Вычислительные технологии, 2024, том 29, № 4, с. 55–70. © ФИЦ ИВТ, 2024
Computational Technologies, 2024, vol. 29, no. 4, pp. 55–70. © FRC ICT, 2024
ISSN 1560-7534
eISSN 2313-691X

DOI:10.25743/ICT.2024.29.4.005

Медиана интервальных данных

А. В. Пролубников
Омский государственный университет, 644077, Омск, Россия
Контактный автор: Пролубников Александр Вячеславович, e-mail: a.v.prolubnikov@mail.ru

Поступила 11 сентября 2023 г., доработана 17 октября 2023 г., принята в печать 14 марта 2024 г.

Медиана набора чисел — это число, которое находится в середине этого набора, если его упорядочить по неубыванию. В работе производится обобщение определения медианы числовых значений для данных с интервальной неопределенностью. В этом случае при определении медианы необходимо задать отношение линейного порядка для интервальных данных и найти для них частоту встречаемости. Эта задача решается нами за счет перехода от исходных интервалов к набору их подынтервалов, объединением которых могут быть получены исходные интервалы. Обобщение производится с соблюдением принципа соответствия: при стремлении ширин интервалов к нулю значение медианы, рассчитанной для интервалов, стремится к значению медианы точечных значений, к которым стремятся сужаемые интервалы. Производится сравнение определяемой в работе медианы с интервальной медианой, определяемой как интервал, концы которого являются соответственно медианами левых и правых концов интервалов выборки. Показывается, что в случае если не одно, а все или некоторое множество допускаемых интервалами выборки точечных значений могут быть истинными, то определяемая интервальная медиана более адекватно оценивает их точечное медианное значение.

Ключевые слова: интервальный анализ.

Цитирование: Пролубников А.В. Медиана интервальных данных. Вычислительные технологии. 2024; 29(4):55–70. DOI:10.25743/ICT.2024.29.4.005.

Введение

Медиана набора чисел — это серединное значение этого набора, т. е. число, которое находится в середине набора, если его упорядочить по неубыванию. Выборка — это конечное множество значений, полученных при измерении некоторой величины. Медиана — это статистическая характеристика выборки (статистика), используемая при эмпирическом оценивании распределения через серединное значение при центрировании данных и их нормализации.

В отличие от среднего арифметического, медиана является робастной статистикой выборки, т. е. она нечувствительна к выбросам — результатам измерений, сильно отличающимся от значений большинства элементов выборки. Это значит, что если в выборку включить слишком малое или слишком большое значение относительно уже содержащихся в ней, то медиана не изменится или изменится мало, тогда как среднее арифметическое сильно колеблется при включении таких значений в выборку.

Кроме того, эмпирическое распределение, значительно отличающееся от нормального, не позволяет применять к нему методы, разработанные для выборок, распределение для которых предполагается близким к нему. Среднее арифметическое пригодно в ка-

честве выбора серединного значения, только если эмпирическое распределение близко к нормальному. Выбор медианы в качестве серединного значения выборки позволяет адекватнее оценивать распределение эмпирических данных еще и потому, что на него не влияет его асимметрия.

Числовые и интервальные данные, для которых может быть вычислена медиана, рассматриваются нами далее как выборка, хотя это могут быть данные, не являющиеся полученной в результате эксперимента частью некоторой генеральной совокупности. Это может быть набор компонент вектора, набор интервалов, являющийся множеством значений функции или его интервальной оценкой, и др.

Под точечными данными мы понимаем выборку $X = \{x_1, x_2, \ldots, x_N\} = \{x_i\}_{i=1}^N$, элементы которой являются вещественными оценками неизвестных истинных значений, т. е. $x_i \in \mathbb{R}$. Для определения медианы выборки X строится *вариационный ряд* $\tilde{X}$, состоящий из элементов выборки, упорядоченных по неубыванию. Элемент вариационного ряда называется *вариантой*. Элементы выборки и построенного для нее вариационного ряда могут повторяться. Количество повторений элемента $x_i \in X$ — *частоту (частоту встречаемости)* значения x_i в выборке — будем обозначать как $f(x_i)$.

В литературе по статистике даются следующие равнозначные определения медианы.

Определение 1. Медиана $\operatorname{med} X$ выборки X — это значение варианты $x_m \in \tilde{X}$, для которой половина вариант из $\tilde{X}$ с учетом их частот лежит слева, а половина — справа.

Определение 2. Медиана $\operatorname{med} X$ выборки X — это значение варианты $x_m \in \tilde{X}$, для которой минимальна сумма расстояний от нее до других вариант из $\tilde{X}$ с учетом их частот.

Отдельно оговаривается случай четного N, когда в качестве медианы выбирается либо одна из вариант с индексом $m = \lfloor N/2 \rfloor$ или $m = \lceil N/2 \rceil$, либо обе эти варианты, либо их арифметическое среднее.

Далее нами также используется определение медианы сгруппированных данных med_g. med_g — это статистика выборки, данные в которой заданы не точно, а указана лишь их принадлежность некоторым интервалам.

При использовании точечных данных предполагается, что либо данные нам известны точно, т. е. известны их истинные значения, либо даны их достаточно точные оценки. Либо данные получены в результате усреднения, под которым обычно понимается взятие среднего арифметического от результатов многократно проведенных измерений. Как правило, предполагается, что имеющаяся точность оценок истинных значений достаточна для решения задачи. Однако во многих случаях это не так, и неопределенность во входных данных оказывает существенное влияние как на методы решения, которые могут быть применены к задаче, так и на соответствие действительности получаемых с их помощью результатов.

По этой причине часто на практике мы имеем дело не с точечной, а с интервальной выборкой, элементы которой — это интервалы из $\mathbb{R}$. В этой работе производится обобщение понятия медианы на случай таких — *интервальных* — данных. Интервальные значения будем далее обозначать жирным шрифтом: $\boldsymbol{a} = [\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}] = \{a \in \mathbb{R} \mid \underline{\boldsymbol{a}} \le a \le \overline{\boldsymbol{a}}\}$, где $\underline{\boldsymbol{a}}$ — нижняя граница интервала, $\overline{\boldsymbol{a}}$ — верхняя, $\underline{\boldsymbol{a}} \le \overline{\boldsymbol{a}}$. Будем говорить, что интервал $\boldsymbol{a}$ *вырожден*, если $\underline{\boldsymbol{a}} = \overline{\boldsymbol{a}}$. Как $\mathbb{IR}$ будем обозначать заданное таким образом на $\mathbb{R}$ множество интервалов. Арифметические операции для интервалов $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{IR}$ определяются исходя из принципа $\boldsymbol{a} \star \boldsymbol{b} := \{a \star b \mid a \in \boldsymbol{a}, b \in \boldsymbol{b}\}$, где $\star \in \{+, -, \cdot, /\}$. При этом требуется, чтобы точечная операция $a \star b$ имела смысл для любых $a \in \boldsymbol{a}$ и $b \in \boldsymbol{b}$.

В этой работе при определении интервальной медианы рассматриваются только ограниченные интервалы. Интервальная выборка $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{IR}$, представляет собой выборку оценок неизвестных нам истинных значений $x_i : \underline{\boldsymbol{x}}_i \le x_i \le \overline{\boldsymbol{x}}_i$.

Как отмечается в [1], выборка с точечными значениями, далекими от истинных, может давать значения статистик, близкие к получаемым по выборке с точечными элементами, значения которых близки к истинным. В случае интервальных данных разница в качестве данных, т. е. в точности интервальных оценок, выражаемая шириной интервалов выборки, влияет на результат вычисления интервальных статистик и, в частности, на результат вычисления медианы. Интервальные значения медианы, получаемой для интервалов небольшой ширины, т. е. для более точных оценок, и медианы, получаемой для достаточно широких интервалов с теми же центрами, что и исходные, будут значительно различаться.

