Поговорим об описательной статистике перформанс-распределений

Андрей Акиньшин

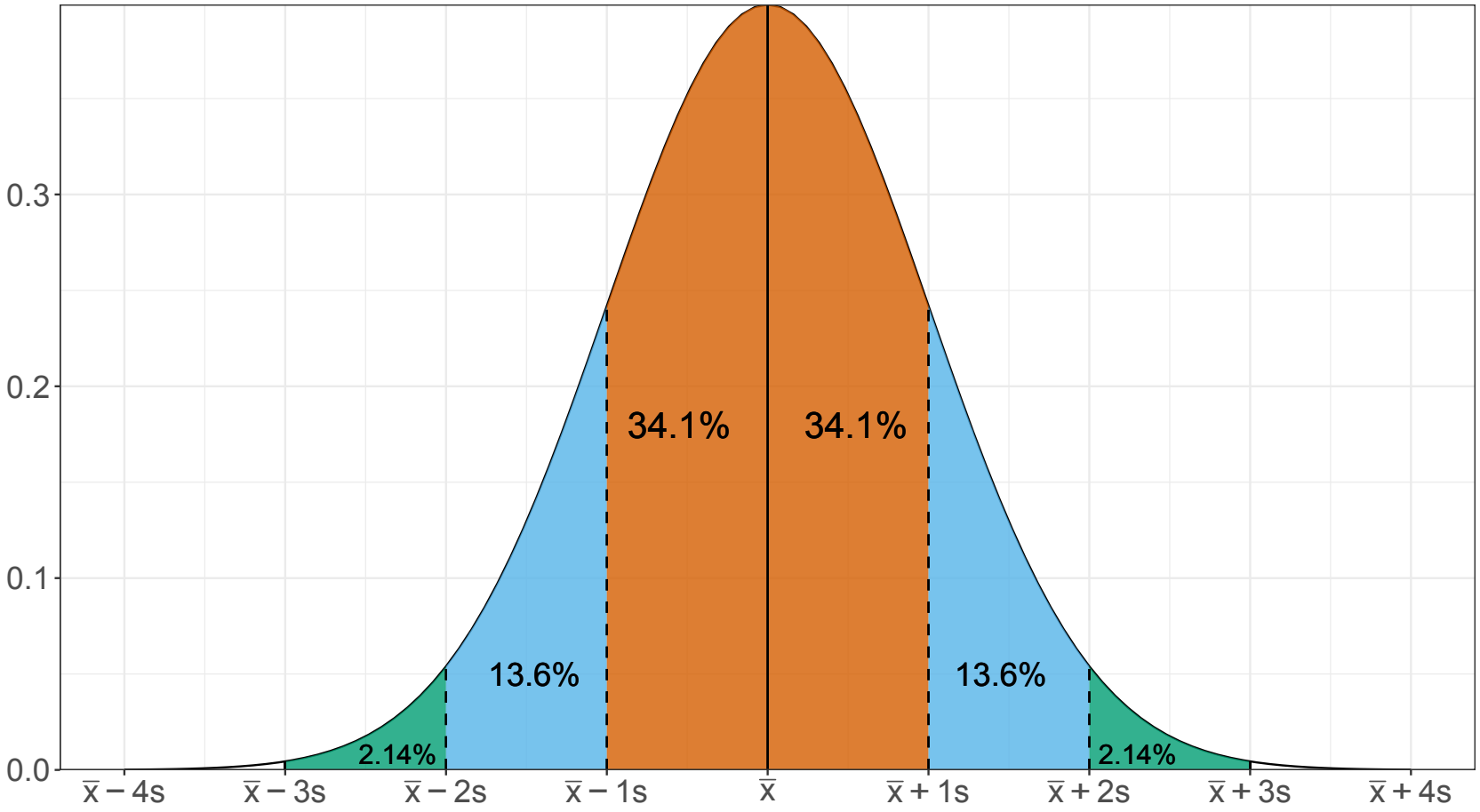
Heisenbug 2022 Autumn

