МЕДИАНА

ДЛЯ ИНТЕРВАЛЬНЫХ ДАННЫХ

Александр Вячеславович Пролубников

Омский государственный университет

a.v.prolubnikov@mail.ru

23.01.23

Определения медианы для точечных данных

Вариационный ряд: *варианты* — x_i. $x_1 \leq x_2 \leq \ldots \leq x_N$.

Для вариационого ряда $\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}$:

(1) Me — значение варианты*,
для которой половина вариант с учётом их частот лежит слева, а половина — справа.

(2) Me — значение варианты*,
для которой минимальна сумма расстояний от неё до других вариант с учётом их частот: $\sum |x_i - \text{Me}| \to \min$.

Me — одна из характеристик

- **выборки (вариационного ряда)**;
- вероятностного распределения.

Не чувствительна к выбросам.

(используется при оценивании распределения данных и нормализации (центрировании) данных)

Определения медианы интервального вар. ряда

Определение медианы Me интервального вариационного ряда требует задания:

(1) *линейного* порядка на $\mathbb{R}$,

(2) расстояния ρ между интервалами.

ПРИНЦИП СООТВЕТСТВИЯ:

- точечное значение — это *точное значение*;
- интервал — *приближённая оценка* точного значения.

— при уточнении измерений — *при узких интервалах* — значения медианы для интервальных данных рассчитываются эквивалентно рассчёту медианы для точечных данных.

Способы задания отношения порядка на $\mathbb{IR}$

Возможные способы задания отношения порядка на $\mathbb{IR}$ определены стандартом 1788 IEEE:

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) *в слабом смысле*, если $(\exists a \in \boldsymbol{a})\ (\exists b \in \boldsymbol{b})(a \leq b)$; $(\underline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}})$;

в) *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) *в $\exists\forall$-смысле*, если $(\exists a \in \boldsymbol{a})\ (\forall b \in \boldsymbol{b})(a \leq b)$; $(\underline{\boldsymbol{a}} \leq \underline{\boldsymbol{b}})$;

д) *(используем далее) в центральном смысле*, если

$$(\overline{\boldsymbol{a}} + \underline{\boldsymbol{a}})/2 \leq (\overline{\boldsymbol{b}} + \underline{\boldsymbol{b}})/2.$$

Выбор отношения на $\mathbb{IR}$, задающего **линейный** *порядок, определяет получаемое значение медианы интервальной выборки*

Медиана **интервальной** выборки

Дана выборка: $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, $\boldsymbol{x}_i \in \mathbb{IR}$.

- 1) **Как упорядочить $\boldsymbol{X}$?**
- 2) **Как задать частоту интервальных данных?**

— возможно, все интервалы $\boldsymbol{x}_i$ окажутся разными, хотя и окажутся близки друг к другу.

От исходных интервалов $\boldsymbol{x}_i$ переходим к работе с их подинтервалами — *интервалами разбиения* с порядком *в центральном смысле* на них.

Минимальное разбиение $\boldsymbol{R}$, дающее множество линейно упорядоченных интервалов:

$\boldsymbol{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N\}$ — множество упорядоченных *элементарных подинтервалов* таких, что

$$(\forall \boldsymbol{x}_i \in \boldsymbol{X})\ (\exists j_1, \ldots, j_t)\ (\boldsymbol{x}_i = \cup_{i=1}^{t} \boldsymbol{r}_{j_i}).$$

Частота элементарного подинтервала $\boldsymbol{r}_i$ — кол-во $\boldsymbol{x}_i \in \boldsymbol{X}$ т.ч. $\boldsymbol{r}_i \in \boldsymbol{x}_i$.

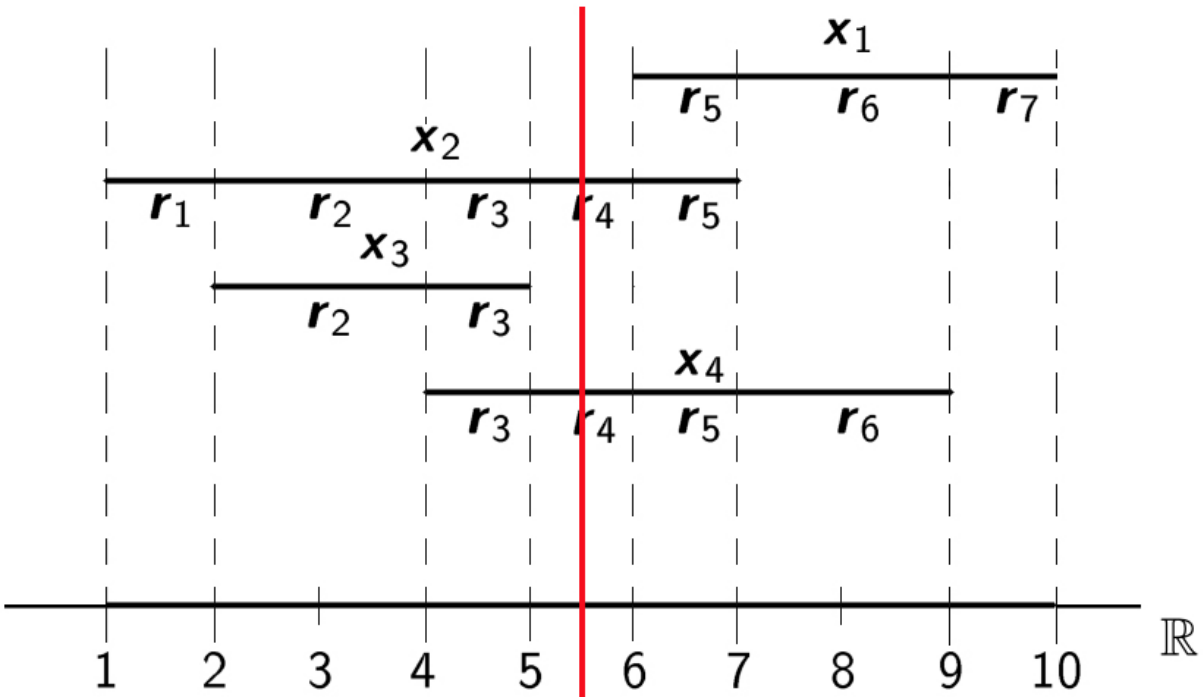
Разбиение на элементарные подинтервалы

Разбиение интервалов на подинтервалы концами интервалов:

Выборка: $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\} \Rightarrow$ упорядоченные подинтервалы:

$$\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4, \boldsymbol{r}_5, \boldsymbol{r}_6, \boldsymbol{r}_7\}.$$

$\boldsymbol{R}$ — вариационный ряд $\Rightarrow$ может быть определена медиана Me.