Неопределенность в значениях входных данных может присутствовать по причине неполноты информации на момент начала исследований и возможности ее изменений с течением времени. Неопределенность может быть обусловлена также неустранимой погрешностью измерений. При этом вероятностное распределение величины ошибки для измерительных устройств часто не является нормальным, и достаточного объема информации для выбора распределения и его параметров может не быть [2, 3].

По тем же причинам, что и в точечном случае, для оценивания интервальной выборки "в среднем", ее центрирования и нормализации необходимо использовать медиану, определенную с учетом специфики интервального представления данных. В частности, необходимо учесть то обстоятельство, что размеры точечных выборок, по которым формируются интервальные данные, как правило, невелики для того, чтобы предполагать наличие некоторого известного распределения ошибки измерений на интервале.

В [1] даны определения арифметического среднего, медианы, моды, среднеквадратичного отклонения и других статистик интервальной выборки. Базовая концепция, используемая в [1] при определении статистик для интервальных данных, — это концепция *конфигурации точек*, под которой понимается множество точек $c = \{x_1, \ldots, x_N\}$, взятых по одной из каждого интервала выборки $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$: $x_i \in \boldsymbol{x}_i$. Благодаря такому подходу решается проблема введения линейного порядка на интервалах, возникающая при определении медианы для интервальных данных, поскольку при работе с конфигурациями точек используется линейный порядок на $\mathbb{R}$. Однако для величин, представляемых интервалами из $\mathbb{IR}$, порядок, в том числе линейный, может быть введен по-разному и, в соответствии с решаемой задачей, может по-разному отражать соотношения интервалов выборки.

Медиана интервальной выборки в [1] определяется следующим образом. Пусть дана выборка $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$. Пусть $\underline{c} = \{\underline{\boldsymbol{x}}_i\}_{i=1}^N$, $\overline{c} = \{\overline{\boldsymbol{x}}_i\}_{i=1}^N$ — конфигурации точек, составленные соответственно из левых и правых концов интервалов из $\boldsymbol{X}$.

Определение 3. Медиана $\mathbf{med}_k \boldsymbol{X}$ интервальной выборки $\boldsymbol{X}$ — это интервал

$$\mathbf{med}_k \boldsymbol{X} = [\operatorname{med} \underline{c}, \operatorname{med} \overline{c}].$$

Сложность вычисления $\mathbf{med}_k \boldsymbol{X}$ составляет $O(N)$ [4].

Определение интервальной медианы, которое дается в нашей работе, является более обоснованным в случае, если интервалы выборки содержат не одно точечное истинное значение величины, измеряемое с помощью какого-либо измерительного прибора, а представляют собой совокупности значений, каждое из которых эта величина гарантированно принимает при некоторых условиях. Используемый нами подход позволяет

учесть частоту ее возможных точечных значений, тогда как медиана, определяемая по концам интервалов (определение 3), не всегда позволяет это сделать.

Давая определения медианы интервальных данных, мы исходим из *принципа соответствия*, который может быть сформулирован так: при стремлении ширин интервалов к нулю значения медианы, рассчитанные по ее определению для интервальных данных, стремятся к значениям, рассчитанным по определению для точечных значений, к которым стремятся интервалы.

Интервальная медиана может быть использована для тех же целей, что и медиана точечных данных, т. е. для выбора серединного значения при оценивании распределения эмпирических и теоретических данных с интервальной неопределенностью, их центрировании и нормализации.

Так, например, в [5, 6] дается интервальный жадный алгоритм — обобщение жадного алгоритма решения задач дискретной оптимизации на графах и гиперграфах на случай интервальных весов ребер (гиперребер), которые являются коэффициентами линейной интервальной функции. Алгоритм дает множество всех возможных оптимальных решений для значений весов из заданных интервалов, как, например, для задачи о минимальном остовном дереве с интервальными весами ребер, или множество всех ее приближенных решений с гарантированной оценкой точности, как, например, для общего случая задачи о покрытии с интервальными весами множеств. Он позволяет получить интервалы возможных значений целевой функции (функции потерь) для них и подынтервалы весов, при которых эти значения достигаются. Даже для задач относительно небольшой размерности количество таких решений и, соответственно, интервалов возможных точечных значений целевой функции для этих решений может быть весьма велико, что затрудняет анализ. Чтобы интервально оценить точечное серединное значение целевой функции для всех возможных решений задачи и всех возможных точечных значений весов, может быть использована интервальная медиана. Далее мы вернемся к этому примеру.

1. Каким требованиям должна удовлетворять медиана интервальных данных

Определение интервальной медианы должно удовлетворять следующим требованиям. Как математическое понятие она должна обобщать понятие точечной медианы и удовлетворять принципу соответствия. Как определяемая статистическая характеристика интервальной выборки она должна давать обоснованную оценку серединного значения возможных точечных данных, допускаемых интервалами выборки. Кроме того, мы требуем, чтобы определяемая нами интервальная медиана содержала только допускаемые интервалами выборки точечные значения и не содержала значений, выборкой не допускаемых.

Определяемая нами интервальная медиана, обозначаемая далее как $\mathbf{med}_p$, обобщает как определение медианы точечной выборки, так и определение медианы для сгруппированных точечных данных. Определения интервальной медианы, аналогичные приведенным определениям медианы точечной, мы даем с помощью элементарных подынтервалов, на которые разбиваем элементы интервальной выборки. Использование элементарных подынтервалов позволяет:

– задать линейный порядок для интервальных данных;
– определить частоту интервальных данных.

2. Элементарные подынтервалы

Для того чтобы распространить определение медианы точечных данных на интервальные данные, требуется задать для них линейный порядок ($\leq$) и частоту.

Возможные способы задания отношения порядка на $\mathbb{IR}$ определены стандартом 1788 IEEE [7]. Говорят, что неравенство $\boldsymbol{a} \leq \boldsymbol{b}$ выполняется:

а) в сильном смысле, если $(\forall a \in \boldsymbol{a})\ (\forall b \in \boldsymbol{b})\ (a \leq b)$, $(\overline{\boldsymbol{a}} \leq \underline{\boldsymbol{b}})$;
б) слабом смысле, если $(\exists a \in \boldsymbol{a})\ (\exists b \in \boldsymbol{b})(a \leq b)$, $(\underline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}})$;
в) $\forall\exists$-смысле, если $(\forall a \in \boldsymbol{a})\ (\exists b \in \boldsymbol{b})(a \leq b)$, $(\overline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}})$;
г) $\exists\forall$-смысле, если $(\exists a \in \boldsymbol{a})\ (\forall b \in \boldsymbol{b})(a \leq b)$, $(\underline{\boldsymbol{a}} \leq \underline{\boldsymbol{b}})$;
д) центральном смысле, если $(\overline{\boldsymbol{a}} + \underline{\boldsymbol{a}})/2 \leq (\overline{\boldsymbol{b}} + \underline{\boldsymbol{b}})/2$.

Отношения порядка "$\leq$" в слабом смысле, в $\forall\exists$-смысле, в $\exists\forall$-смысле не являются отношениями линейного порядка на $\mathbb{IR}$, т. е. для $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}$, $\boldsymbol{a} \neq \boldsymbol{b}$, может быть одновременно выполнено как $\boldsymbol{a} \leq \boldsymbol{b}$, так и $\boldsymbol{b} \leq \boldsymbol{a}$. Для отношения порядка в сильном смысле два неравных интервала могут быть не сравнимы. Любые интервалы из $\mathbb{IR}$ сравнимы в соответствии с порядком в центральном смысле, что позволяет рассчитывать точечную медиану интервальной выборки как медиану для центров интервалов выборки.