Андрей Акиньшин — Анализируем перформанс с пользой для себя и окружающих

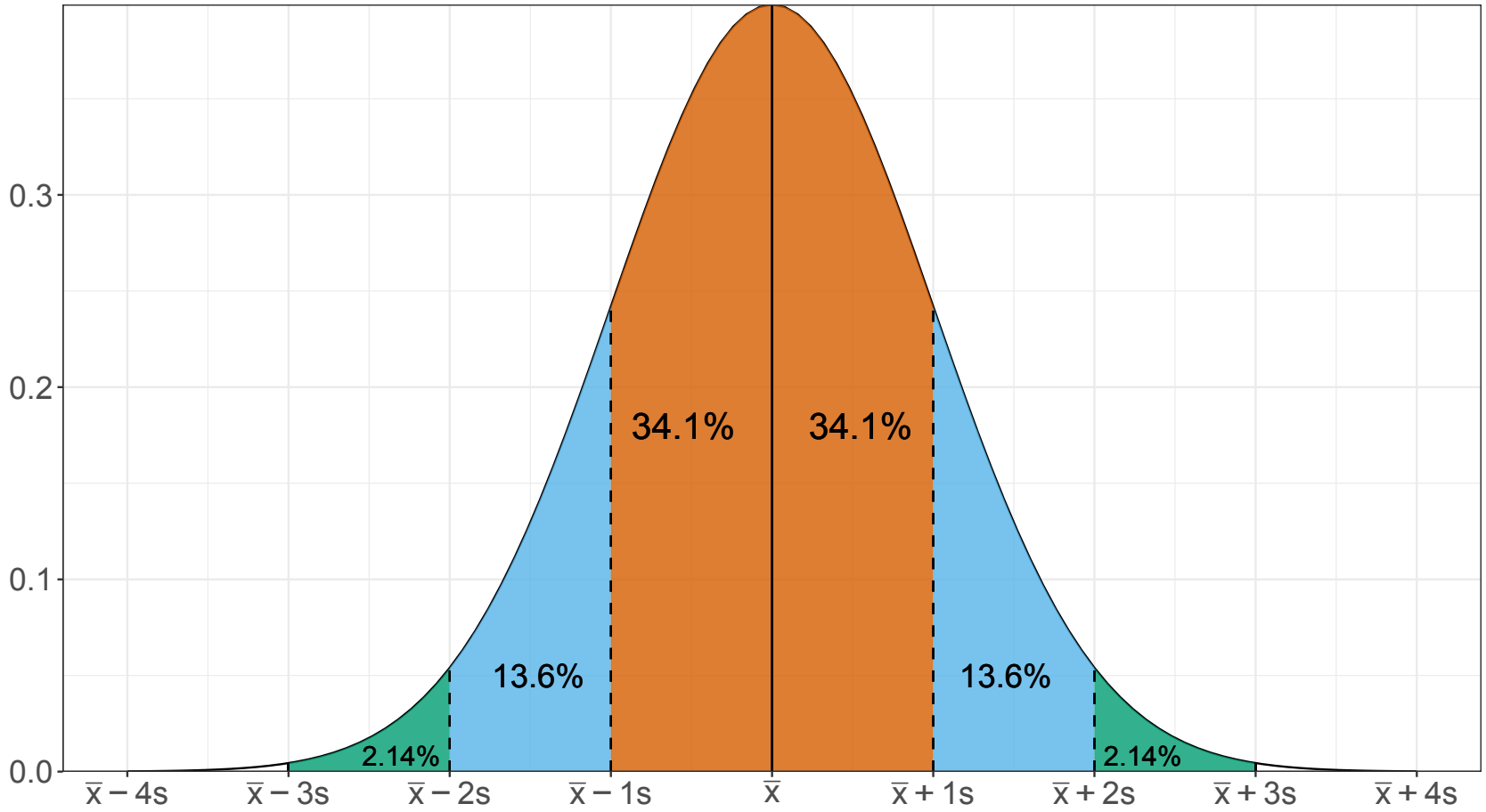
https://www.youtube.com/watch?v=jZ0quqA1Fn8

В прошлой серии...