Хаусдорфово расстояние

Пусть

- A и B — компактные множества в $\mathbb{R}^n$;
- $\rho : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ — расстояние на $\mathbb{R}^n$.

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

$\sup\limits_{x \in A} \inf\limits_{y \in B} \rho(x, y)$ — *максимум минимального расстояния*

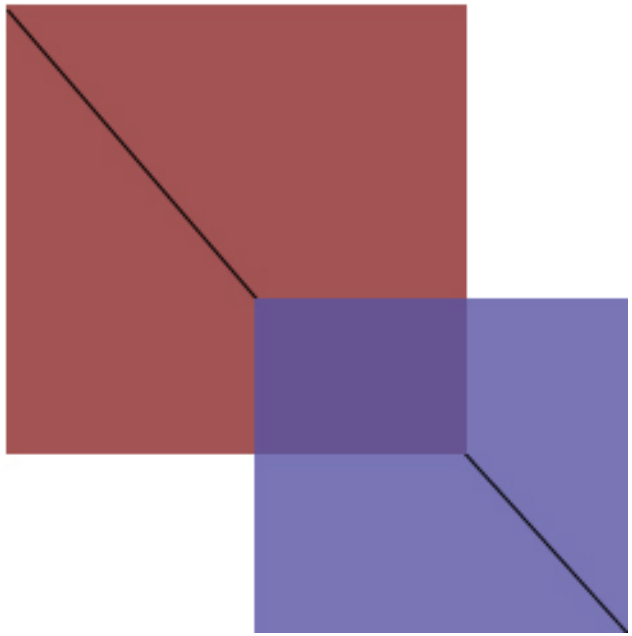
по точкам из B до точек из A.

Если $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}$, то

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

Хаусдорфово расстояние между интервалами

$\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}^2$:



Если $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}$, то

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

ТОЧЕЧНАЯ ИНТЕРВАЛЬНАЯ МЕДИАНА

ИНТЕРВАЛЬНОЙ ВЫБОРКИ

Точечная медиана для сгруппированных данных

Интервалы задают разбиение всего интервала возможных значений, по элементам разбиения (интервалам) распределяются точечные значения элементов выборки.

Пример. Распределение студентов по возрастным группам в соответствии с заданным значением некоторого признака:

Возр. гр.	Кол-во студ.	Σ накопл. частот
<20	346	346
20–25	872	1218
25–30	1054	2272
30–35	781	3053
35–40	212	3265
40–45	121	3386
≥ 45	76	3462

Точечная медиана для сгруппированных данных

Вариационный ряд: *варианты* — это значения возраста студентов, точное значение которого в каждом случае не известно, а известна только его принадлежность некоторому интервалу разбиения интервала возможных значений.

$\Rightarrow$ вводим разбиение, чтобы упорядочить *точно не заданные значения* вариант.

Частота интервала разбиения — частота попадания неточно заданных вариант в интервал.

В рассматриваемом примере:

$X = \{x_1, \ldots, x_{3462}\}$, x_i — возраст i-го студента.

Точечная медиана для сгруппированных данных

Шаг 1. Определяем **медианный интервал** — интервал, содержащий варианту (либо две), которая делит вариационный ряд на две равные части: m — номер такой варианты x_m.

Шаг 2. Вычисляем **медиану** Me:

$$\mathrm{Me} = b_0 + h\,\frac{\frac{1}{2}\sum_{i=1}^{N} f_i - S_{m-1}}{f_m},$$

b_0 — нижняя граница медианного интервала;

h — ширина этого интервала;

S_{m-1} — сумма накопленных частот для интервалов, предшествующих медианному;

f_m — частота медианного интервала — сколько неточно измеренных вариант x_i попадает в медианный интервал.

Точечная медиана для сгруппированных данных

В точечном случае:

$$\mathrm{Me} = b_0 + h\,\frac{\frac{1}{2}\sum_{i=1}^{N} f_i - S_{m-1}}{f_m}.$$

— точные значения некой величины (возраста) неизвестны, известна только их принадлежность *интервалам разбиения* всей области её значений

Me — ПРИБЛИЖЁННАЯ ОЦЕНКА МЕДИАНЫ ТОЧНЫХ ТОЧЕЧНЫХ ДАННЫХ

(которых нет — точный возраст каждого студента не известен)

— точка в медианном интервале, смещённая относительно его начала соответственно разнице суммарных частот до медианного интервала и точного значения полусуммы частот.