Как отмечается в [1], при определении медианы, моды и других статистик интервальных данных возникают сложности с тем, что считать их частотой. Например, пересечение двух интервалов может быть велико, если они соответствуют одному и тому же истинному значению, измеренному с некоторыми незначительно различающимися погрешностями. Если такие не равные, но близкие друг к другу интервалы выборки рассматривать как отдельные совокупности точечных данных, то частоты всех элементов выборки могут оказаться равными единице, что будет плохо отражать статистическую структуру интервальных данных.

Для того чтобы определить для интервальных данных отношение линейного порядка и частоту, мы используем разбиение $\tilde{\boldsymbol{X}}$ исходных интервалов выборки $\boldsymbol{X}$ на интервалы $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_T$, где $\tilde{\boldsymbol{x}}_i$ — подынтервалы интервалов из $\boldsymbol{X}$. Это разбиение задается концами интервалов выборки. $\tilde{\boldsymbol{X}}$ — линейно упорядоченное множество *элементарных подынтервалов* интервалов из $\boldsymbol{X}$ в соответствии с любым отношением порядка из приведенных выше.

Разбиение $\tilde{\boldsymbol{X}}$ — это минимальное по мощности множество упорядоченных подынтервалов интервалов из $\boldsymbol{X}$, такое, что

$$(\forall \boldsymbol{x}_i \in \boldsymbol{X})\ (\exists j_1, \ldots, j_t)\ \left(\boldsymbol{x}_i = \bigcup_{i=1}^{t} \tilde{\boldsymbol{x}}_{j_i}\right),$$

т. е. интервалы исходной выборки представляются в виде объединений элементарных подынтервалов.

Частота встречаемости точечного значения x такого, что $x \in \boldsymbol{x}_i$, $\boldsymbol{x}_i \in \boldsymbol{X}$, т. е. допускаемого интервальной выборкой, — это количество интервалов выборки $\boldsymbol{X}$, которые содержат x. *Частота встречаемости элементарного подынтервала* $\tilde{\boldsymbol{x}} \in \tilde{\boldsymbol{X}}$ — это количество интервалов выборки, которые содержат $\tilde{\boldsymbol{x}}$.

Элементарный подынтервал составляют точечные значения, допускаемые выборкой и имеющие одинаковую частоту встречаемости. Каждый интервал из $\tilde{\boldsymbol{X}}$ — это максимальный по включению подынтервал одного или нескольких интервалов из $\boldsymbol{X}$, содержащий возможные точечные значения с одинаковой частотой. Так, например, для выборки $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}\}$ (рис. 1) имеем разбиение, состоящее из пяти элементарных

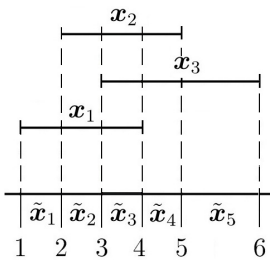

Рис. 1. Разбиение на элементарные подынтервалы для выборки $\boldsymbol{X}$
Fig. 1. Splitting into elementary subintervals of the sample $\boldsymbol{X}$

подынтервалов $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \tilde{\boldsymbol{x}}_3, \tilde{\boldsymbol{x}}_4, \tilde{\boldsymbol{x}}_5\}$. Любой интервал, включающий, например, элементарный интервал $\tilde{\boldsymbol{x}}_3$, в отличие от самого $\tilde{\boldsymbol{x}}_3$, будет содержать возможные точечные значения с различной частотой встречаемости, а не только равной 3.

Алгоритм получения разбиения $\tilde{\boldsymbol{X}}$

Вход: $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ — интервальная выборка.

Шаг 1. $\tilde{\boldsymbol{X}} \leftarrow \varnothing$.

Шаг 2. Получить множество $\dot{\boldsymbol{X}}$ всех вырожденных интервалов из $\boldsymbol{X}$:
2.1) $\dot{\boldsymbol{X}} \leftarrow \varnothing$;
2.2) для всех $i = \overline{1, N}$:
если $\underline{\boldsymbol{x}}_i = \overline{\boldsymbol{x}}_i$, то $\dot{\boldsymbol{X}} \leftarrow \dot{\boldsymbol{X}} \cup \{\boldsymbol{x}_i\}$.

Шаг 3. Сформировать неупорядоченное множество M_0 концов невырожденных интервалов из $\boldsymbol{X}$:
3.1) $M_0 \leftarrow \varnothing$;
3.2) для всех $i = \overline{1, N}$:
если $\underline{\boldsymbol{x}}_i \neq \overline{\boldsymbol{x}}_i$ и $\underline{\boldsymbol{x}}_i \notin M_0$, то $M_0 \leftarrow M_0 \cup \{\underline{\boldsymbol{x}}_i\}$;
если $\underline{\boldsymbol{x}}_i \neq \overline{\boldsymbol{x}}_i$ и $\overline{\boldsymbol{x}}_i \notin M_0$, то $M_0 \leftarrow M_0 \cup \{\overline{\boldsymbol{x}}_i\}$.

Шаг 4. Получить множество M упорядочиванием по возрастанию элементов множества M_0, $\operatorname{card} M_0 = S$:
$M = \{x_1, \ldots, x_S\}$, $x_j \in M_0$, $x_j < x_{j+1}$, $j = \overline{1, S-1}$.

Шаг 5. Сформировать упорядоченное по возрастанию множество $\tilde{\boldsymbol{X}}$:
5.1) для всех $j = \overline{1, S}$:
если $\exists \boldsymbol{x}_i \in \boldsymbol{X}$ такой, что $[x_j, x_{j+1}] \subseteq \boldsymbol{x}_i$, то $\tilde{\boldsymbol{X}} \leftarrow \tilde{\boldsymbol{X}} \cup \{[x_j, x_{j+1}]\}$;
5.2) поместить элементы из $\dot{\boldsymbol{X}}$ в $\tilde{\boldsymbol{X}}$ в соответствии с порядком на $\tilde{\boldsymbol{X}}$.

Выход: упорядоченное множество элементарных подынтервалов $\tilde{\boldsymbol{X}}$.

3. Медиана med_p интервальных данных

Группировка данных является одним из источников возникновения интервальной неопределенности. При группировке данных происходит интервализация точечных данных: не зная истинных значений элементов точечной выборки, имеем разбиение интервала возможных значений измеряемой величины на его подынтервалы, которым принадлежат истинные значения. Медиана сгруппированных данных med_g — это точечная оценка выборки с интервальной неопределенностью.

Далее производим модификацию med_g, получая в результате определение интервальной медианы, обобщающее и определение 1 на случай интервальных данных.

3.1. Интервальное обобщение медианы точечных данных

Разбиение интервалов выборки $\boldsymbol{X}$ на элементарные подынтервалы $\tilde{\boldsymbol{X}}$ позволяет рассматривать ее как выборку, полученную группировкой точечных данных, с тем отличием, что входящие в нее вырожденные интервалы рассматриваются нами как отдельные элементарные подынтервалы. Иначе говоря, если $\tilde{\boldsymbol{x}}_i = [x, x]$, $\tilde{\boldsymbol{x}}_j = [a, b]$ и $a \leq x \leq b$, то частоты $\tilde{\boldsymbol{x}}_i$ и $\tilde{\boldsymbol{x}}_j$ рассчитываются отдельно. Для расчета частот вырожденных интервалов используются только точечные элементы выборки.

