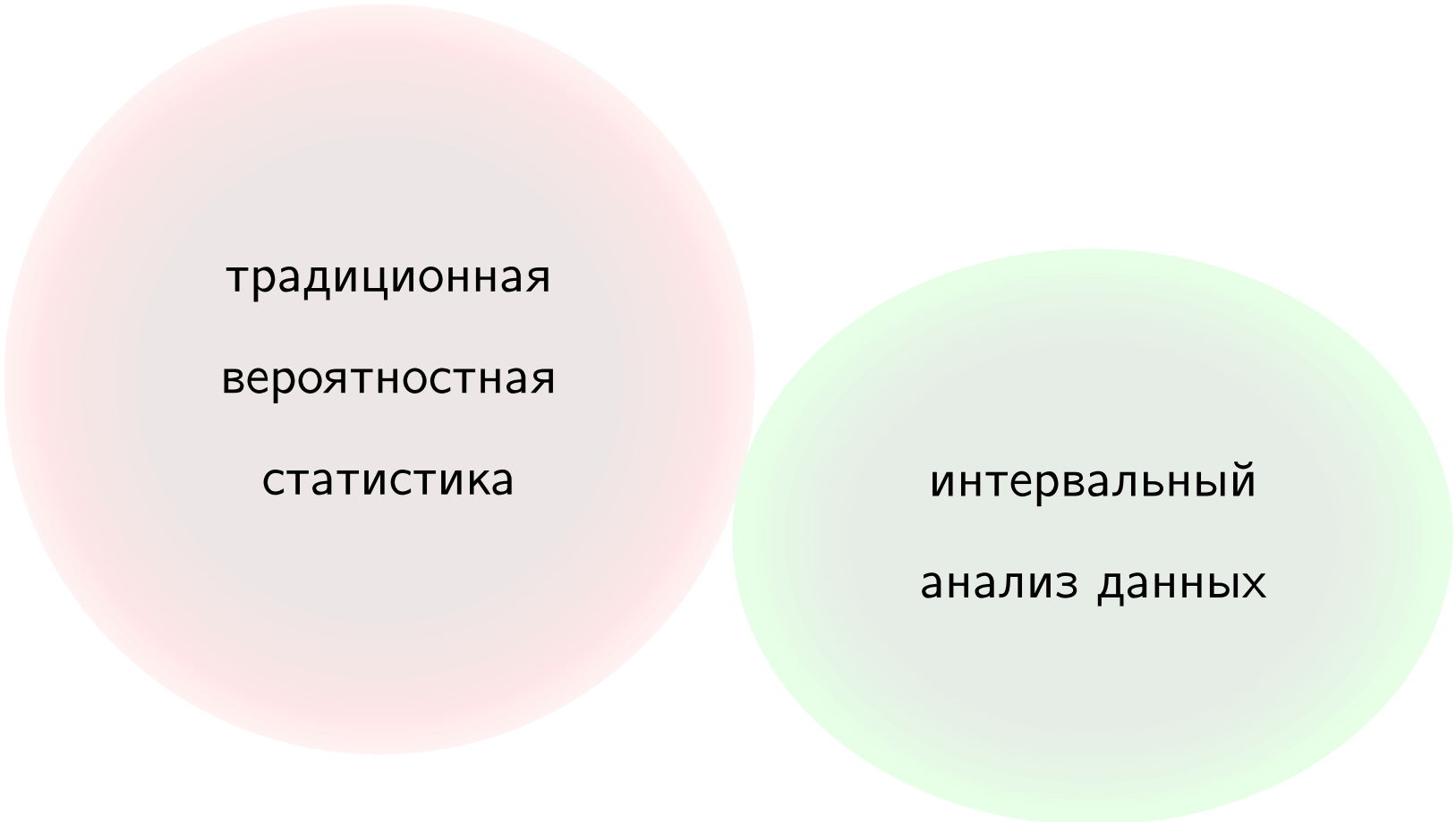
Новая книга по интервальному анализу данных



Издательство ИКИ-РХД,
Ижевск-Москва, 2024

Взаимное отношение

. . . интервального анализа данных и традиционной статистики?



Когда интервальный анализ данных полезен

Интервальные методы обязательно нужно применять, если

- малые выборки,
- нет уверенности в вероятностном характере погрешностей,
- данные имеют интервальный характер,
- . . . .

Каково взаимоотношение интервального анализа данных

и традиционной математической статистики?

«Традиционная математическая статистика»

= теоретико-вероятностная статистика

Каково взаимоотношение интервального анализа данных

и традиционной математической статистики?

«Традиционная математическая статистика»

= теоретико-вероятностная статистика

Методический вопрос: на какой основе сравнивать?

На интервалах данных назначим вероятностные распределения и сравним результаты теоретико-вероятностной статистики и интервального анализа данных.

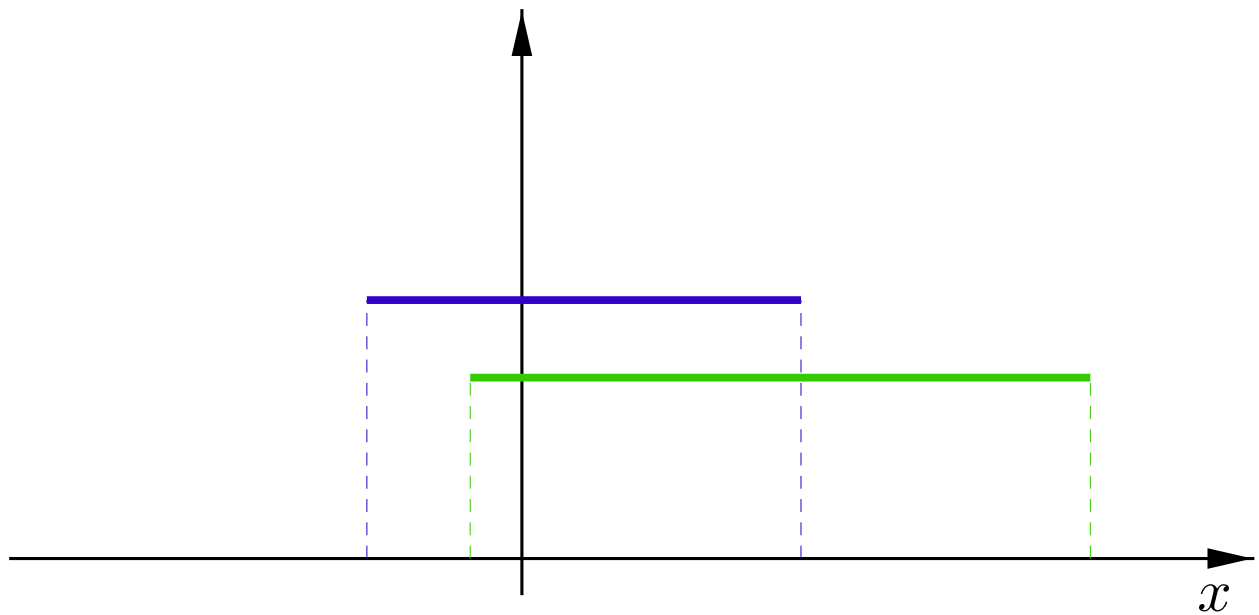
Но какие распределения взять на интервалах?

Равномерные?

На интервалах данных назначим вероятностные распределения и сравним результаты теоретико-вероятностной статистики и интервального анализа данных.

Но какие распределения взять на интервалах?

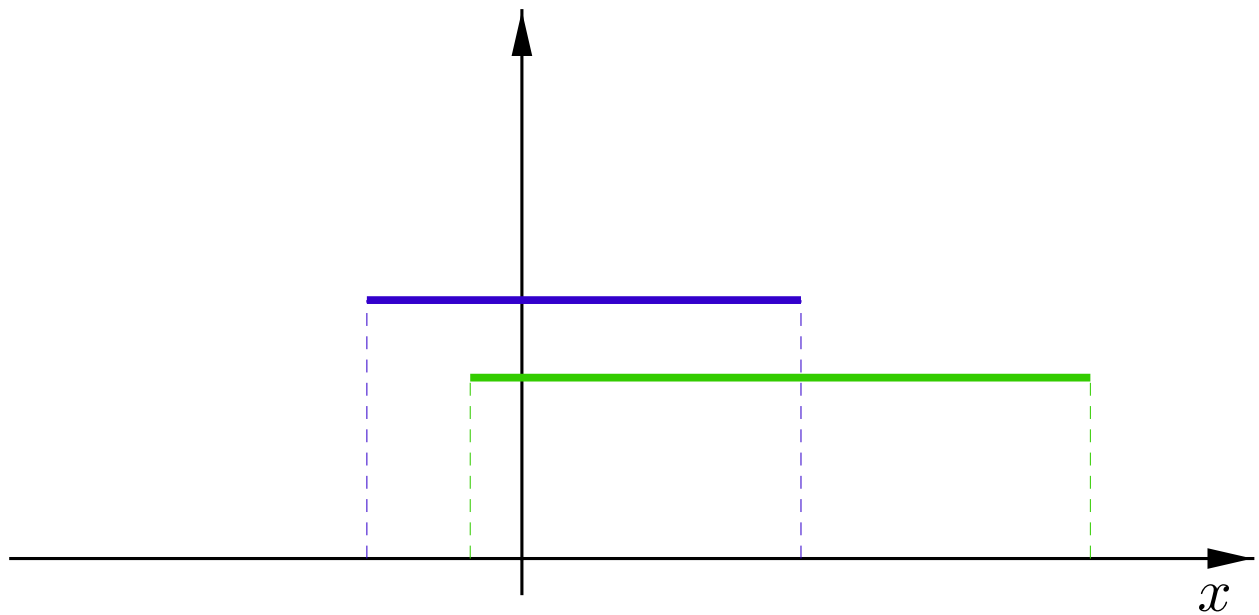
Равномерные?