Точечная медиана для сгруппированных данных

Возр. гр.	Кол-во студ.	Σ накопл. частот
<20	346	346
20–25	872	1218
25–30	1054	2272
30–35	781	3053
35–40	212	3265
40–45	121	3386
≥ 45	76	3462

Интервалы —

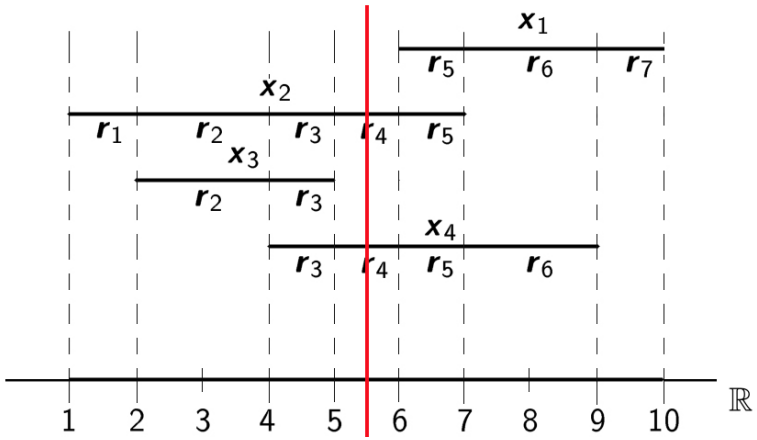
$$[0, 20], [20, 25], [25, 30], [30, 35], [35, 40], [40, 45], [45, 60]$$

1. *Медианный интервал* — $\sum f_i = 1731 \Rightarrow$ это $[25, 30]$.
2. $$\text{Me} = b_0 + h\frac{\sum f_i/2 - S_{m-1}}{f_m} = 25 + 5\frac{\frac{3462}{2} - 1218}{1054} = 27.4$$

Точечная медиана Me **интервальной** выборки

Интервал разбиения — элементарный подинтервал —
— *это оценка значения, заданного с неопределённостью.*

Частота f_i элементарного подинтервала $\boldsymbol{r}_i$ —
— это частота попадания $\boldsymbol{r}_i$ в $x_i \in \boldsymbol{X}$.

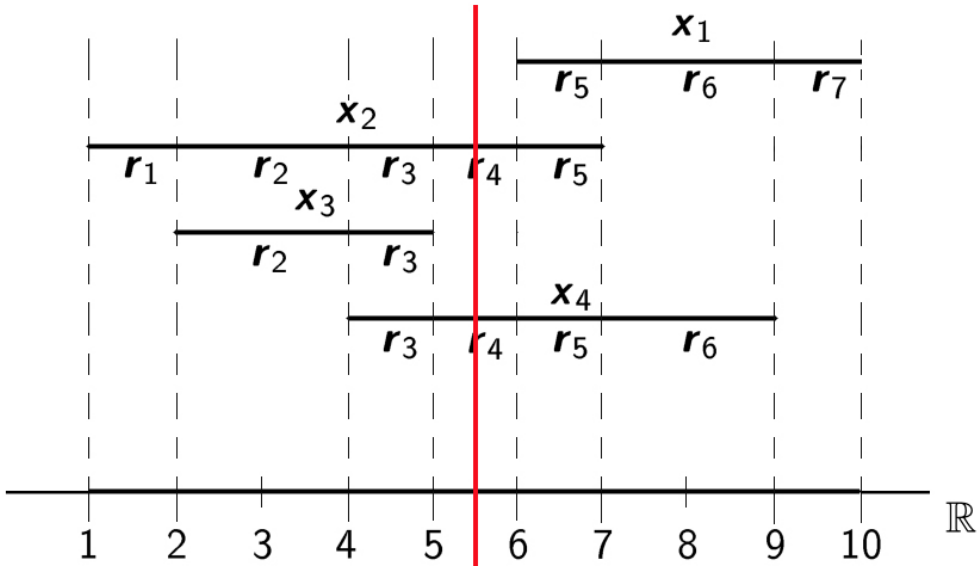


Частоты интервалов разбиения $\boldsymbol{r}_i$:

$\boldsymbol{r}_1$	$\boldsymbol{r}_2$	$\boldsymbol{r}_3$	$\boldsymbol{r}_4$	$\boldsymbol{r}_5$	$\boldsymbol{r}_6$	$\boldsymbol{r}_7$
1	2	3	2	3	2	1

Точечная медиана Me **интервальной** выборки

Разбиение:



Частоты интервалов разбиения r_i:

r_1	r_2	r_3	r_4	r_5	r_6	r_7
1	2	3	2	3	2	1

$\Rightarrow$ Me принадлежит интервалу r_4,

$$\text{Me} = 5 + \frac{\frac{14}{2} - 6}{2} = 5.5$$

Точечная медиана Me **интервальной** выборки

Me:

$$\mathrm{Me} = b_0 + h\,\frac{\frac{1}{2}\sum_{i=1}^{N} f_i - S_{m-1}}{f_m},$$

b_0 — нижняя граница медианного элементарного подинтервала $\boldsymbol{r}_m$;

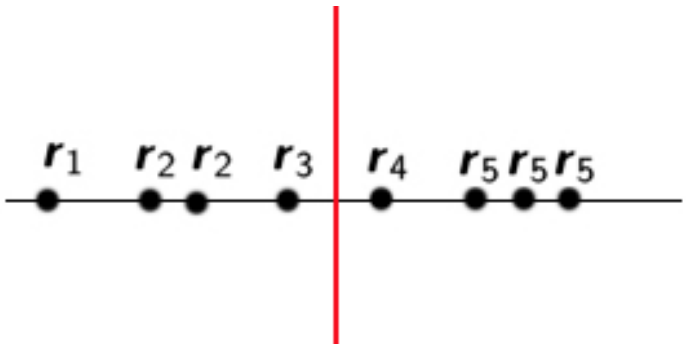
h — ширина $\boldsymbol{r}_m$;

S_{m-1} — сумма частот элементарных подинтервалов, предшествующих $\boldsymbol{r}_m$;

f_m — частота $\boldsymbol{r}_m$ — для скольких неточно измеренных значений $\boldsymbol{x}_i$ в них имеются значения, которые попадают в $\boldsymbol{r}_m$.