Пример 1. В табл. 1 задано распределение по возрастным группам студентов, обучающихся на заочных отделениях крупного университета. Этой таблице соответствует точечная выборка $X = \{x_i\}_{i=1}^{3462}$, где x_i — возраст i-го студента. Однако в результате группировки данных значение x_i задается не точно, а задана лишь принадлежность его одному из интервалов $\tilde{\boldsymbol{x}}_i$ разбиения $\tilde{\boldsymbol{X}}$ всего интервала возможных значений возраста $[16, 60]$: $\tilde{\boldsymbol{X}} = \{[16, 20],\ [20, 25],\ [25, 30],\ [30, 35],\ [35, 40],\ [40, 45],\ [45, 60]\}$. То есть фактически вместо выборки X получаем интервальную выборку $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_i\}_{i=1}^{3462}$, состоящую из упорядоченных интервалов $\tilde{\boldsymbol{x}}_i$, которые могут пересекаться только по их границам. Частота $f(\tilde{\boldsymbol{x}}_i)$ — это количество студентов, попадающих в задаваемый интервалом $\tilde{\boldsymbol{x}}_i$ возрастной диапазон.

В литературе по статистике дается следующее определение точечной медианы сгруппированных данных med_g для таких по своей сути интервальных выборок. Пусть истинные значения элементов точечной выборки $X = \{x_i\}_{i=1}^{N}$ неизвестны, а известна лишь их принадлежность заданным интервалам разбиения $\tilde{\boldsymbol{X}}$ их области возможных значений, т. е. $x_i \in \tilde{\boldsymbol{x}}_j$ для некоторого $\tilde{\boldsymbol{x}}_j \in \tilde{\boldsymbol{X}}$. Пусть $f(\tilde{\boldsymbol{x}}_i)$ — частота интервала $\tilde{\boldsymbol{x}}_i$, равная количеству $x_i \in X$, попадающих в интервал $\tilde{\boldsymbol{x}}_i$. $\operatorname{wid} \boldsymbol{x} = \overline{\boldsymbol{x}} - \underline{\boldsymbol{x}}$ — ширина интервала $\boldsymbol{x}$, $S_n = \sum\limits_{i=1}^{n} f(\tilde{\boldsymbol{x}}_i)$.

Заметим, что для элементов выборки $\tilde{\boldsymbol{X}}$ определен линейный порядок в соответствии с любым из пяти приведенных выше отношений порядка на $\mathbb{IR}$, т. е. если $i \neq j$, то либо $\tilde{\boldsymbol{x}}_i \leq \tilde{\boldsymbol{x}}_j$, либо $\tilde{\boldsymbol{x}}_i \geq \tilde{\boldsymbol{x}}_j$ для любого из этих отношений порядка. Это позволяет рассматривать $\tilde{\boldsymbol{X}}$ как интервальный вариационный ряд.

Определение 4. *Медиана сгруппированных данных* med_g определяется как значение, вычисляемое по следующему алгоритму:

Шаг 1. Определяем *медианный интервал* — варианту из $\tilde{\boldsymbol{X}}$, которая делит $\tilde{\boldsymbol{X}}$ на две равные части. Пусть m — номер такой варианты $\tilde{\boldsymbol{x}}_m$.
Шаг 2. Вычисляем медиану $\operatorname{med}_g \tilde{\boldsymbol{X}}$:

$$\operatorname{med}_g \tilde{\boldsymbol{X}} = \min\ rm\tilde{\boldsymbol{x}}_m + \operatorname{wid} \tilde{\boldsymbol{x}}_m \frac{\frac{1}{2}\sum\limits_{i=1}^{N} f(\tilde{\boldsymbol{x}}_i) - S_{m-1}}{f(\tilde{\boldsymbol{x}}_m)}.$$

Значение медианы med_g — это точка из медианного интервала $\tilde{\boldsymbol{x}}_m$, смещенная относительно его начала вправо тем больше, чем меньше $f(\tilde{\boldsymbol{x}}_m)$ и чем дальше от $\tilde{\boldsymbol{x}}_m$ находятся элементы выборки, равные $\tilde{\boldsymbol{x}}_{m-1}$.

Для данных выборки, приведенных в табл. 1, имеем $\sum\limits_{i=1}^{N} f(\tilde{\boldsymbol{x}}_i)/2 = 1731$, следовательно, медианный интервал — это интервал $[25, 30]$. В результате получаем

$$\operatorname{med}_g \tilde{\boldsymbol{X}} = 25 + 5\frac{1731 - 1218}{1054} = 27.4.$$

Т а б л и ц а 1. Распределение студентов по возрастным группам
Table 1. Distribution of students by age groups

Возрастная группа $\tilde{\boldsymbol{x}}_i$	Количество студентов в возрастной группе $f(\tilde{\boldsymbol{x}}_i)$	Сумма частот предшествующих интервалов и данного интервала
16–20	346	346
20–25	872	1218
25–30	1054	2272
30–35	781	3053
35–40	212	3265
40–45	121	3386
45–60	76	3462

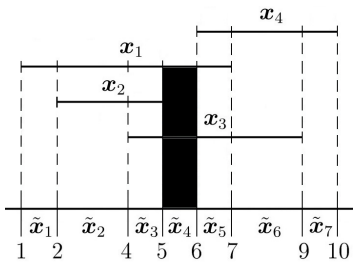

Рис. 2. Пример 2, иллюстрирующий расчет $\mathbf{med}_p$
Fig. 2. Example 2 that illustrates calculation of $\mathbf{med}_p$

Т а б л и ц а 2. Разбиение $\tilde{\boldsymbol{X}}$
Table 2. The partition $\tilde{\boldsymbol{X}}$

$\tilde{\boldsymbol{x}}_i$	$\tilde{\boldsymbol{x}}_1$	$\tilde{\boldsymbol{x}}_2$	$\tilde{\boldsymbol{x}}_3$	$\tilde{\boldsymbol{x}}_4$	$\tilde{\boldsymbol{x}}_5$
$f(\tilde{\boldsymbol{x}}_i)$	1	2	3	2	1

Т а б л и ц а 3. Разбиение $\tilde{\boldsymbol{X}}$
Table 3. The partition $\tilde{\boldsymbol{X}}$

$\tilde{\boldsymbol{x}}_i$	$\tilde{\boldsymbol{x}}_1$	$\tilde{\boldsymbol{x}}_2$	$\tilde{\boldsymbol{x}}_3$	$\tilde{\boldsymbol{x}}_4$	$\tilde{\boldsymbol{x}}_5$	$\tilde{\boldsymbol{x}}_6$	$\tilde{\boldsymbol{x}}_7$	$\tilde{\boldsymbol{x}}_8$	$\tilde{\boldsymbol{x}}_9$	$\tilde{\boldsymbol{x}}_{10}$
$f(\tilde{\boldsymbol{x}}_i)$	1	2	1	2	1	1	2	1	2	1

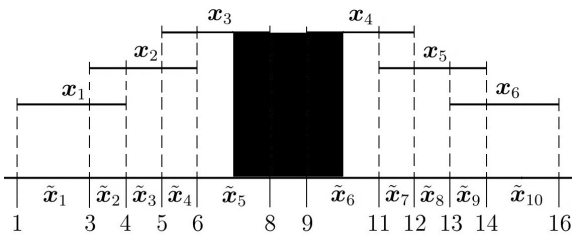

Рис. 3. Пример 3. Пункт А: $\mathbf{med}_p\,\boldsymbol{X} = [7, 10]$
Fig. 3. Example 3. Paragraph A: $\mathbf{med}_p\,\boldsymbol{X} = [7, 10]$

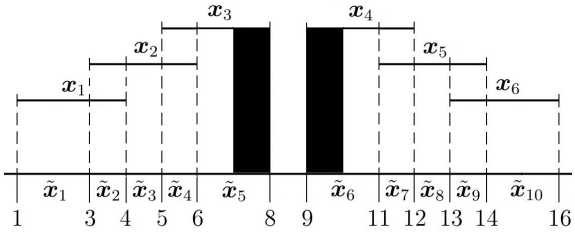

Рис. 4. Пример 3. Пункт Б: $\mathbf{med}_p\,\boldsymbol{X} = [7, 8] \cup [9, 10]$
Fig. 4. Example 3. Paragraph Б: $\mathbf{med}_p\,\boldsymbol{X} = [7, 8] \cup [9, 10]$