На интервалах данных назначим вероятностные распределения и сравним результаты теоретико-вероятностной статистики и интервального анализа данных.

Но какие распределения взять на интервалах?

Равномерные наиболее естественны. Но не всегда.



Энтропия вероятностного распределения

Дифференциальная энтропия —

$$H(p) := -\int_X p(x)\ \log\, p(x)\, \mathrm{d}x,$$

где $p(x)$ — плотность распределения,

X — носитель распределения.

— мера неинформативности и неупорядоченности,

которые несёт вероятностное распределение.

Энтропия вероятностного распределения

Дифференциальная энтропия —

$$H(p) := -\int_X p(x)\ \log\, p(x)\, \mathrm{d}x,$$

где $p(x)$ — плотность распределения,

X — носитель распределения.

Факт

Равномерное распределение имеет наибольшую энтропию среди всех распределений на интервале, т. е. даёт наименьшую информацию о том, какие предпочтения в значениях заданы на этом интервале.

Часто говорят: интервал равносилен равномерному распределению . . .

Это в принципе неверно

Вероятностное распределение — нетривиальная конструкция:

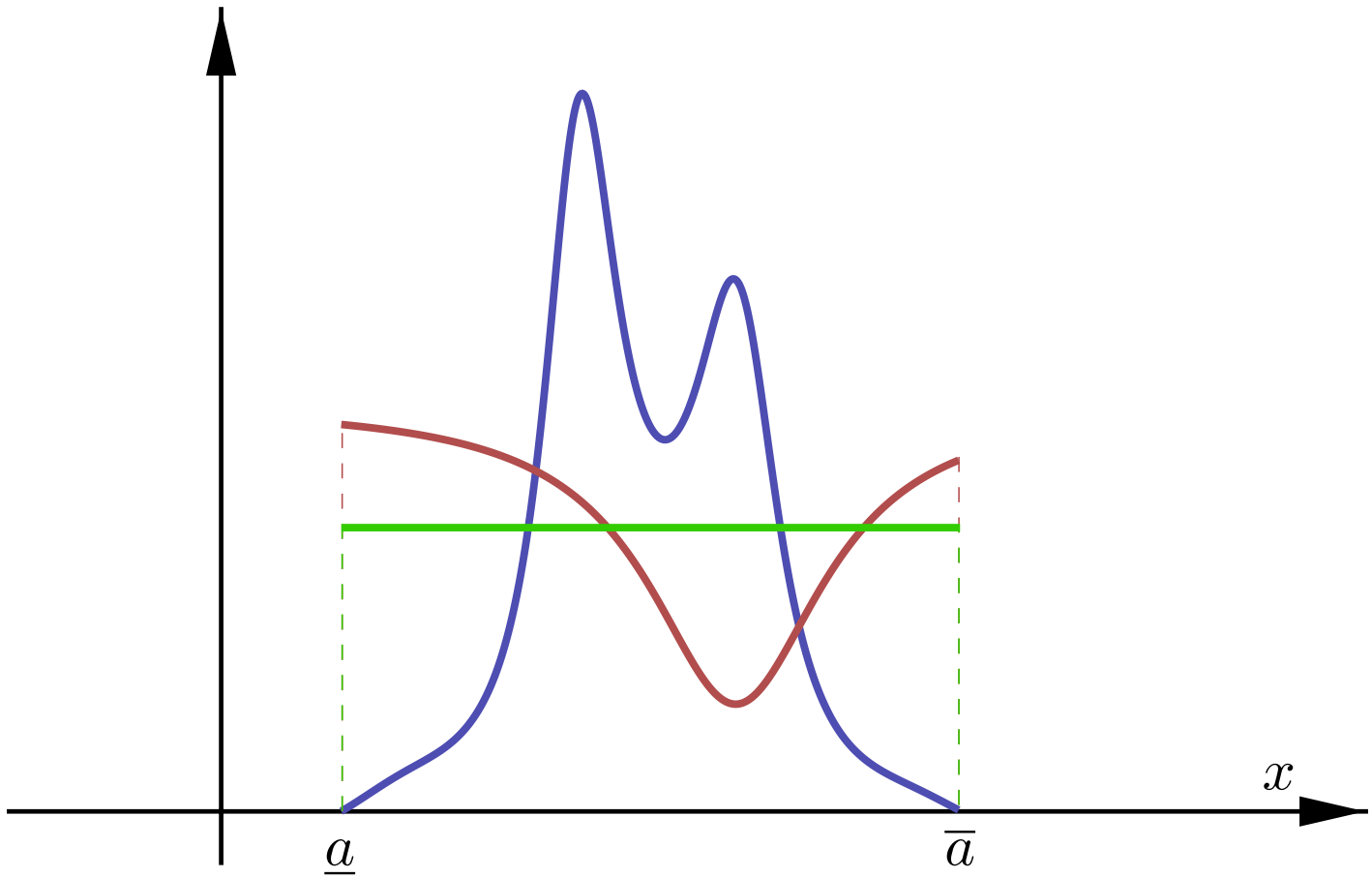
эмпирические условия + математический объект

На интервале может не быть

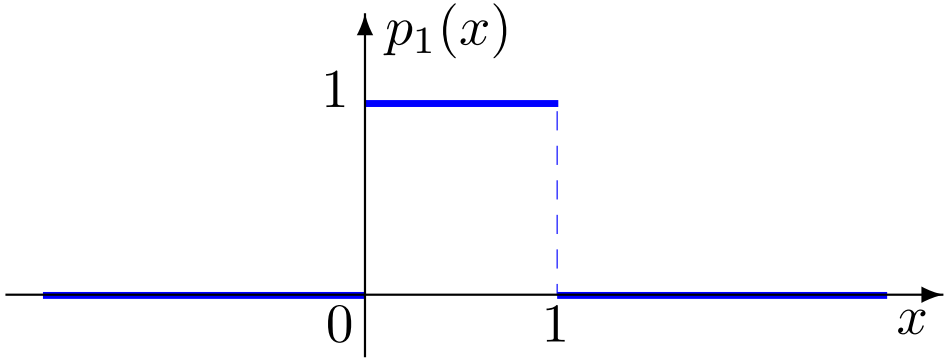
никакого вероятностного распределения

Часто говорят: интервал равносилен равномерному распределению . . .

Это в принципе неверно

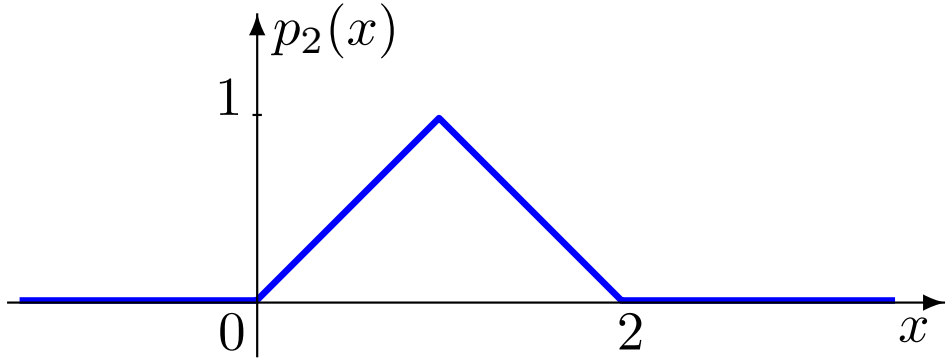


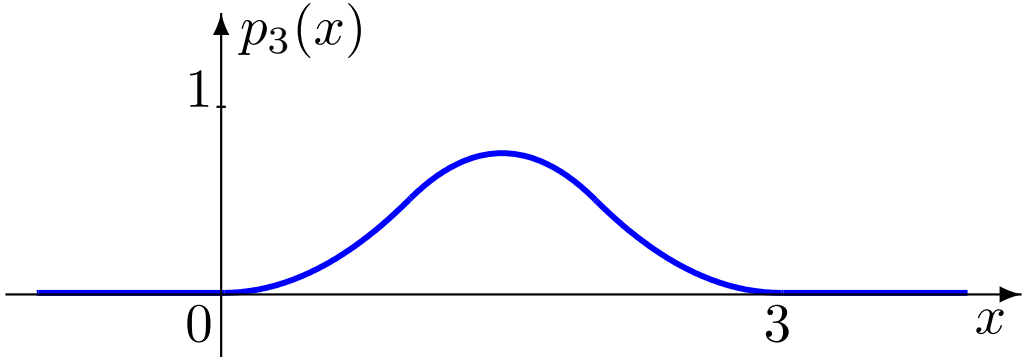
Эволюция плотности вероятности на сумме интервалов