Точечная медиана Me **интервальной** выборки

В случае, если **нет** элементарного подинтервала, который с учётом частот интервалов лежал бы точно посередине вариационного ряда, то есть середина в упорядоченной последовательности значений с повторениями лежит между r_{i-1} и r_i:



Частоты интервалов разбиения r_i:

r_1	r_2	r_3	r_4	r_5
1	2	1	1	3

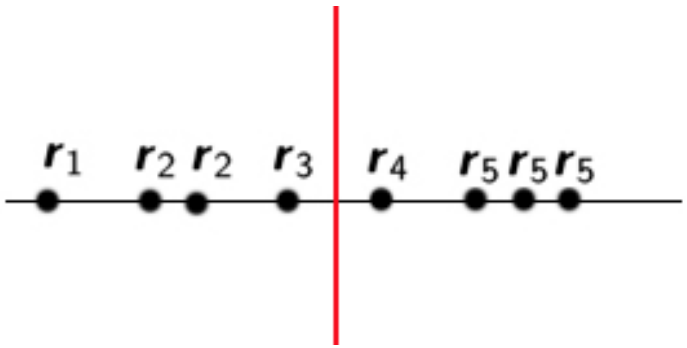
Точечная медиана Me **интервальной** выборки

В случае, если **нет** элементарного подинтервала, который с учётом частот интервалов, лежал бы точно посередине вариационного ряда, в формуле

$$\mathrm{Me} = b_0 + h\,\frac{\frac{1}{2}\sum_{i=1}^{N} f_i - S_{m-1}}{f_m}$$

используем значения

1) $b_0 = \underline{\boldsymbol{r}}_{i-1}$;
2) $h = \overline{\boldsymbol{r}}_i - \underline{\boldsymbol{r}}_{i-1}$;
3) $f_m = f_{i-1} + f_i$.



ИНТЕРВАЛЬНАЯ МЕДИАНА

ИНТЕРВАЛЬНОЙ ВЫБОРКИ

Выборка точечных значений: $X = \{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}$.

(1) Me — значение варианты*, для которой половина членов ряда с учётом их частот лежит слева, а половина — справа.

(2) Me — значение варианты*, для которой минимальна сумма расстояний от него до других вариант с учётом их частот.

Медиана **интервальной** выборки $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, $\boldsymbol{x}_i \in \mathbb{IR}$:

(1) $\mathbf{Me}_f = \boldsymbol{r}_m$ — накопленная сумма частот слева от $\boldsymbol{r}_m$ равна накопленной сумме частот справа от $\boldsymbol{r}_m$:

$$\sum_{i=1}^{m-1} f_i = \sum_{i=m+1}^{N} f_i;$$

(2) $\mathbf{Me}_\rho = \boldsymbol{r}_m$ — элементарный подинтервал $\boldsymbol{r}_m$, для которого минимальна сумма *хаусдорфовых* расстояний от него до других элеметарных интервалов с учётом их частот.

Интервальная выборка, содержащая точечные данные

Интервальная выборка $\boldsymbol{X}$ содержит точечные данные —

— в $\boldsymbol{X}$ есть вырожденные интервалы: $\boldsymbol{x}_i = [v, v]$.

$\boldsymbol{r}_j = [v, v]$ — отдельный элементарный подинтервал.

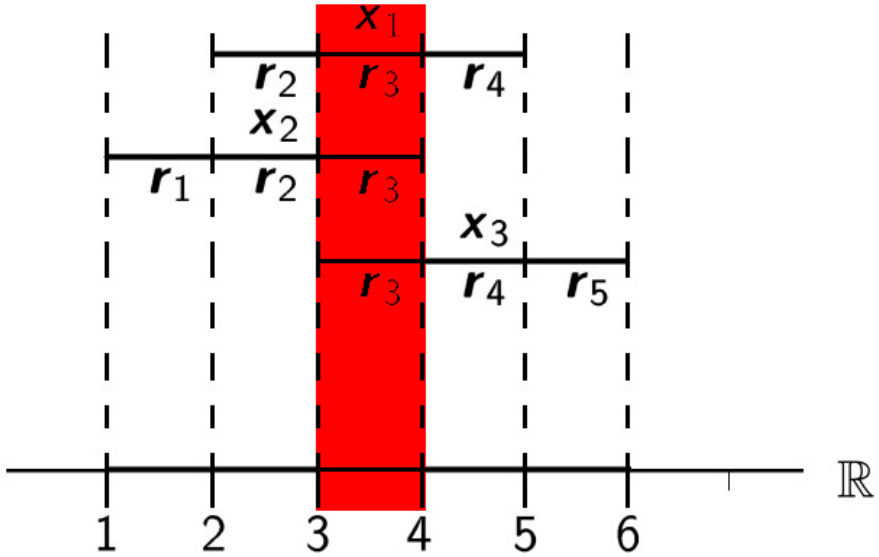
Если есть $\boldsymbol{r}_k = [v, w]$, то частоты $\boldsymbol{r}_j$ и $\boldsymbol{r}_k$ считаются отдельно.

- $\mathbf{Me}_\rho$: расстояние: $\rho(\boldsymbol{r}_j, \boldsymbol{r}_k) = |\overline{\boldsymbol{r}}_k - \underline{\boldsymbol{r}}_j| = |w - v|$.

ПРИМЕРЫ ВЫЧИСЛЕНИЯ

ИНТЕРВАЛЬНОЙ МЕДИАНЫ

Вычисление $\mathbf{Me}_f$



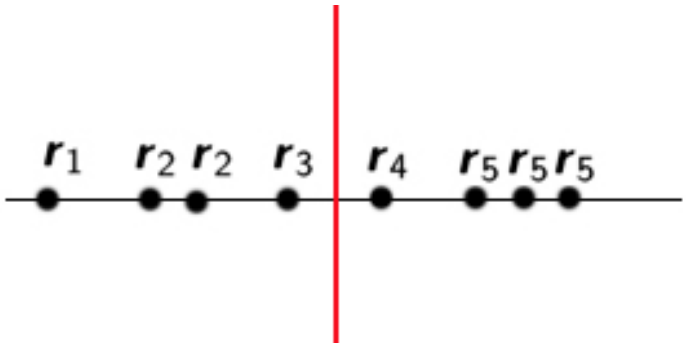
Частоты интервалов разбиения r_i:

r_1	r_2	r_3	r_4	r_5
1	2	3	2	1

$$\mathbf{Me}_f = r_3 = [3, 4]$$

Вычисление $\mathbf{Me}_f$

В случае, если **нет** элементарного подинтервала, который с учётом частот подинтервалов, лежал бы точно посередине вариационного ряда:

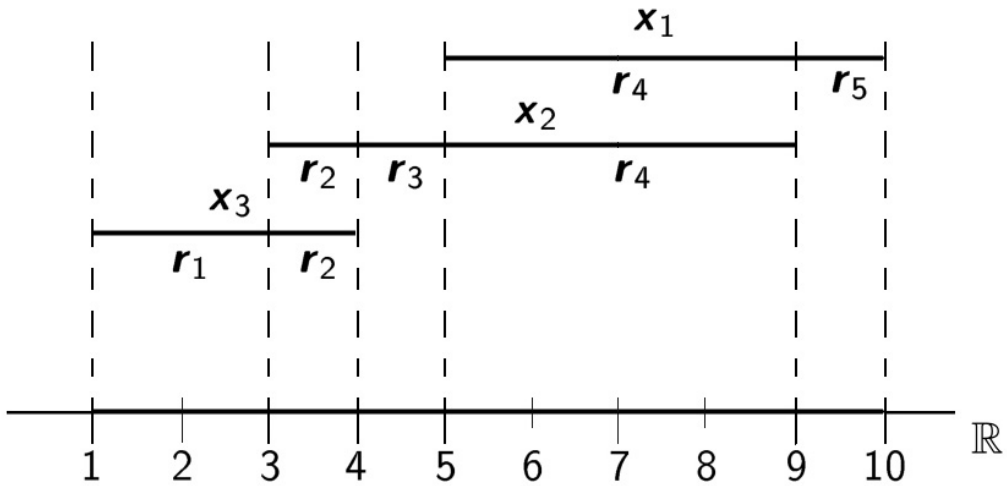


Частоты интервалов разбиения r_i:

r_1	r_2	r_3	r_4	r_5
1	2	1	1	3

$$\mathbf{Me}_f = r_3 \cup r_4$$

Вычисление $\mathbf{Me}_\rho$

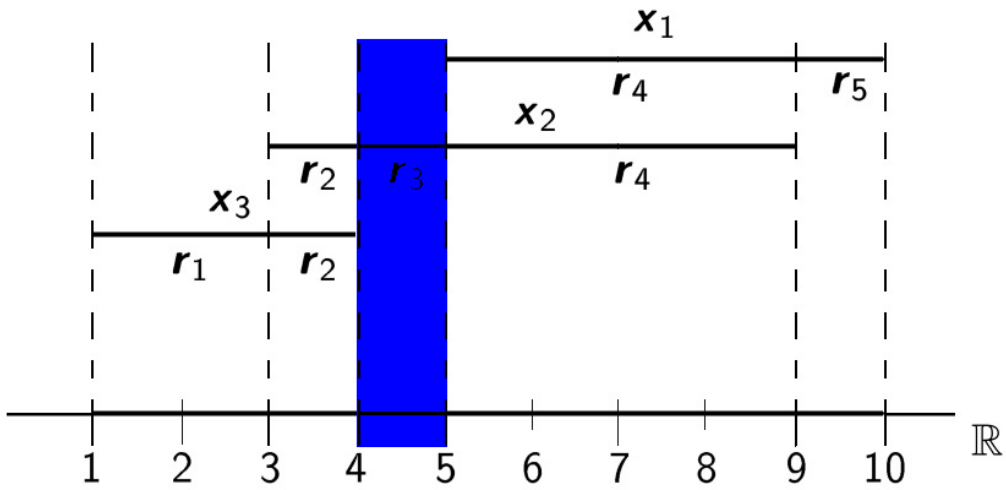


Частоты интервалов разбиения r_i:

r_1	r_2	r_3	r_4	r_5
1	2	1	2	1

$$\mathbf{Me}_\rho = r_3 = [4, 5]$$

Вычисление $\mathbf{Me}_\rho$



$$\sum_{k\neq 1}(f_k\cdot\rho(\boldsymbol{r}_1,\boldsymbol{x}_k))=27,\ \sum_{k\neq 2}(f_k\cdot\rho(\boldsymbol{r}_2,\boldsymbol{x}_k))=19,\ \sum_{k\neq 3}(f_k\cdot\rho(\boldsymbol{r}_3,\boldsymbol{x}_k))=18,$$

$$\sum_{k\neq 4}(f_k\cdot\rho(\boldsymbol{r}_4,\boldsymbol{x}_k))=24,\ \sum_{k\neq 5}(f_k\cdot\rho(\boldsymbol{r}_5,\boldsymbol{x}_k))=33.$$