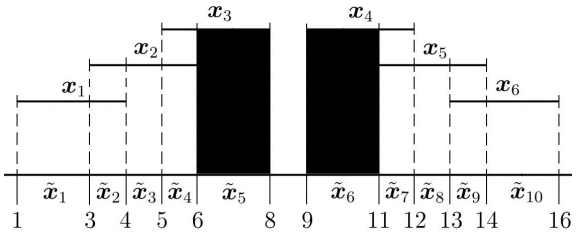

Рис. 5. Пример 3. Пункт В: $\mathbf{med}_p\,\boldsymbol{X} = [6, 8] \cup [9, 11]$
Fig. 5. Example 3. Paragraph В: $\mathbf{med}_p\,\boldsymbol{X} = [6, 8] \cup [9, 11]$

Определим интервальную медиану $\mathbf{med}_p$.

Определение 5. Медиана $\mathbf{med}_p \boldsymbol{X}$ выборки $\boldsymbol{X}$ — это значение варианты $\tilde{\boldsymbol{x}}_m \in \tilde{\boldsymbol{X}}$, для которой половина вариант из $\tilde{\boldsymbol{X}}$ с учетом их частот лежит слева, а половина — справа.

Пример 2 (рис. 2) иллюстрирует расчет интервальной медианы $\mathbf{med}_p \boldsymbol{X}$.

Пример 2. Пусть $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^3$, $\boldsymbol{x}_1 = [1,4]$, $\boldsymbol{x}_2 = [2,5]$, $\boldsymbol{x}_3 = [3,6]$. $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_i\}_{i=1}^5$, $\tilde{\boldsymbol{x}}_i = [i, i+1]$. Частоты интервалов разбиения $\tilde{\boldsymbol{X}}$ заданы в табл. 2. $\mathbf{med}_p \boldsymbol{X} = \tilde{\boldsymbol{x}}_3 = [3,4]$.

В ситуации, когда имеются два элементарных подынтервала $\tilde{\boldsymbol{x}}_m$ и $\tilde{\boldsymbol{x}}_{m+1}$, расположенных посередине вариационного ряда, и $\tilde{\boldsymbol{x}}_m \neq \tilde{\boldsymbol{x}}_{m+1}$ (пример 3), медиана может быть определена одним из следующих способов.

А. Естественным обобщением взятия полусуммы точечных значений, расположенных посередине вариационного ряда из точечных значений, в случае интервальной выборки является взятие полусуммы интервалов $\tilde{\boldsymbol{x}}_m$ и $\tilde{\boldsymbol{x}}_{m+1}$:

$$\mathbf{med}_p \boldsymbol{X} = (\tilde{\boldsymbol{x}}_m + \tilde{\boldsymbol{x}}_{m+1})/2.$$

Однако в случае, если $\tilde{\boldsymbol{x}}_m \cap \tilde{\boldsymbol{x}}_{m+1} = \varnothing$, т. е. эти интервалы не пересекаются по границе, получаемый таким образом медианный интервал будет содержать точечные значения, не содержащиеся в интервалах исходной выборки. Для примера 3 это значения из интервала $]8,9[$ (рис. 3).

Б. Более обоснованным считаем следующий вариант обобщения точечной медианы, лишенный этого недостатка. Значение интервальной медианы в случае двух неравных элементарных подынтервалов, расположенных посередине вариационного ряда, определим как

$$\mathbf{med}_p \boldsymbol{X} = \big((\tilde{\boldsymbol{x}}_m + \tilde{\boldsymbol{x}}_{m+1})/2\big) \cap (\tilde{\boldsymbol{x}}_m \cup \tilde{\boldsymbol{x}}_{m+1}). \tag{1}$$

Если $\tilde{\boldsymbol{x}}_m \cap \tilde{\boldsymbol{x}}_{m+1} = \varnothing$, $\mathbf{med}_p \boldsymbol{X}$ является мультиинтервалом (пример 3, рис. 4).

В. $\mathbf{med}_p \boldsymbol{X}$ может быть также определена как объединение двух элементарных подынтервалов, расположенных посередине вариационного ряда (пример 3, рис. 5):

$$\mathbf{med}_p \boldsymbol{X} = \tilde{\boldsymbol{x}}_m \cup \tilde{\boldsymbol{x}}_{m+1}.$$

Таким образом, при наличии двух элементарных подынтервалов $\tilde{\boldsymbol{x}}_m$ и $\tilde{\boldsymbol{x}}_{m+1}$, расположенных посередине вариационного ряда, определим $\mathbf{med}_p$ так, как указано в п. Б, поскольку в этом случае серединные значения выбираются более обоснованно. Так, для выборки, состоящей из двух непересекающихся интервалов, интервальная медиана, определяемая так, как указано в п. В, не даст никакой информации о серединном интервальном значении выборки, поскольку в качестве серединных значений она будет содержать все возможные точечные значения, допускаемые выборкой.

Пример 3. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^6$, $\boldsymbol{x}_1 = [1,4]$, $\boldsymbol{x}_2 = [3,6]$, $\boldsymbol{x}_3 = [5,8]$, $\boldsymbol{x}_4 = [9,12]$, $\boldsymbol{x}_5 = [11,14]$, $\boldsymbol{x}_6 = [13,16]$ (см. рис. 3–5). Частоты интервалов разбиения $\tilde{\boldsymbol{X}}$ заданы в табл. 3.

4. Интервальная медиана $\mathbf{med}_d$ и принцип соответствия

Для того чтобы распространить определение 2 медианы точечной выборки на случай интервальных данных, необходимо ввести функцию расстояния для интервалов интервального вариационного ряда. В качестве такого расстояния удобно выбрать *хаусдорфово расстояние* dist, определяемое следующим образом. Пусть A и B — компактные множества в $\mathbb{R}^n$, $\mathrm{dist} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ — расстояние на $\mathbb{R}^n$, тогда

$$\operatorname{dist}(A,B)=\max\{\sup_{x\in A}\inf_{y\in B}\operatorname{dist}(x,y),\sup_{y\in B}\inf_{x\in A}\operatorname{dist}(x,y)\}.$$

Значение $\sup\limits_{x\in A}\inf\limits_{y\in B}\operatorname{dist}(x,y)$ равно максимуму минимального расстояния от точек из B до точек из A. Если $\boldsymbol{a},\boldsymbol{b}\in\mathbb{IR}$, то

$$\operatorname{dist}(\boldsymbol{a},\boldsymbol{b})=\max\{\sup_{x\in\boldsymbol{a}}\inf_{y\in\boldsymbol{b}}\operatorname{dist}(x,y),\sup_{y\in\boldsymbol{b}}\inf_{x\in\boldsymbol{a}}\operatorname{dist}(x,y)\}=\max\{|\underline{\boldsymbol{b}}-\underline{\boldsymbol{a}}|,|\overline{\boldsymbol{b}}-\overline{\boldsymbol{a}}|\}.$$

Выбор хаусдорфова расстояния на $\mathbb{IR}$ для определения медианы обусловлен его согласованностью с интервальной арифметикой [8]. Эта согласованность имеет место, поскольку хаусдорфово расстояние между интервалами естественным образом связано с растоянием между отдельными точками, им принадлежащими.

Определение 6. Интервальная медиана $\mathbf{med}_d$ интервальной выборки $\boldsymbol{X}$ — это такой элементарный подынтервал $\tilde{\boldsymbol{x}}_m\in\tilde{\boldsymbol{X}}$, для которого сумма хаусдорфовых расстояний от него до других элеметарных интервалов с учетом их частот минимальна.