— исходное распределение

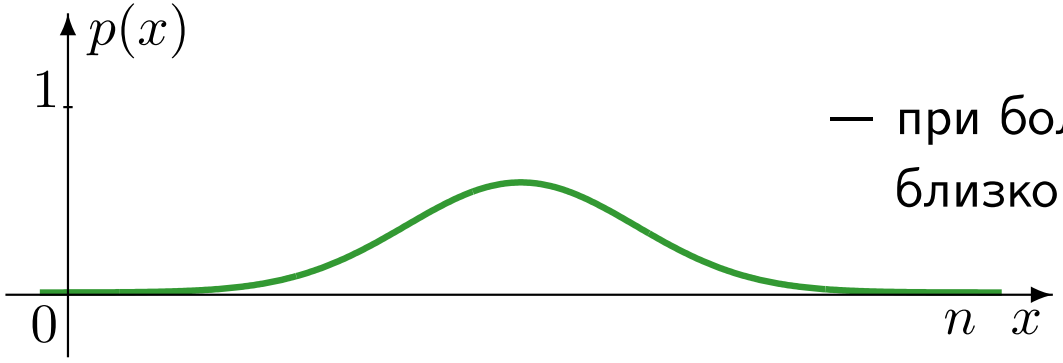
сумма 2-х слагаемых —





— сумма 3-х слагаемых

. . . . . .



— при больших n распределение суммы близко к нормальному гауссовскому

Лемма Е.М. Бронштейна

Е.М. Бронштейн
Об одной возможной вероятностной интерпретации интервальной величины // Вычислительные технологии. – 2014. – Том 19, № 5. – С. 12–14.

Не существует вероятностного распределения двумерной случайной величины (X, Y), определённой на квадрате $[0, 1]$, для которой случайные величины $X + Y$ и XY распределены равномерно.

Лемма Е.М. Бронштейна

Вывод

«. . . естественная теоретико-вероятностная интерпретация интервальных величин несовместима с алгебраическими операциями, т. е. с этой точки зрения интервальный анализ является самостоятельным инструментом описания неопределённостей»

Мы можем лишь назначать вероятностные распределения на исходных интервалах данных — не более.

Исследование С.И. Жилина

Жилин Сергей Иванович

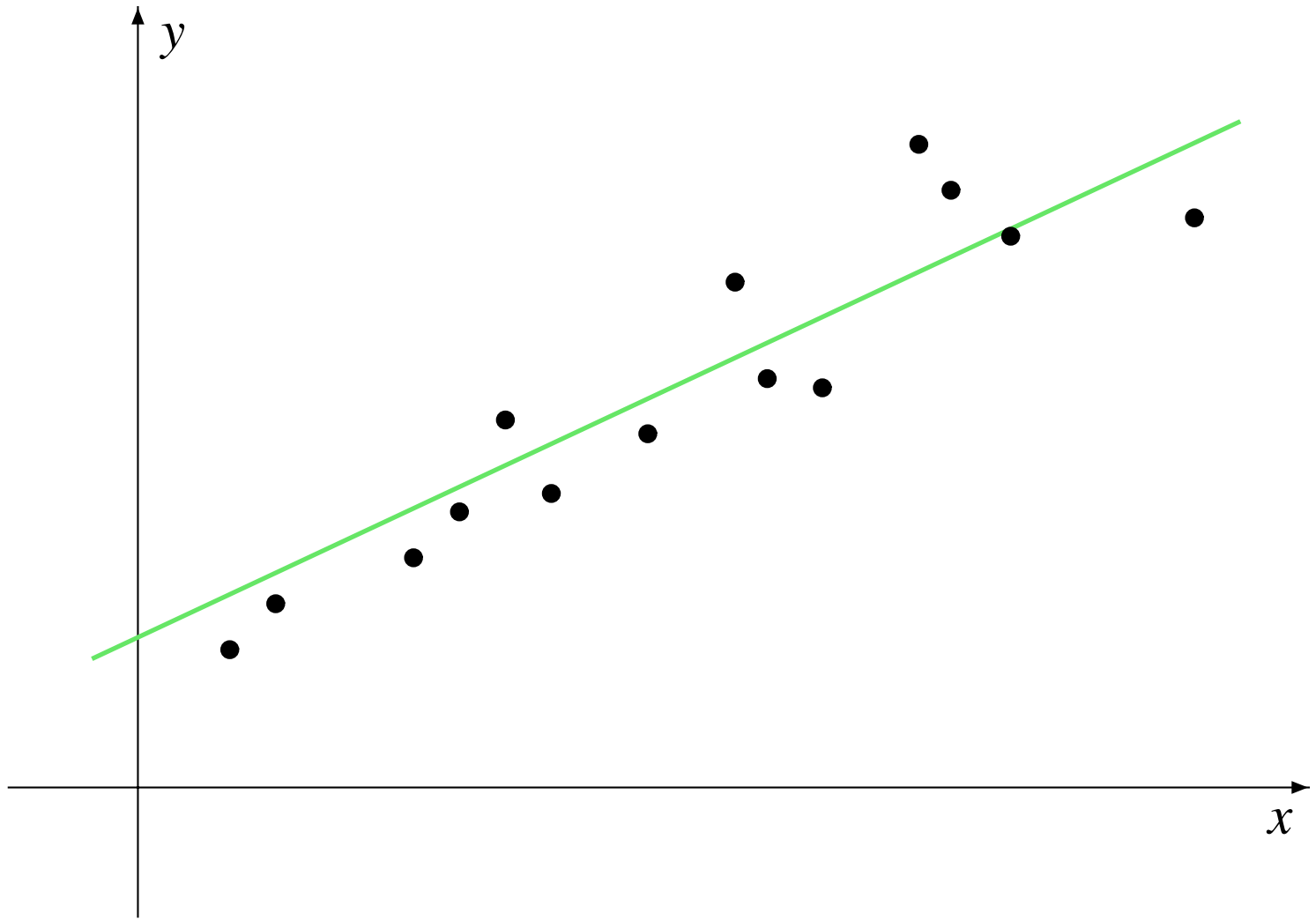
«Нестатистические модели и методы
построения и анализа зависимостей»

Диссертация . . . к. ф.-м. н. по специальности 05.13.01. Барнаул, 2004.

Свободно доступна на веб-сайте
«Интервальный анализ и его приложения» или по прямой ссылке
`http://www.nsc.ru/interval/Library/ApplDiss/Zhilin.pdf`