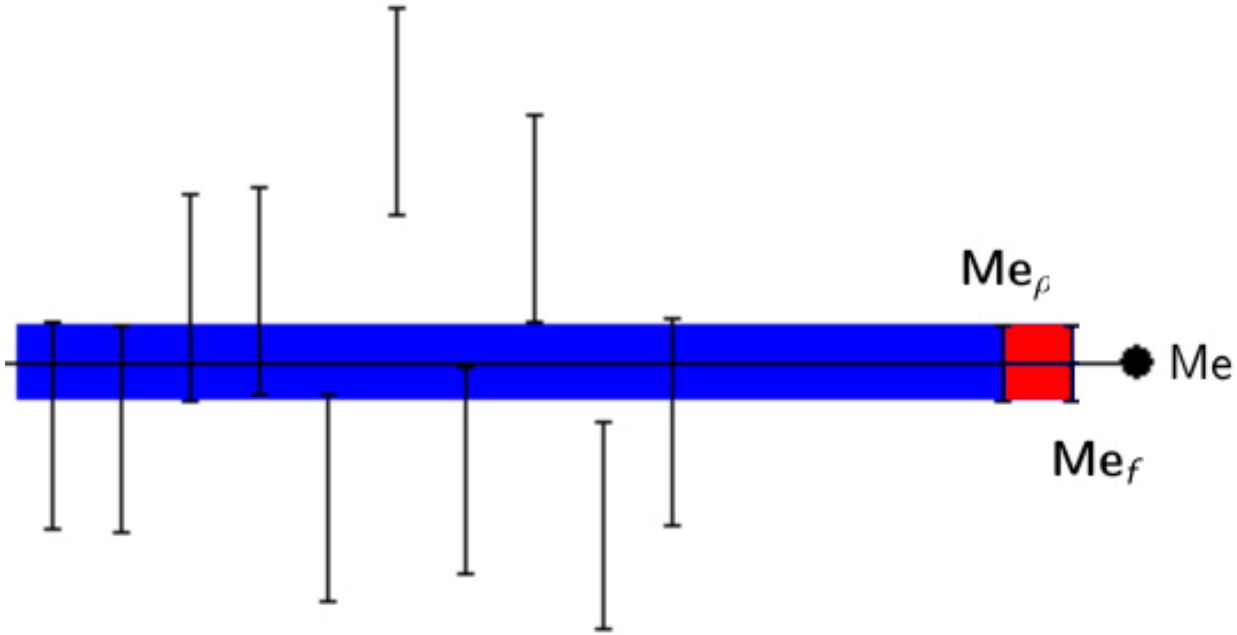
$$\mathbf{Me}_\rho=\boldsymbol{r}_3=[4,5]$$

КАК СООТНОСЯТСЯ Me, Me_f И Me_ρ?

Me, $\mathbf{Me}_f$, $\mathbf{Me}_\rho$

Генерирование интервалов: интервалы исходной выборки $\boldsymbol{X}_0 = \{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i = [-1, 1]$, случайно смещаются вверх и вниз.

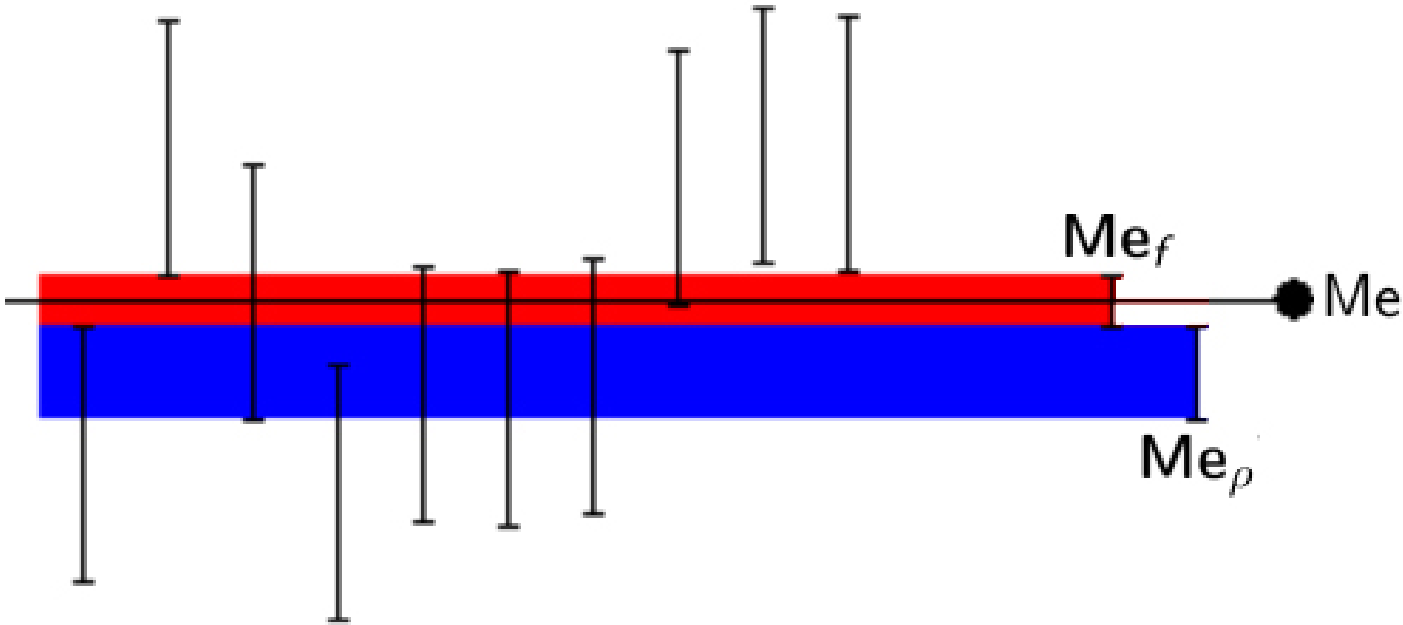
Наиболее типичная ситуация:



$\mathbf{Me}_f$ и $\mathbf{Me}_\rho$ — совпадающие элементарные подинтервалы

Me, $\mathbf{Me}_f$, $\mathbf{Me}_\rho$

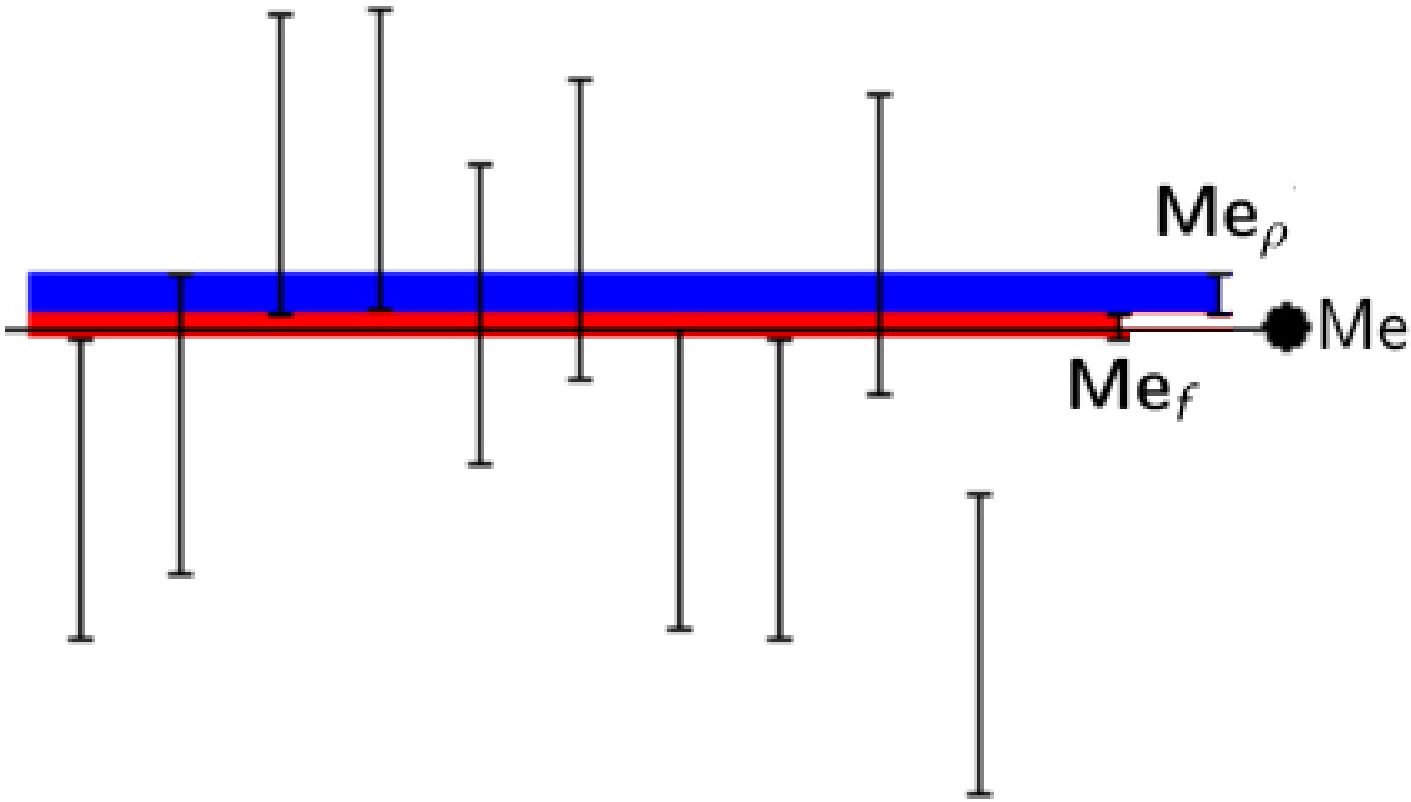
Возможно:



$\mathbf{Me}_f$ и $\mathbf{Me}_\rho$ — соседние элементарные подинтервалы

Me, $\mathbf{Me}_f$, $\mathbf{Me}_\rho$

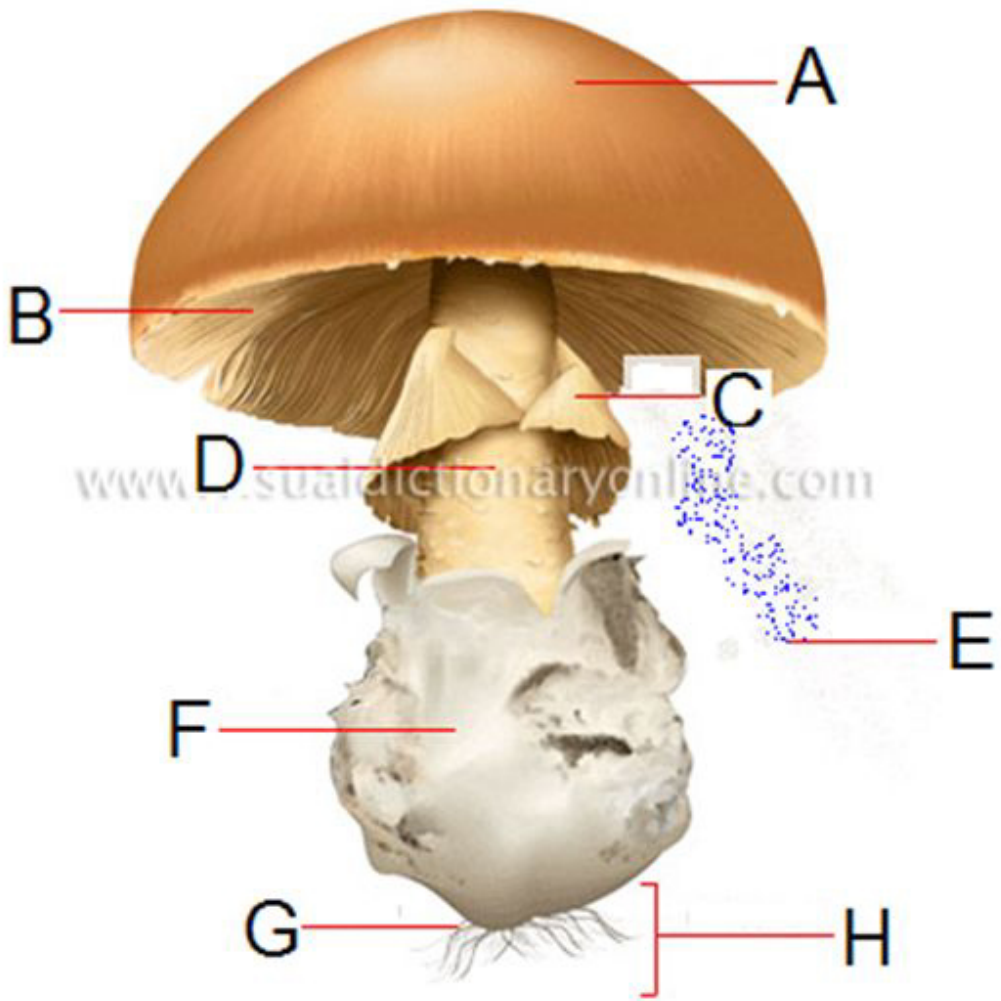
Возможно:



$\mathbf{Me}_f$ и $\mathbf{Me}_\rho$ — соседние элементарные подинтервалы

Выборка параметров грибов

Mushroom structure



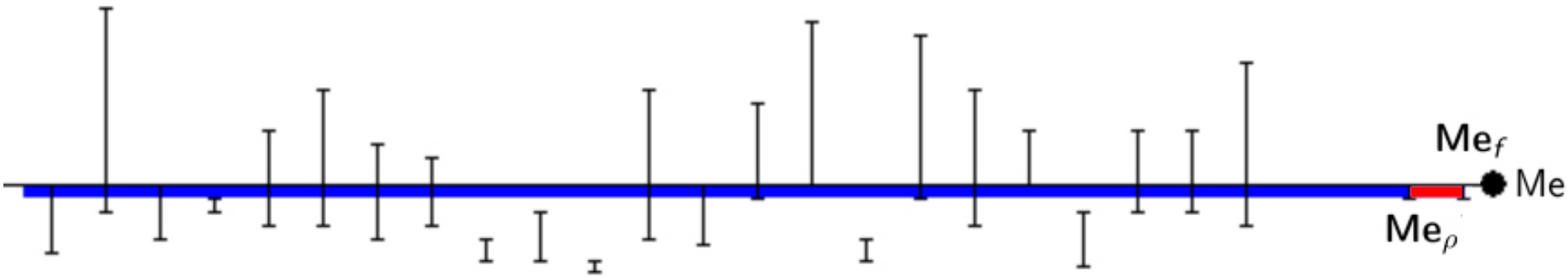
A: Cap(pileus)
B: Gill
C: Ring
D: Stalk,Stipe

E: Spores
F: Volva
G: Hypa(e)
H: Mycelium

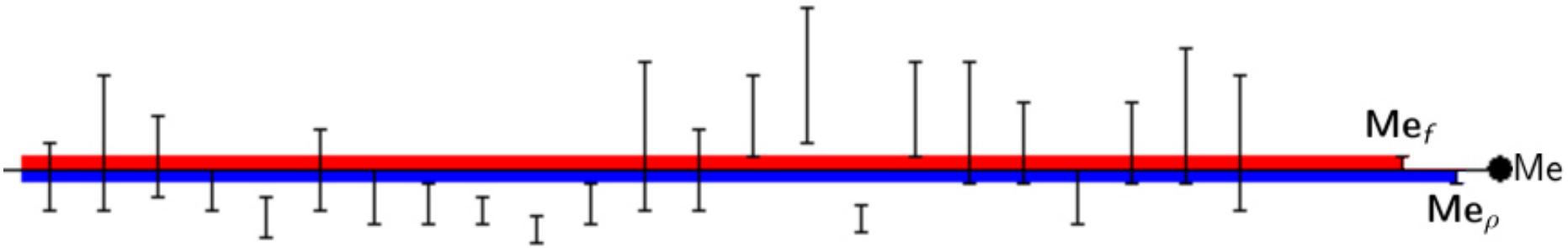
Выборка параметров грибов

n	Species	Pileus Cap Width	Stipe Length	Stipe Thickness
1	arorae	[3.0, 8.0]	[4.0, 9.0]	[0.50, 2.50]
2	arvenis	[6.0, 21.0]	[4.0, 14.0]	[1.00, 3.50]
3	benesi	[4.0, 8.0]	[5.0, 11.0]	[1.00, 2.00]
4	bernardii	[7.0, 6.0]	[4.0, 7.0]	[3.00, 4.50]
5	bisporus	[5.0, 12.0]	[2.0, 5.0]	[1.50, 2.50]
6	bitorquis	[5.0, 15.0]	[4.0, 10.0]	[2.00, 4.00]
7	californinus	[4.0, 11.0]	[3.0, 7.0]	[0.40, 1.00]
⋮	⋮	⋮	⋮	⋮
21	silvicola	[6.0, 12.0]	[6.0, 12.0]	[1.50, 2.00]
22	subrutilescens	[6.0, 12.0]	[6.0, 16.0]	[1.00, 2.00]
23	xanthodermus	[5.0, 17.0]	[4.0, 14.0]	[1.00, 3.50]

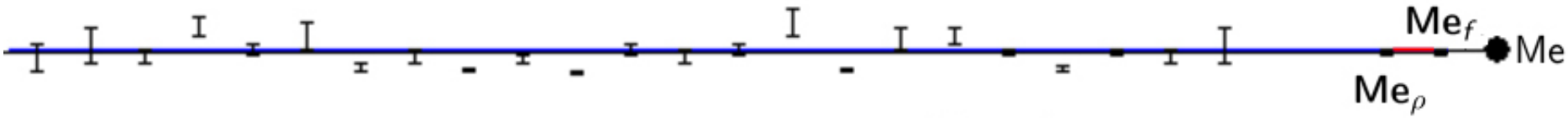
Выборка параметров грибов. Размер шляпы.



Выборка параметров грибов. Длина ножки.



Выборка параметров грибов. Ширина ножки.



Обобщение на случай многомерных интервалов

$\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, $\boldsymbol{x}_i \in \mathbb{IR}^n$.

ВАРИАНТ 1.

$$\mathbf{Me} = (\mathbf{Me}_1, \ldots, \mathbf{Me}_n) \in \mathbb{IR}^n,$$

где $\mathbf{Me}_i$ — медианы компонент $\boldsymbol{x}_i$.

ВАРИАНТ 2.

1). Находим разбиение всех интервалов выборки $\boldsymbol{X}$ концами интервалов $\boldsymbol{x}_i$ на элементарные подбрусы.

2). $\mathbf{Me} \in \mathbb{IR}^n$ — брус, выбираемый из элементарных подбрусов, такой, что расстояние от него до всех остальных элементарных подбрусов минимально.