В отличие от точечного случая, когда для любой выборки медианные значения, рассчитанные по определениям 1 и 2, совпадают, для интервальной медианы $\mathbf{med}_p$, задаваемой определением 5, и интервальной медианы $\mathbf{med}_d$, задаваемой определением 6, это, вообще говоря, не так. Однако определение 6 также задает интервальную медиану, удовлетворяющую приниципу соответствия. Значения интервальных медиан $\mathbf{med}_p$, $\mathbf{med}_d$ и $\mathbf{med}_k$ совпадают для выборок, содержащих узкие интервалы, т. е. таких выборок, что никакие два интервала из выборки не пересекаются, и стремятся к точечным значениям медианы при стремлении ширин интервалов выборки к нулю.

5. Вычислительная сложность нахождения $\mathbf{med}_p$

Сложность вычисления $\mathbf{med}_p\,\boldsymbol{X}$ для выборки $\boldsymbol{X}$ определяется вычислительной сложностью нахождения разбиения $\tilde{\boldsymbol{X}}$. Вычислительная сложность приведенного выше алгоритма нахождения $\tilde{\boldsymbol{X}}$ складывается из следующих значений вычислительной сложности его шагов: шаг 1 — вычислительная сложность $O(1)$, шаг 2 — $O(N)$, шаг 3 — $O(N)$, шаг 4 — $O(N\log N)$, шаг 5 — $O(N^2)$. Таким образом, суммарная вычислительная сложность нахождения $\tilde{\boldsymbol{X}}$ составляет $O(N^2)$. Сложность вычисления $\mathbf{med}_p\,\boldsymbol{X}$ по найденному разбиению $\tilde{\boldsymbol{X}}$, имеющему мощность $\operatorname{card}\tilde{\boldsymbol{X}}=O(N)$, составляет $O(N)$. Общая сложность вычисления $\mathbf{med}_p\,\boldsymbol{X}$, таким образом, составляет $O(N^2)$.

6. Когда необходимо использовать $\mathbf{med}_p$

6.1. Интервальная медиана как оценка точечных медиан для конфигураций возможных точечных значений

Интервальная медиана — это оценка медианного значения для всех возможных точечных значений, допускаемых выборкой. Рассмотрим два типа конфигураций таких значений, допускаемых интервальной выборкой $\boldsymbol{X}$:

1) конфигурации точек $x_i\in\boldsymbol{x}_i$, взятых по одной из каждого интервала $\boldsymbol{X}$:

$$c=\{x_1,\ldots,x_i,\ldots,x_N\};$$

2) конфигурации точек $x_i \in \tilde{\boldsymbol{x}}_i$, взятых из каждого элементарного подынтервала из $\tilde{\boldsymbol{X}}$ столько раз, какова его частота:

$$c = \{x_1^{f_1}, \ldots, x_i^{f_i}, \ldots, x_T^{f_T}\},$$

где $f_i = f(\tilde{\boldsymbol{x}}_i)$ — частота встречаемости $\tilde{\boldsymbol{x}}_i$ и принадлежащих ему точечных значений, $x_i^{f_i}$ — обозначает f_i точечных значений из $\tilde{\boldsymbol{x}}_i \in \tilde{\boldsymbol{X}}$, $T = \operatorname{card} \tilde{\boldsymbol{X}}$.

Обозначим множество всех конфигураций первого типа для выборки $\boldsymbol{X}$ как C_1, второго — как C_2. Имеем

$$(\forall c \in C_1)(\operatorname{med} c \in \mathbf{med}_k \boldsymbol{X}),$$
$$(\forall c \in C_2)(\operatorname{med} c \in \mathbf{med}_p \boldsymbol{X}),$$

т. е. $\mathbf{med}_k$ содержит точечные медианы для всех конфигураций из C_1, $\mathbf{med}_p$ — точечные медианы для всех конфигураций из C_2.

Оценивая медиану возможных точечных значений для кофигураций типа 1, составляющих множество C_1, предполагаем, что наблюдаемым является только одно возможное точечное значение из каждого интервала выборки. Тогда как оценивая ее для кофигураций типа 2, составляющих множество C_2, предполагаем, что наблюдаемыми являются все точечные значения, допускаемые выборкой, и используем содержащуюся в выборке информацию об их частоте.

Эта разница в подходах может быть сформулирована как разница в типах интервальной неопределенности, к которым могут быть применены $\mathbf{med}_k$ и $\mathbf{med}_p$. Под *неопределенностью типа I* будем понимать неопределенность следующего вида:

> *I. Каждый интервал $\boldsymbol{x}_i \in \boldsymbol{X}$ содержит одно наблюдаемое (истинное) значение измеряемой с погрешностью величины.*

Под *неопределенностью типа II* будем понимать неопределенность вида

> *II. Все точечные значения из каждого интервала $\boldsymbol{x}_i \in \boldsymbol{X}$ или некоторое заданное подмножество принадлежащих ему точечных значений являются наблюдаемыми, т. е. истинными при определенных условиях.*

В ситуации, когда имеется некоторое подмножество наблюдаемых точечных значений, допускаемых интервалами выборки, заданная нами интервальная медиана $\mathbf{med}_p$ также определена. Частоты таких значений будут учтены так же, как это делается при вычислении точечной медианы сгруппированных данных med_g.

Находя $\mathbf{med}_p$, оцениваем серединное значение для $\boldsymbol{X}$, используя $(T = \operatorname{card} \tilde{X})$ *интервальных* оценок для всех возможных конфигураций типа 2 — интервалы $\tilde{\boldsymbol{x}}_i \in \tilde{\boldsymbol{X}}$. Находя $\mathbf{med}_k$, мы используем $2N$ $(N = \operatorname{card} X)$ *точечных* оценок — концы интервалов $\boldsymbol{x}_i \in \boldsymbol{X}$. По этой причине значение $\mathbf{med}_k$ может содержать завышенные или заниженные относительно $\mathbf{med}_p$ серединные значения интервальной выборки, поскольку учет частот концов интервалов не всегда эквивалентен учету частот возможных точечных данных, входящих в интервалы выборки.

Поэтому использование $\mathbf{med}_k$ для оценки серединного значения интервальной выборки будет более обоснованным в случае интервальной неопределенности типа I. Тогда, как и в случае неопределенности типа II, обоснованным будет использование $\mathbf{med}_p$, что иллюстрируется примерами 4 и 6, приведенными ниже.

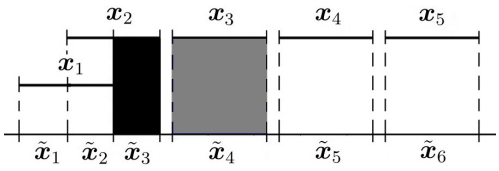

Рис. 6. Интервалы возможных значений целевой функции
Fig. 6. Intervals of possible values of the objective function

Пример 4. Рассмотрим пример выбора серединного значения целевой функции (функции потерь) для задач дискретной оптимизации на графах и гиперграфах в случае интервальных весов ребер (гиперребер). Предположим, что для задачи имеется 5 возможных оптимальных решений с интервалами возможных значений целевой функции, представленными на рис. 6: $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4,\boldsymbol{x}_5\}$, где $\mathbf{med}_p\,\boldsymbol{X}$ отмечена черным цветом, $\mathbf{med}_k\,\boldsymbol{X}$ — серым.