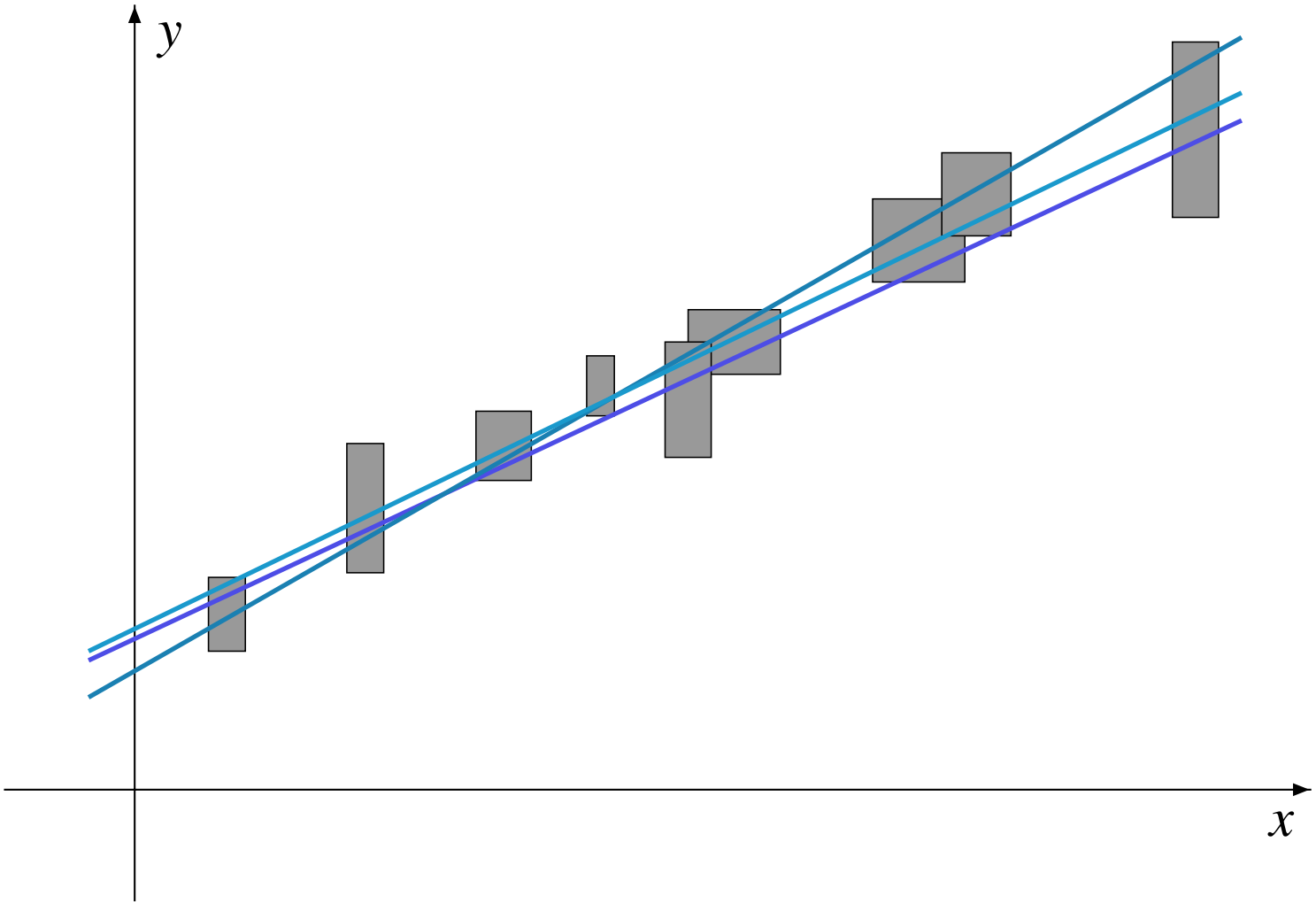
Восстановление зависимостей

Вместо



Восстановление зависимостей по интервальным данным

. . . нужно решить



Исследования С.И. Жилина

На интервалах данных назначались различные распределения и экспериментально сравнивались результаты интервальных методов с традиционными теоретико-вероятностными.

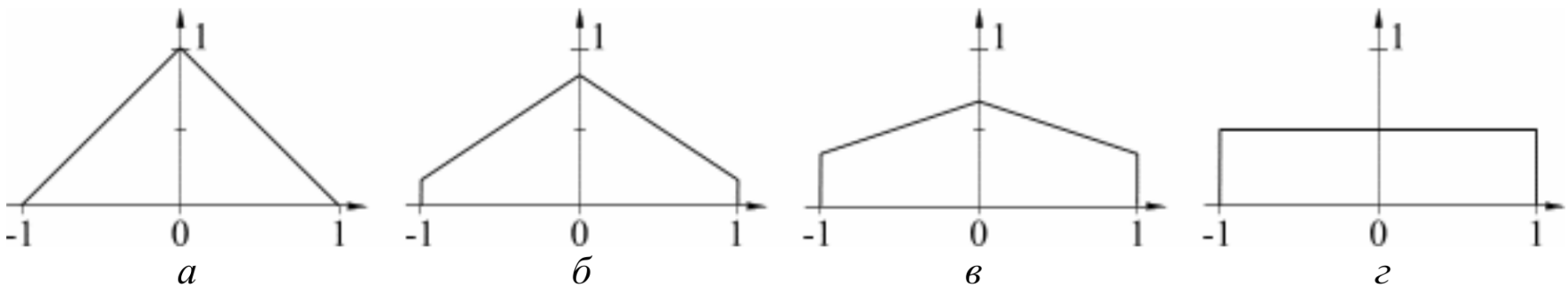


Рис. 2.8. Графики функции плотности p_α^2 при $\varepsilon = 1$ и *а*) $\alpha = 0$,

... ...

Тестировался метод центра неопределённости (МЦН)

Резюме результатов С.И. Жилина

1. Эксперименты свидетельствуют о более высокой эффективности МНК при оценивании параметров, если выполняются стандартные условия нормальности и независимости погрешностей измерений.

2. При нарушении стандартных условий на распределение погрешностей и увеличении кратности наблюдений МЦН не уступает в качестве оценок.

3. При распределении погрешностей, близком к равномерному, более эффективными являются МЦН-оценки.

Исследования В.А. Суханова

Суханов Вячеслав Анатольевич

Исследование эмпирических зависимостей: нестатистический подход

– Барнаул: Издательство Алтайского госуниверситета, 2007, 288 с.

— рассмотрение интервальных методов анализа данных (МЦН) с теоретико-вероятностной точки зрения.

Исследование В.Я. Крейновича и С.П. Шарого

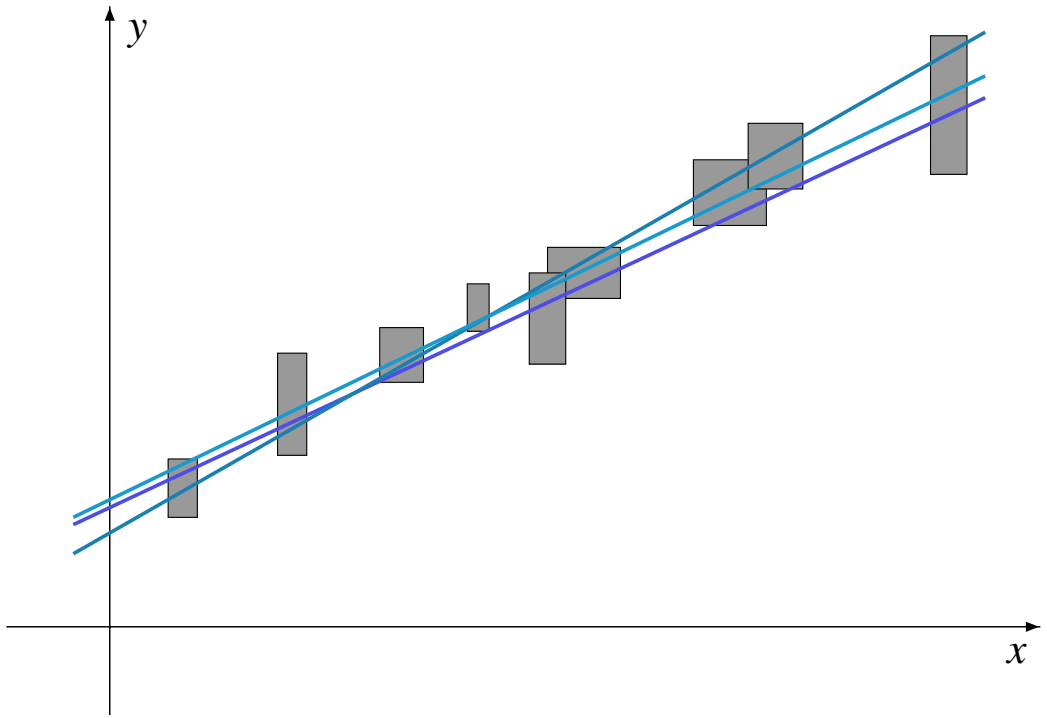
Vladik Kreinovich, Sergey P. Shary

Interval methods for data fitting under uncertainty:
A probabilistic treatment

// Reliable Computing. – 2016. – Vol. 23. – P. 105–140.

Свободно доступна на веб-сайте журнала или по прямой 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.

Вероятностные гарантии агрегирования интервала

В интервальном анализе данных популярный способ получения интервалов измеряемых величин — их *агрегирование* или *группировка*.

Математически, если результаты повторных измерений интересующей нас величины равны $x_1, x_2, \ldots, x_n$, то интервальным результатом серии следует взять

$$\boldsymbol{x} = \left[\min_{1 \le i \le n} x_i, \ \max_{1 \le i \le n} x_i \right].$$

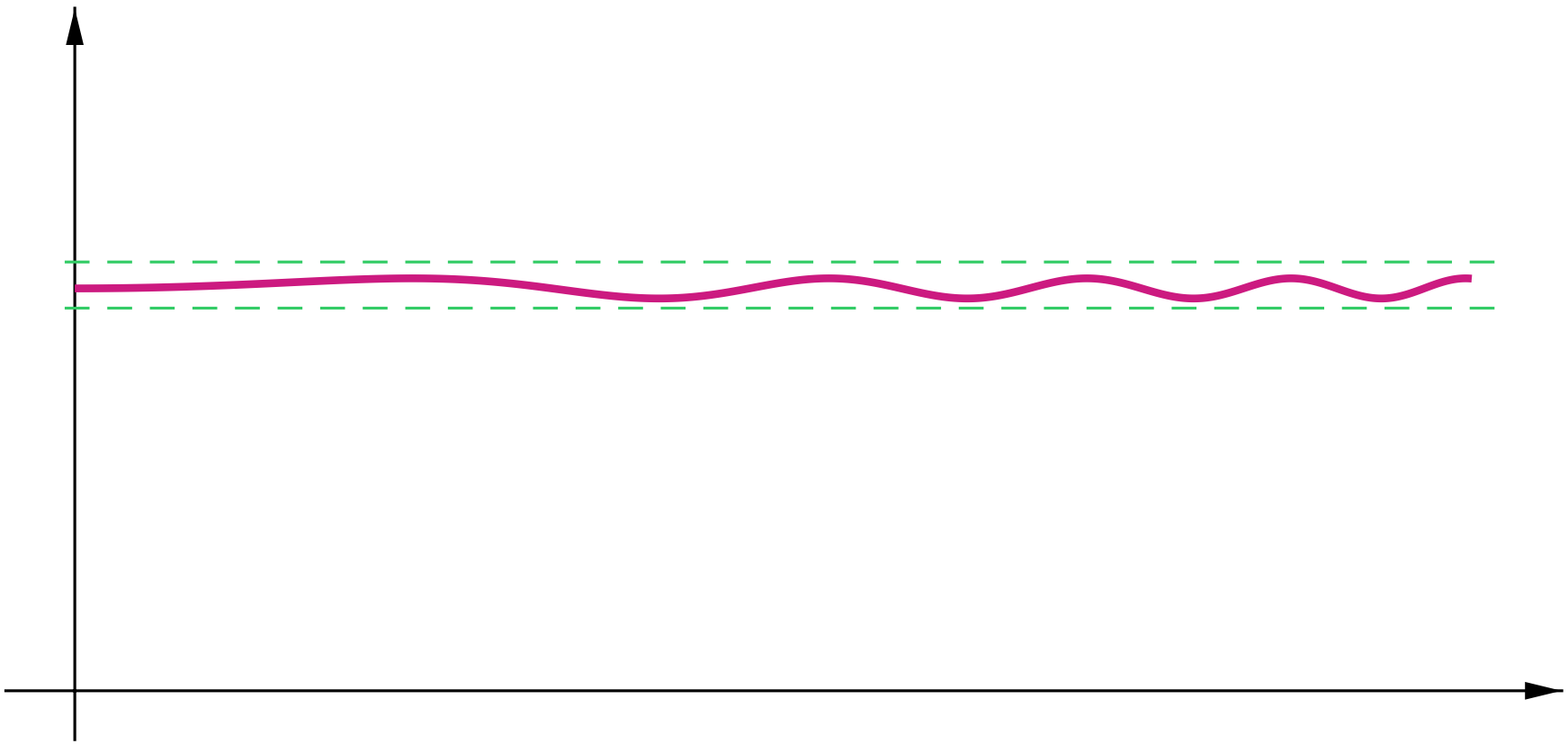
В практической метрологии результат серии повторяющихся измерений часто представляют одним числом, например, средним отдельных результатов.

Но представление результата серии одним числом
теряет информацию о разбросе данных.

Если хотим указать её в результате измерения, то должны вводить второе число, которое может быть стандартным отклонением и т. п.

Хорошая альтернатива — интервальное агрегирование результатов.

Непрерывный аналог агрегирования



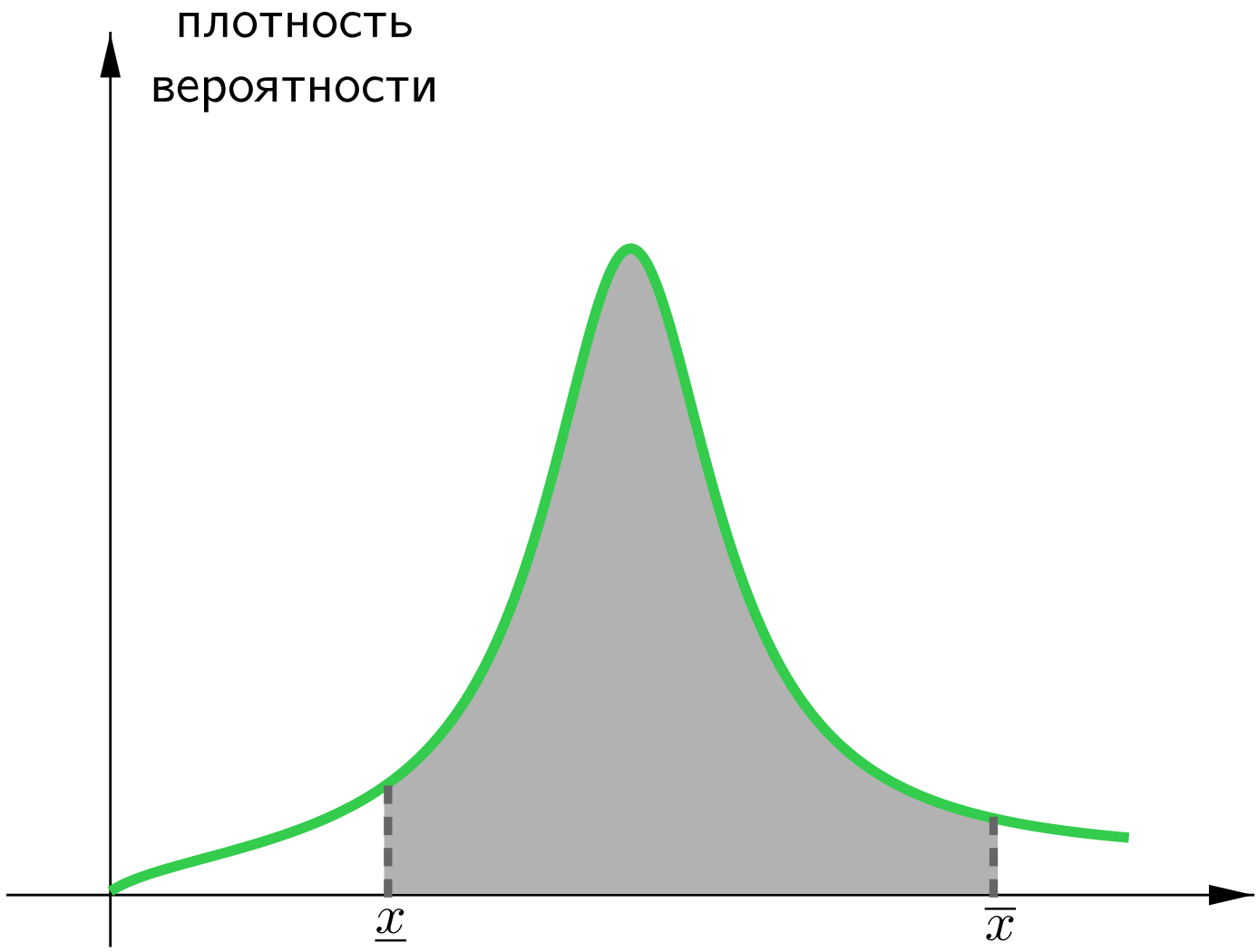
Вероятностные гарантии агрегирования интервала

Пусть измеряемая величина подчинена законам теории вероятностей, т. е. является вероятностно-случайной, а измерения

$$x_1,\ x_2,\ \ldots\ , x_n$$

— это её отдельные реализации.

Если вероятностное распределение — «вероятностная масса» 1, рассредоточенная на $\mathbb{R}$, то какую её часть покроем агрегированием?