В рассматриваемом случае элементарные подынтервалы $\tilde{\boldsymbol{x}}_i$ представляют собой совокупности точечных значений, каждое из которых гарантированно принимает целевая функция при некоторых точечных значениях коэффициентов, принадлежащих заданным интервалам. Поскольку точки из элементарного подынтервала $\tilde{\boldsymbol{x}}_2$ встречаются среди таких значений вдвое чаще, чем значения из других элементарных подынтервалов $\tilde{\boldsymbol{x}}_i$, $i\neq 2$, это обстоятельство должно быть учтено при нахождении серединного интервального значения целевой функции для задачи дискретной оптимизации с интервальной целевой функцией. Оно учитывается при нахождении $\mathbf{med}_p\,\boldsymbol{X}$, когда серединным интервалом является $\tilde{\boldsymbol{x}}_3$: $\mathbf{med}_p\,\boldsymbol{X}=\tilde{\boldsymbol{x}}_3$, и не учитывается при нахождении $\mathbf{med}_k\,\boldsymbol{X}$, когда серединным интервалом является $\tilde{\boldsymbol{x}}_4$: $\mathbf{med}_k\,\boldsymbol{X}=\tilde{\boldsymbol{x}}_4$. То есть при вычислении $\mathbf{med}_k\,\boldsymbol{X}$ не учитывается повторяемость возможных точечных значений целевой функции из $\tilde{\boldsymbol{x}}_2$ и серединное значение интервальной целевой функции для задачи оптимизации из рассматриваемого примера оказывается завышенным.

Кроме того, отметим, что в соответствии с конфигурацией 1 $\mathbf{med}_p$ не содержит точечных значений, не принадлежащих выборке, тогда как $\mathbf{med}_k$ их содержать может (см. пример 3).

6.2. Примеры вычисления медиан для интервальных данных

Пример 5. Дана таблица распределения значений заработной платы по региону, т. е. заданы выборка $\boldsymbol{X}=\tilde{\boldsymbol{X}}$ и частоты интервалов $\tilde{\boldsymbol{x}}_i$ (табл. 4).

Пусть μX обозначает среднее арифметическое выборки $X=\{x_i\}_{i=1}^N$. X — выборка, получаемая из интервальной выборки $\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^N$ заменой интервалов $\boldsymbol{x}_i$ их центрами: $x_i=(\underline{\boldsymbol{x}}_i+\overline{\boldsymbol{x}}_i)/2$.

Для интервальной выборки, представленной в табл. 4, будем иметь: 1) $\mu X=741.25$; 2) $\operatorname{med} X=30$; 3) $\operatorname{med}_g X=28.6$; 4) $\mathbf{med}_p\,\boldsymbol{X}=[20,40]$; 5) $\mathbf{med}_k\,\boldsymbol{X}=[20,40]$.

Для данных из примера 1 (распределение студентов по возрастам): 1) $\mu X=32.1$; 2) $\operatorname{med} X=32.1$; 3) $\operatorname{med}_g X=27.4$; 4) $\mathbf{med}_p\,\boldsymbol{X}=[25,30]$; 5) $\mathbf{med}_k\,\boldsymbol{X}=[25,30]$.

Пример 6. В табл. 5 представлены результаты измерений (в сантиметрах) размеров ножки и шляпки для видов грибов из рода Agaricies. Эти данные взяты из видового каталога грибов Калифорнии [9]. Например, ширина шляпки гриба вида arorae принадлежит интервалу от 3 до 8 см, длина его ножки принадлежит интервалу от 4 до 9 см, а толщина — интервалу от 0.5 до 2.5 см. Очевидно, что в этом случае выполнено

Т а б л и ц а 4. Распределение значений заработной платы по региону
Table 4. Distribution of salaries by region

$\tilde{\boldsymbol{x}}_i$, тыс. руб.	$f(\tilde{\boldsymbol{x}}_i)$
[0,10]	489
[10,20]	3000
[20,40]	3500
[40,60]	2000
[60,100]	700
[100,200]	300
[200,500]	10
[500,10000]	1

Т а б л и ц а 5. Параметры видов грибов из рода Agaricies, см
Table 5. Parameters of species of fungi of the genus Agaricies, cm

$\boldsymbol{X}$	Вид	Ширина шляпки	Длина ножки	Толщина ножки
$\boldsymbol{x}_1$	arorae	[3.0, 8.0]	[4.0, 9.0]	[0.50, 2.50]
$\boldsymbol{x}_2$	arvenis	[6.0, 21.0]	[4.0, 14.0]	[1.00, 3.50]
$\boldsymbol{x}_3$	benesi	[4.0, 8.0]	[5.0, 11.0]	[1.00, 2.00]
$\boldsymbol{x}_4$	bernardii	[7.0, 6.0]	[4.0, 7.0]	[3.00, 4.50]
$\boldsymbol{x}_5$	bisporus	[5.0, 12.0]	[2.0, 5.0]	[1.50, 2.50]
$\boldsymbol{x}_6$	bitorquis	[5.0, 15.0]	[4.0, 10.0]	[2.00, 4.00]
$\boldsymbol{x}_7$	califorinus	[4.0, 11.0]	[3.0, 7.0]	[0.40, 1.00]
$\boldsymbol{x}_8$	campestris	[5.0, 10.0]	[3.0, 6.0]	[1.00, 2.00]
$\boldsymbol{x}_9$	comtulus	[2.5, 4.0]	[3.0, 5.0]	[0.40, 0.70]
$\boldsymbol{x}_10$	cupreo-brunneus	[2.5, 6.0]	[1.5, 3.5]	[1.00, 1.50]
$\boldsymbol{x}_11$	diminutives	[1.5, 2.5]	[3.0, 6.0]	[0.25, 0.35]
$\boldsymbol{x}_12$	fuseo-fibrillosus	[4.0, 15.0]	[4.0, 15.0]	[1.50, 2.50]
$\boldsymbol{x}_13$	fuscovelatus	[3.5, 8.0]	[4.0, 10.0]	[1.00, 2.00]
$\boldsymbol{x}_14$	hondensis	[7.0, 14.0]	[8.0, 14.0]	[1.50, 2.50]
$\boldsymbol{x}_15$	lilaceps	[8.0, 20.0]	[9.0, 19.0]	[3.00, 5.00]
$\boldsymbol{x}_16$	micromegathus	[2.5, 4.0]	[2.5, 4.5]	[0.40, 0.70]
$\boldsymbol{x}_17$	praeclaresquamosus	[7.0, 19.0]	[8.0, 15.0]	[2.00, 3.50]
$\boldsymbol{x}_18$	pattersonae	[5.0, 15.0]	[6.0, 15.0]	[2.50, 3.50]
$\boldsymbol{x}_19$	perobscurus	[8.0, 12.0]	[6.0, 12.0]	[1.50, 2.00]
$\boldsymbol{x}_20$	semotus	[2.0, 6.0]	[3.0, 7.0]	[0.40, 0.80]
$\boldsymbol{x}_21$	silvicola	[6.0, 12.0]	[6.0, 12.0]	[1.50, 2.00]
$\boldsymbol{x}_22$	subrutilescens	[6.0, 12.0]	[6.0, 16.0]	[1.00, 2.00]
$\boldsymbol{x}_23$	xanthodermus	[5.0, 17.0]	[4.0, 14.0]	[1.00, 3.50]

условие II, поскольку эти параметры изменяются непрерывно в зависимости от условий произрастания грибов и их возраста. Значит, истинное значение этих параметров для гриба из рода Agaricies может быть любым точечным значением, допускаемым интервалами выборки.