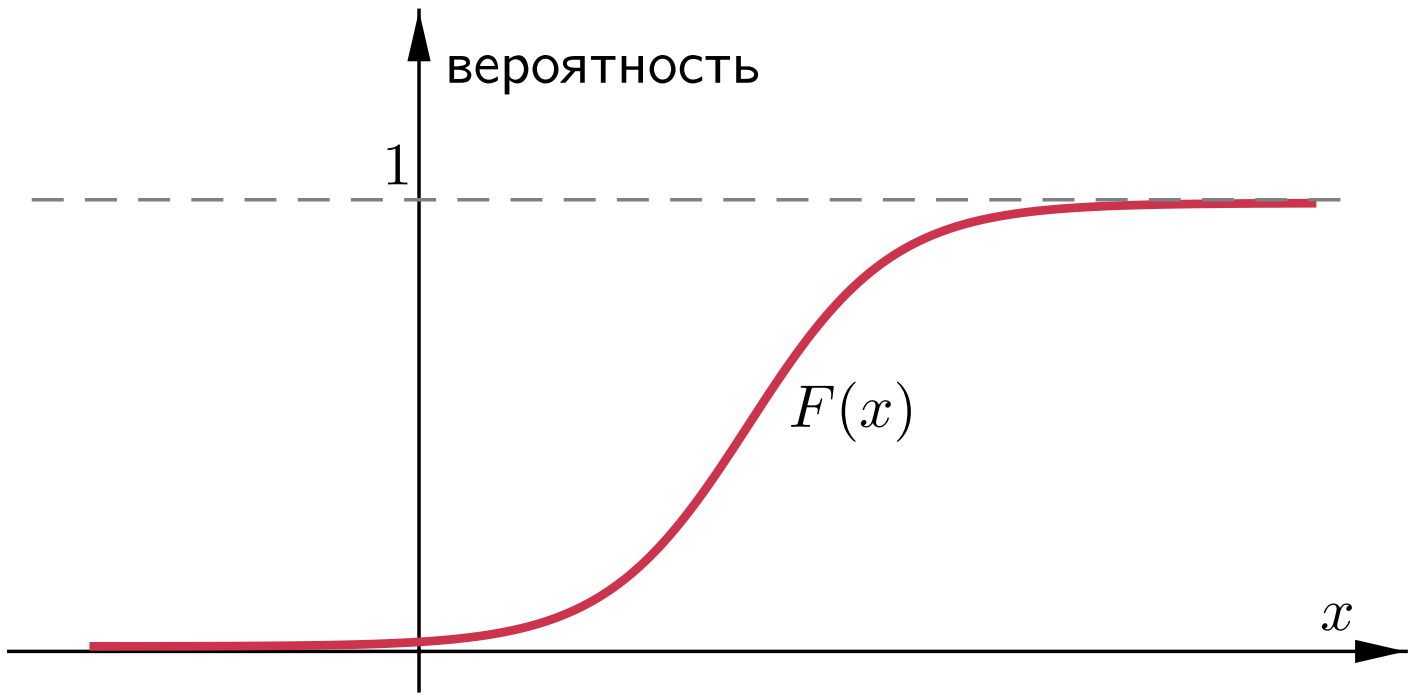

плотность
вероятности
$\underline{x}$
$\overline{x}$

Функция распределения вероятностей

Если ξ — случайная величина, то

$$F(x) = \mathrm{P}(\xi \leq x)$$

— её функция распределения.



Вероятностные гарантии агрегирования интервала

Если $F(x)$ — функция распределения вероятностей измеряемой величины и обозначаем

$$\underline{x} := \min_{1 \le i \le n} x_i \qquad \text{и} \qquad \overline{x} := \max_{1 \le i \le n} x_i,$$

то ответ на вопрос требует нахождения приращения

$$F(\overline{x}) - F(\underline{x}).$$

С другой стороны, эту величину можно трактовать как вероятность того, что следующая реализация x_{n+1} нашей случайной величины, попадёт в интервал от $\underline{x}$ до $\overline{x}$ уже измеренных значений.

Постановка задачи

Предположим, что случайно берутся

вещественные числа — $x_1, x_2, \ldots, x_n$.

Фиксируем границы их изменения —

нижнюю $\underline{x} := \min\{x_1, x_2, \ldots, x_n\}$

и верхнюю $\overline{x} := \max\{x_1, x_2, \ldots, x_n\}$.

Какова вероятность того, что случайно взятое,

по тому же распределению, число x_{n+1} попадёт в интервал $[\underline{x}, \overline{x}]$?

Взятые в порядке возрастания элементы выборки значений случайной величины называются, как известно, *порядковыми статистиками*.

Им посвящена обширная литература
и большое количество результатов теории вероятностей.


ван дер Варден Б.Л. Математическая статистика. – Москва: Издательство иностранной литературы, 1960.


Галамбош Я. Асимптотическая теория экстремальных порядковых статистик. – Москва: Наука, 1984.


Дэйвид Г. Порядковые статистики. – Москва: Наука, 1979.

Уточнение постановки

Задана вещественная функция распределения вероятностей $\Phi(x)$.

Пусть задаваемое ею распределение абсолютно непрерывно, т. е. существует функция $p(x) \geq 0$ — плотность распределения, такая что

$$\Phi(x) = \int_{-\infty}^{x} p(x)\,\mathrm{d}x.$$

Тогда

$$\Phi'(x) = p(x)$$

для любой точки непрерывности $p(x)$.

Функцией распределения максимума случайных величин с одной общей функцией распределения $\Phi(x)$ является $(\Phi(x))^n$.

Плотность вероятности максимума

$$n\,(\Phi(x))^{n-1}p(x).$$

Функцией распределения минимума случайных величин с общей функцией распределения $\Phi(x)$ является $1-(1-\Phi(x))^n$.

Плотность вероятности минимума

$$n\,(1-\Phi(x))^{n-1}p(x).$$

Если на $\mathbb{R}$ случайно, по распределению с этой плотностью, берётся число, то вероятность его попадания в интервал $[x, x + \mathrm{d}x]$ равна

$$n \left(\Phi(x)\right)^{n-1} p(x)\,\mathrm{d}x$$

для максимума

и

$$n \left(1 - \Phi(x)\right)^{n-1} p(x)\,\mathrm{d}x$$

для минимума.

Вероятность того, что взятое затем случайно число x_{n+1} будет бóльшим взятого x, есть

$$\big(1-\Phi(x)\big)\,n\,\big(\Phi(x)\big)^{n-1}p(x)\,\mathrm{d}x.$$

Полная вероятность сложного события

$$x_{n+1} > \max\{\,x_1, x_2, \ldots, x_n\}$$

получается суммированием этих значений по всей $\mathbb{R}$:

$$n\int_{-\infty}^{+\infty}\big(1-\Phi(x)\big)\big(\Phi(x)\big)^{n-1}p(x)\,\mathrm{d}x.$$

Имеем

$$n\int_{-\infty}^{+\infty}\big(1-\Phi(x)\big)\big(\Phi(x)\big)^{n-1}p(x)\,\mathrm{d}x \qquad (\spadesuit)$$

$$= n\int_{-\infty}^{+\infty}\big(\Phi(x)\big)^{n-1}p(x)\,\mathrm{d}x - n\int_{-\infty}^{+\infty}\big(\Phi(x)\big)^{n}p(x)\,\mathrm{d}x$$

$$= n\int_{-\infty}^{+\infty}\big(\Phi(x)\big)^{n-1}\,\mathrm{d}\Phi(x) - n\int_{-\infty}^{+\infty}\big(\Phi(x)\big)^{n}\,\mathrm{d}\Phi(x)$$

$$= n\,\frac{\big(\Phi(x)\big)^{n}}{n}\,\Bigg|_{-\infty}^{+\infty} - n\,\frac{\big(\Phi(x)\big)^{n+1}}{n+1}\,\Bigg|_{-\infty}^{+\infty}$$

$$= \big(1^n-0^n\big) - \frac{n}{n+1}\big(1^{n+1}-0^{n+1}\big) = \frac{1}{n+1}.$$

Для нахождения вероятности события

$$x_{n+1} < \min\{x_1, x_2, \ldots, x_n\}$$

заметим, что это эквивалентно

$$-x_{n+1} > -\max\{-x_1, -x_2, \ldots, -x_n\}.$$

Искомая вероятность тоже равна $1/(n+1)$.

Для нахождения вероятности события

$$x_{n+1} < \min\{\, x_1, x_2, \ldots, x_n\}$$

заметим, что это эквивалентно

$$-x_{n+1} > -\max\{\, -x_1, -x_2, \ldots, -x_n\}.$$

Искомая вероятность тоже равна $1/(n+1)$.

Непопадание x_{n+1} в $[\underline{x}, \overline{x}]$ есть сумма несовместимых событий

$$x_{n+1} < \min\{\, x_1, x_2, \ldots, x_n\} \qquad \text{и} \qquad x_{n+1} > \max\{\, x_1, x_2, \ldots, x_n\}.$$

Поэтому их вероятности складываются

$$\frac{1}{n+1} + \frac{1}{n+1} = \frac{2}{n+1}.$$

Вероятность дополнительного события,

т. е. принадлежности интервалу $[\underline{x}, \overline{x}]$, равна

$$\mathsf{P}\left(x_{n+1} \in \left[\min_{1 \le i \le n} x_i, \max_{1 \le i \le n} x_i \right] \right) = 1 - \frac{2}{n+1}$$

Результат не зависит от вида распределения.

Пример

Один из популярных стандартных уровней гарантии — 95%.

Чтобы получить его, необходимо столько измерений n, что

$$1 - \frac{2}{n+1} \geq 0.95,$$

т. е.

$$\frac{2}{n+1} \leq 0.05 \qquad \Longrightarrow \qquad n+1 \geq 40.$$

Итого, нужно $n \geq 39$ измерений.

Другой вывод результата

Н.И. Чернова заметила, что наша оценка может быть выведена просто.

В наборе независимых случайных величин x_1, x_2, ..., x_n, x_{n+1}, имеющих общее абсолютно непрерывное распределение, вероятность того, что x_{n+1} будет наибольшей, равна всегда $1/(n+1)$.

Другой вывод результата

Н.И. Чернова заметила, что наша оценка может быть выведена просто.

В наборе независимых случайных величин x_1, x_2, ..., x_n, x_{n+1}, имеющих общее абсолютно непрерывное распределение, вероятность того, что x_{n+1} будет наибольшей, равна всегда $1/(n+1)$.

Это следует из равновозможности всех вариантов взаимных расположений чисел x_1, x_2, ..., x_n, x_{n+1} на $\mathbb{R}$, откуда вытекает равновероятность того, что x_{n+1} — самое большое, второе и т. д.

Поэтому искомая вероятность $= \dfrac{1}{n+1}$.

Абсолютная непрерывность распределения нужна для того, чтобы вариант совпадения значений x_i имел нулевую вероятность.

Обобщения

Предположим, что среди n измерений x_1, x_2, ..., x_n присутствует одно «измерение-примесь» с распределением вероятностей $\Psi(x)$, которое отличается от $\Phi(x)$ — распределения других измерений.

Какова теперь вероятность попадания очередного измерения x_{n+1} с распределением $\Phi(x)$ в интервал от минимума до максимума,

$$\boldsymbol{x} = \left[\min_{1 \le i \le n} x_i, \max_{1 \le i \le n} x_i \right],$$

уже полученных значений x_1, x_2, ..., x_n измеряемой величины?

Предполагаем, что Ψ также абсолютно непрерывно,
и $q(x)$ — его плотность.

Общий случай одного «измерения-примеси»

Теперь функция распределения максимума всех n измерений равна $(\Phi(x))^{n-1}\Psi(x)$, а её плотность —

$$(n-1)\big(\Phi(x)\big)^{n-2}\Psi(x)\,p(x) + \big(\Phi(x)\big)^{n-1}\,q(x).$$

Вероятность попадания числа, случайно взятого по этому распределению, в $[x, x+\mathrm{d}x]$

$$\Big((n-1)\big(\Phi(x)\big)^{n-2}\Psi(x)\,p(x) + \big(\Phi(x)\big)^{n-1}\,q(x)\Big)\,\mathrm{d}x.$$

Вероятность выполнения неравенства $x_{n+1} > x$

$$\big(1-\Phi(x)\big)\Big((n-1)\big(\Phi(x)\big)^{n-2}\Psi(x)\,p(x) + \big(\Phi(x)\big)^{n-1}\,q(x)\Big)\,\mathrm{d}x,$$

Для определения вероятности выполнения $x_{n+1} > \max\{x_1, \ldots, x_n\}$ необходимо взять интеграл по всем $x \in \mathbb{R}$:

$$\int_{-\infty}^{+\infty} \big(1 - \Phi(x)\big)\Big((n-1)\big(\Phi(x)\big)^{n-2}\Psi(x)\,p(x) + \big(\Phi(x)\big)^{n-1}\,q(x)\Big)\,\mathrm{d}x$$

$$= (n-1)\int_{-\infty}^{+\infty} \big(1 - \Phi(x)\big)\big(\Phi(x)\big)^{n-2}\Psi(x)\,p(x)\,\mathrm{d}x$$

$$+ \int_{-\infty}^{+\infty} \big(1 - \Phi(x)\big)\big(\Phi(x)\big)^{n-1}\,q(x)\,\mathrm{d}x.$$

Первый интеграл очень похож на (♠), который уже брали ранее.

Для определения вероятности выполнения $x_{n+1} > \max\{x_1, \ldots, x_n\}$ необходимо взять интеграл по всем $x \in \mathbb{R}$:

$$\int_{-\infty}^{+\infty} \big(1 - \Phi(x)\big)\Big((n-1)\big(\Phi(x)\big)^{n-2}\Psi(x)\,p(x) + \big(\Phi(x)\big)^{n-1}\,q(x)\Big)\,\mathrm{d}x$$

$$= (n-1)\int_{-\infty}^{+\infty} \big(1 - \Phi(x)\big)\big(\Phi(x)\big)^{n-2}\Psi(x)\,p(x)\,\mathrm{d}x$$

$$+ \int_{-\infty}^{+\infty} \big(1 - \Phi(x)\big)\big(\Phi(x)\big)^{n-1}\,q(x)\,\mathrm{d}x.$$

Первый интеграл очень похож на (♠), который уже брали ранее.

И так далее . . .

Получаем

$$1 - \frac{3}{n} \leq$$

$$\mathsf{P}\left(x_{n+1} \in \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right] \right) \qquad (*)$$

$$\leq 1.$$

Получаем

$$1 - \frac{3}{n} \leq \mathsf{P}\left(x_{n+1} \in \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right] \right) \leq 1. \qquad (*)$$

Это в ≈ 1.5 раза хуже оригинальной оценки без «примеси» . . .

Дальнейшие исследования

Можно указать ещё несколько способов оценивания

асимптотики при $n \to \infty$ влияние измерения-примеси.

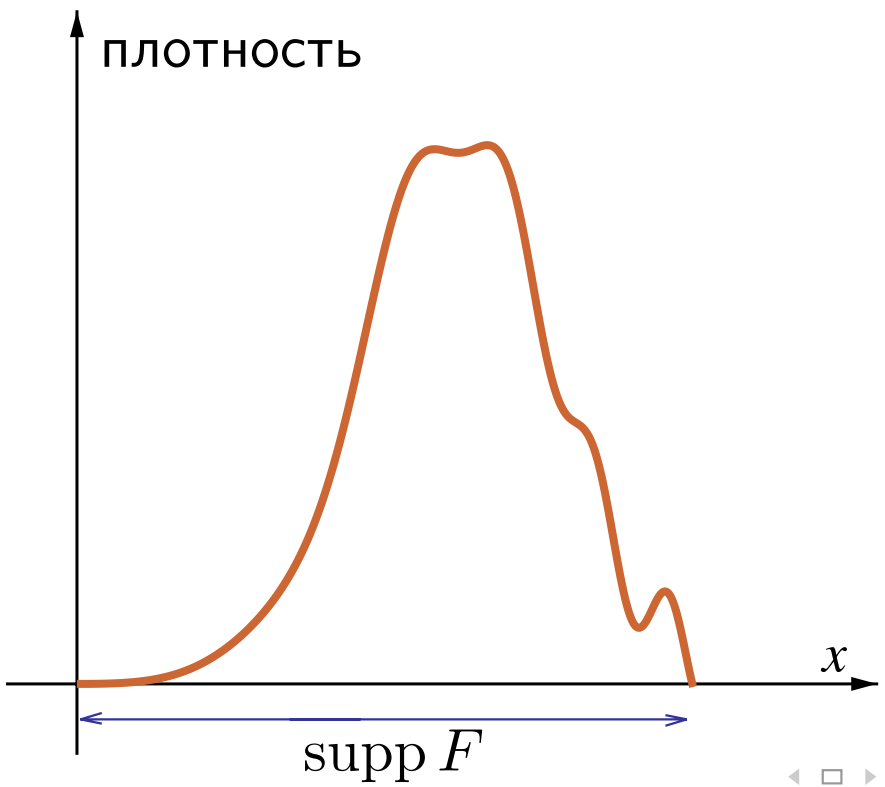
В них на распределения Φ и Ψ накладываются дополнительные условия, но взамен получаем более быстрое стремление оценки к 1.

Носитель распределения

Носитель $\operatorname{supp} F$ вероятностного распределения F — наименьшее замкнутое множество, вероятность принадлежать которому $= 1$.

Для абсолютно непрерывного распределения F

$$\operatorname{supp} F = \text{ замыкание множества } \left\{ x \in \mathbb{R} \mid \text{ плотность } > 0 \right\}.$$



Случай ограниченного носителя «примеси»

Пусть распределение $\Psi(x)$ имеет ограниченный носитель,
т. е. $q(x) \neq 0$ лишь на каком-то интервале $[\underline{\omega}, \overline{\omega}] \subset \mathbb{R}$.

Потребуем также, чтобы носитель распределения Ψ содержался в топологической внутренности носителя распределения Φ, т. е.

$$\operatorname{supp} \Psi \subseteq \operatorname{int} \operatorname{supp} \Phi$$

Из того, что $\Phi(x)$ — неубывающая функция, вытекает

$$0 \leq M := \max_{x \in [\underline{\omega}, \overline{\omega}]} \Phi(x) < 1.$$

Тогда получаем ту же самую асимптотику

стремления достоверности интервального агрегирования к 1:

$$1 - \frac{2}{n} - \frac{M^{n-2}}{2} \leq$$

$$\mathsf{P}\left(x_{n+1} \in \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right] \right)$$

$$\leq 1,$$

Но несколько сильнее оценки ($*$).

Случай близкого распределения «примеси»

Рассмотрим ситуацию, когда распределения $\Phi(x)$ и $\Psi(x)$ «близки» друг к другу, т. е. отличаются «не слишком сильно».

Эту близость можно понимать в разных смыслах, и ниже исследуем случай, в котором отношение плотностей вероятностей ограниченно.

Случай близкого распределения «примеси»

Потребуем, чтобы для плотностей распределений $\Phi(x)$ и $\Psi(x)$,
т. е. для функций $p(x)$ и $q(x)$, было выполнено условие:

для некоторой положительной константы K справедливо

$$0 \leq \frac{q(x)}{p(x)} \leq K \qquad (\blacklozenge)$$

при всех x, в которых $p(x) \neq 0$.