Значения рассматриваемых статистик для соответствующих выборок следующие:

а) выборка $\boldsymbol{X}$ — значения ширины шляпки: 1) $\mu X = 8$, 2) $\operatorname{med} X = 8.75$, 3) $\mathbf{med}_p\, \boldsymbol{X} = [6, 7]$, 4) $\mathbf{med}_k\, \boldsymbol{X} = [5, 12]$;

б) выборка $\boldsymbol{X}$ — значения длины ножки: 1) $\mu X = 7.49$, 2) $\operatorname{med} X = 7.5$, 3) $\mathbf{med}_p\, \boldsymbol{X} = [5, 7]$, 4) $\mathbf{med}_k\, \boldsymbol{X} = [4, 10.5]$;

в) выборка $\boldsymbol{X}$ — значения толщины ножки: 1) $\mu X = 1.82$, 2) $\operatorname{med} X = 1.75$, 3) $\mathbf{med}_p\, \boldsymbol{X} = [1.5, 2]$, 4) $\mathbf{med}_k\, \boldsymbol{X} = [1, 2.25]$.

Заметим, что для этого примера $\mathbf{med}_p$ дает обоснованно более узкие интервальные серединные значения для всех трех выборок, т. е. более точно оценивает серединное значение для множества точечных значений, допускаемых интервальной выборкой.

Выводы

В работе производится обобщение определения медианы точечных данных на случай интервальных данных. Для определения интервальной медианы задается отношение

линейного порядка для интервальных данных и находится их частота. Это производится за счет перехода от исходных интервалов к минимальному набору их подынтервалов (элементарных подынтервалов), объединением которых они могут быть получены. Обобщение медианы точеченых данных производится с соблюдением принципа соответствия: при стремлении ширины интервалов к нулю значение медианы, рассчитанной для интервалов, стремится к значению медианы для точечных значений, к которым стремятся сужаемые интервалы.

Производится сравнение определяемой в работе медианы с интервальной медианой, определяемой как интервал, концы которого являются, соответственно, точечными медианами левых и правых концов интервалов выборки. Показано, что в том случае, если некоторая величина гарантированно может принимать все возможные точечные значения из интервалов выборки или некоторое множество допускаемых этими интервалами точечных значений, а не имеет для каждого интервала выборки только одно ее истинное значение, определяемая в работе интервальная медиана оказывается более адекватной для оценки серединного значения, чем медиана, определяемая по концам интервалов.

Определенная в работе интервальная медиана может быть использована для оценивания распределения интервальных эмпирических данных через их серединное значение, а также для их центрирования и нормализации.

Благодарности. Автор благодарит С.П. Шарого, А.Н. Баженова, С.И. Жилина, С.И. Кумкова и Е.В. Чаусову за ценные советы и замечания.

Список литературы

[1] **Ferson S., Kreinovich V., Hajagos J., Oberkampf W., Ginzburg L.** Experimental uncertainty estimation and statistics for data having interval uncertainty. Sandia Report. SAND 2007-0939. 2007: 162.

[2] **Новицкий П.В., Зограф И.А.** Оценка погрешностей результатов измерений. Л.: Энергоатомиздат. Ленинградское отделение; 1991: 304.

[3] **Орлов А.И.** Часто ли распределение результатов наблюдений является нормальным? Заводская лаборатория. 1991; 57(7):64–66.

[4] **Blum M., Floyd R.W., Pratt V., Rivest R.L., Tarjan R.E.** Time bounds for selection. Journal of Computer and System Sciences. 1973; (7):448–461.

[5] **Пролубников А.В.** Задача о покрытии множества с интервальными весами подмножеств и жадный алгоритм ее решения. Вычислительные технологии. 2015; 20(6):72–86.

[6] **Пролубников А.В.** Об одном подходе к решению задачи о покрытии с интервальными весами и его вычислительной сложности. Вычислительные технологии. 2017; 22(2):115–126.

[7] IEEE Std 1788TM-2015. IEEE standard for interval arithmetic. N.Y.: The Institute of Electrical and Electronics Engineers; 2015: 79.

[8] **Шарый С.П.** Конечномерный интервальный анализ. Адрес доступа: `http://www.nsc.ru/interval/Library/InteBooks/SharyBook.pdf`.

[9] **Billard L., Diday E.** Symbolic data analysis. John Wiley and Sons Ltd; 2006: 325.

Вычислительные технологии, 2024, том 29, № 4, с. 55–70. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 4, pp. 55–70. © FRC ICT, 2024 eISSN 2313-691X

COMPUTATIONAL TECHNOLOGIES

DOI:10.25743/ICT.2024.29.4.005

Median of interval data

A. V. Prolubnikov

Omsk State Universisty, 644077, Omsk, Russia
Corresponding author: Alexandr V. Prolubnikov, e-mail: a.v.prolubnikov@mail.ru
Received September 11, 2023, revised October 17, 2023, accepted March 14, 2024.

Abstract

Purpose. Define the median of the interval sample as a statistical characteristic of the sample elements, generalizing the definition of the median for point data.

Methodology. To determine the median of an interval sample, it is necessary to set a linear order relation for interval data and determine the frequency of occurrence in the sample for them. We solve this problem by switching from the initial sampling intervals to a minimal set of their subintervals (elementary subintervals), sampling intervals can be obtained by combining these subintervals. A generalized definition of the median for point data is carried out by us in compliance with the correspondence principle: when the interval widths tend to zero, the median value calculated for the intervals will tend to the median value of the sample of point values to which the narrowed intervals tend.

Findings. The median of the sample with interval elements is determined. This definition is a natural generalization of both the usual definition of the sample median as the middle value in a variation series composed of sample elements and the definition of the median of grouped data.

Originality/value. The interval median defined in the paper can be used to estimate the distribution of interval empirical data through their median value and for centering and normalization of interval data. If the value specified with interval uncertainty is guaranteed to take on all possible values from given sampling intervals and does not have only one true value for each sampling interval, the interval median determined in the paper turns out to be more adequate for estimating the middle value than the median determined by the ends of the sampling intervals.

Keywords: interval analysis.

Citation: Prolubnikov A.V. Median of interval data. Computational Technologies. 2024; 29(4):55–70. DOI:10.25743/ICT.2024.29.4.005. (In Russ.)

Acknowledgements. The author is grateful to S.P. Shary, A.N. Bazhenov, S.I. Zhilin, S.I. Kumkov, and E.V. Chausova for their valuable advices and comments.

References

1. **Ferson S., Kreinovich V., Hajagos J., Oberkampf W., Ginzburg L.** Experimental uncertainty estimation and statistics for data having interval uncertainty. Sandia Report. SAND 2007-0939. 2007: 162.
2. **Novitskiy P.V., Zorgraf I.A.** Otsenka pogreshnostey rezul'tatov izmereniy [Estimation of errors for measurement results]. Leningrad: Energoatomizdat. Leningradskoe Otdelenie; 1991: 304. (In Russ.)
3. **Orlov A.I.** Is it often happens that the results of observations are normally distributed? Zavodskaya Laboratoria. 1991; 57(7):64–66. (In Russ.)
4. **Blum M., Floyd R.W., Pratt V., Rivest R.L., Tarjan R.E.** Time bounds for selection. Journal of Computer and System Sciences. 1973; (7):448–461.
5. **Prolubnikov A.V.** The set cover problem with interval weights and the greedy algorithm for its solution. Computational Technologies. 2015; 20(6):72–86. (In Russ.)

6. **Prolubnikov A.V.** On an approach to set covering problem with interval weights and its computational complexity. Computational Technologies. 2017; 22(2):115–126. (In Russ.)
7. IEEE Std 1788TM-2015. IEEE standard for interval arithmetic. N.Y.: The Institute of Electrical and Electronics Engineers; 2015: 79.
8. **Shary S.P.** Konechnomernyy interval'nyy analiz [Finite-dimensional interval analysis]. Available at: `http://www.nsc.ru/interval/Library/InteBooks/SharyBook.pdf`. (In Russ.)
9. **Billard L., Diday E.** Symbolic data analysis. John Wiley and Sons Ltd; 2006: 325.