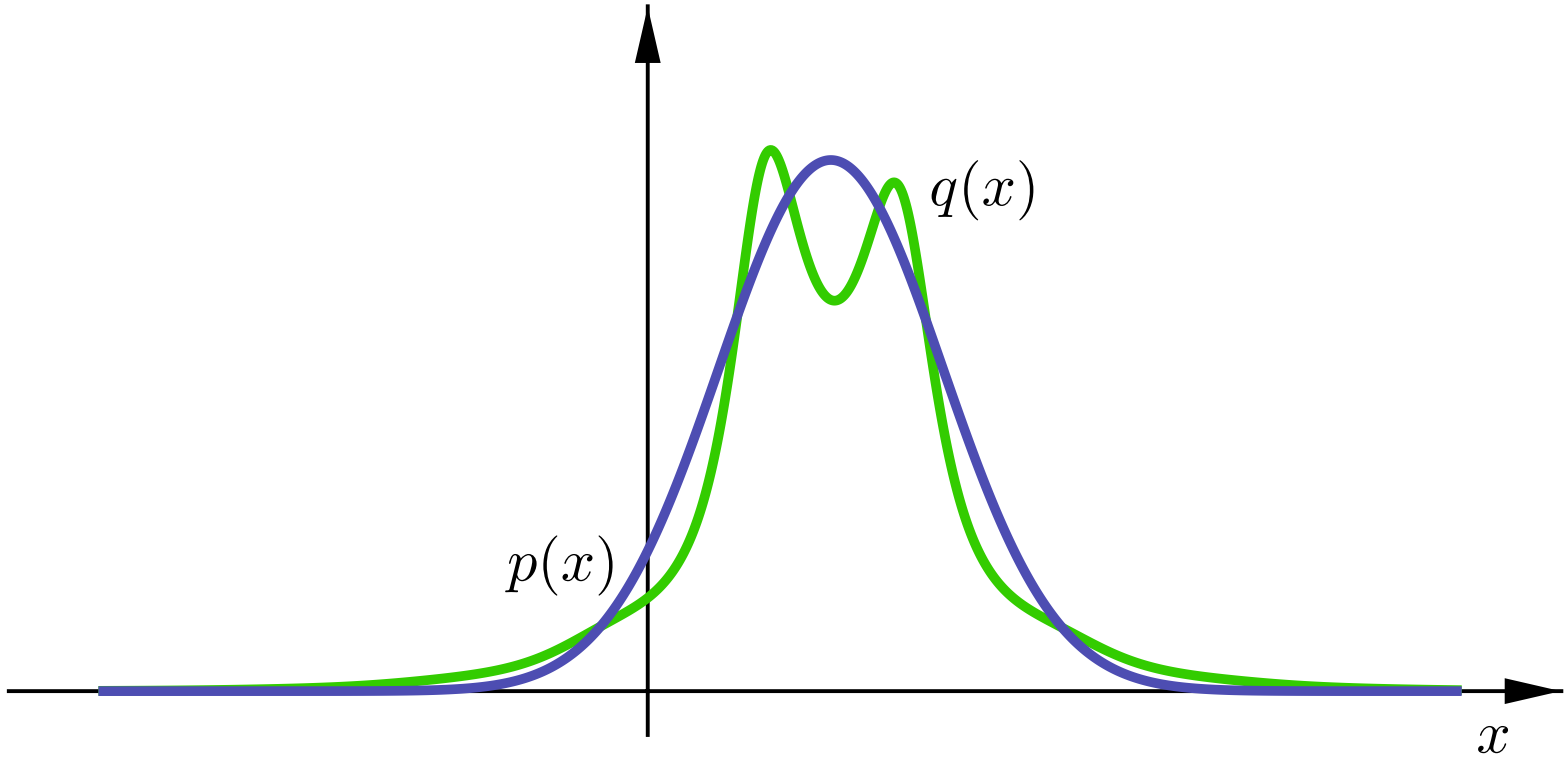


Иллюстрация плотностей вероятностных распределений,

удовлетворяющих условию (♦).

В целом, в условиях (♦) получаем

$$1 - \frac{2}{n} - \frac{2K}{n(n+1)} \leq$$

$$\mathsf{P}\left(x_{n+1} \in \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right] \right)$$

$$\leq 1.$$

Это снова несколько лучше оценки (*).

Условие (♦) близости плотностей заведомо выполнено, например, если выполнено требования предыдущего случая: $\Psi(x)$ имеет ограниченный носитель, и он содержится во внутренности носителя $\Phi(x)$,

$$\operatorname{supp}\Psi \subseteq \operatorname{int}\operatorname{supp}\Phi.$$

Для традиционных вероятностных распределений, имеющих неограниченный носитель, условие (♦) часто не выполняется, хотя нарушается при $|x| \to \infty$, т. е. за границами физического смысла.

На практике можно удовлетворить условию (♦), выполнив стандартное усечение распределения $\Psi(x)$ к какому-нибудь разумному интервалу с «размазыванием» вероятностной массы отсекаемых хвостов.

Пример 1

Рассмотрим два нормальных распределения с одинаковыми средними, которые можно считать нулевыми, и разными дисперсиями σ_1 и σ_2.

Их плотности —

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma_1}\exp\Big(-\frac{x^2}{2\sigma_1^2}\Big), \qquad q(x) = \frac{1}{\sqrt{2\pi}\,\sigma_2}\exp\Big(-\frac{x^2}{2\sigma_2^2}\Big),$$

и отношение

$$\frac{q(x)}{p(x)} \;=\; \frac{\sigma_1}{\sigma_2}\exp\left(\frac{x^2}{2}\,\frac{\sigma_2^2-\sigma_1^2}{(\sigma_1\sigma_2)^2}\right).$$

Показатель в экспоненте отрицателен (неположителен) при $\sigma_1 \geq \sigma_2$, и при таких условиях выписанное выражение ограниченно.

Иначе, если $\sigma_1 < \sigma_2$, экспонента

$$\exp\left(\frac{x^2}{2}\,\frac{\sigma_2^2-\sigma_1^2}{(\sigma_1\sigma_2)^2}\right)$$

неограниченно растёт с ростом $|x|$.

Условие близости плотностей (♦) выполнено, если дисперсия «измерения-примеси» не больше дисперсии основной серии.

Например, это может быть измерение интересующей нас величины при более низкой температуре, когда тепловые флуктуации и шумы уменьшаются.

Рассмотрим два нормальных распределения

с одинаковыми дисперсиями σ, но разными средними μ_1 и μ_2.

Их плотности —

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\Big(-\frac{(x-\mu_1)^2}{2\sigma^2}\Big), \qquad q(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\Big(-\frac{(x-\mu_2)^2}{2\sigma^2}\Big),$$

и отношение

$$\frac{q(x)}{p(x)} \;=\; \exp\left(\frac{2x(\mu_2-\mu_1)+(\mu_1^2-\mu_2^2)}{2\sigma^2}\right).$$

Оно может неограниченно расти при $|x| \to \infty$

при подходящем соотношении μ_1 и μ_2.

Как следствие, условие (♦) не выполнено на всей вещественной оси.

Случай нескольких «измерений-примесей»

Наши оценки обобщаются на случай, когда в выборке присутствует несколько «измерений-примесей» с распределением $\Psi(x)$, которое отличается от распределения $\Phi(x)$ основной массы измерений.

Пусть среди n измерений x_1, x_2, ..., x_n имеются

$(n-k)$ измерений с распределением $\Phi(x)$ и

k «измерений-примесей» с распределением $\Psi(x)$.

Какова вероятность попадания очередного измерения, имеющего распределение $\Phi(x)$, в интервал $[\underline{x}, \overline{x}]$ уже измеренных значений?

Случай нескольких «измерений-примесей»

Здесь получаем асимптотику стремления достоверности интервального агрегирования к 1:

$$1 - \frac{2}{n-k+1} - \frac{1}{n-k} \leq$$

$$\mathsf{P}\left(x_{n+1} \in \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right] \right) \qquad (**)$$

$$\leq 1.$$

Она аналогична найденным ранее.

Случай бесконечной серии «измерений-примесей»

Полученные результаты применимы также к случаю, когда количество «измерений-примесей» неограниченно растёт с увеличением выборки.

Так как никаких условий на конечность или фиксированность числа примесей k при выводе оценки (**) не накладывается, то возможна ситуация, когда k также растёт, не будучи ничем ограниченным.

Неравенства (**) всё равно будут справедливы.

Например, k может быть равным какой-то фиксированной доле от общего количества измерений n, скажем, половине n, когда $k \approx n/2$.

Тогда, в частности, оценка (**) превратится в такую:

$$1 - \frac{4}{n+2} - \frac{2}{n} \leq$$

$$\mathsf{P}\left(x_{n+1} \in \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right] \right)$$

$$\leq 1.$$

При более жёстких условиях на распределение «измерений-примесей» можно получить более тонкие оценки, аналогичные найденным ранее.

Итоги

1. Между интервальным анализом данных и традиционной теоретико-вероятностной статистикой существуют связи и «мосты», которые полезно применять при решении практических задач.

   Многие методы интервального анализа данных можно интерпретировать в вероятностном духе.

2. Интервальное агрегирование данных — популярный инструмент получения интервальных данных.

   Его доверительная вероятность, как правило, стремится к 1 с быстротой, пропорциональной $1/n$, n — объём выборки.

Спасибо за внимание!