Нормальное распределение

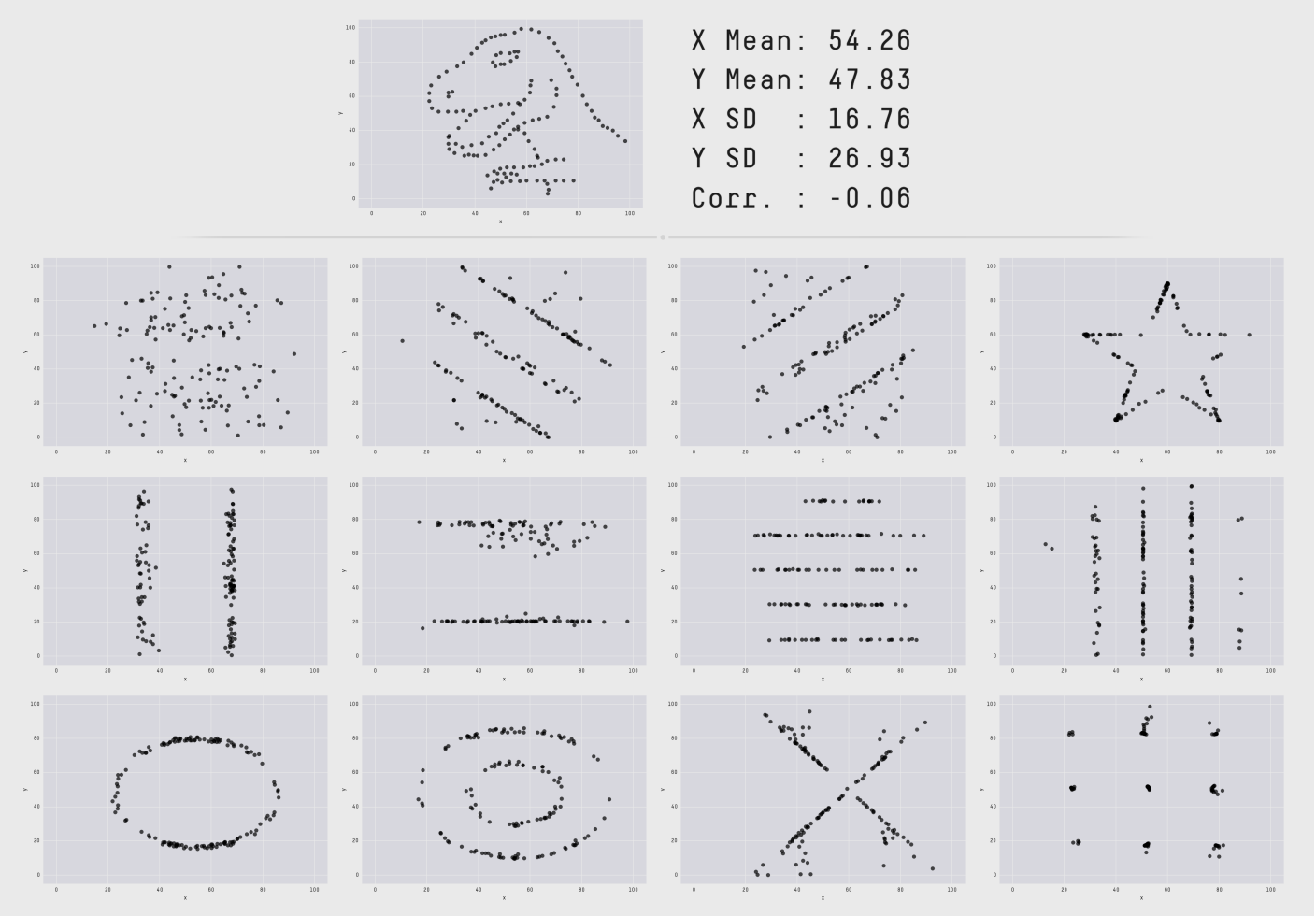


Нормальное распределение



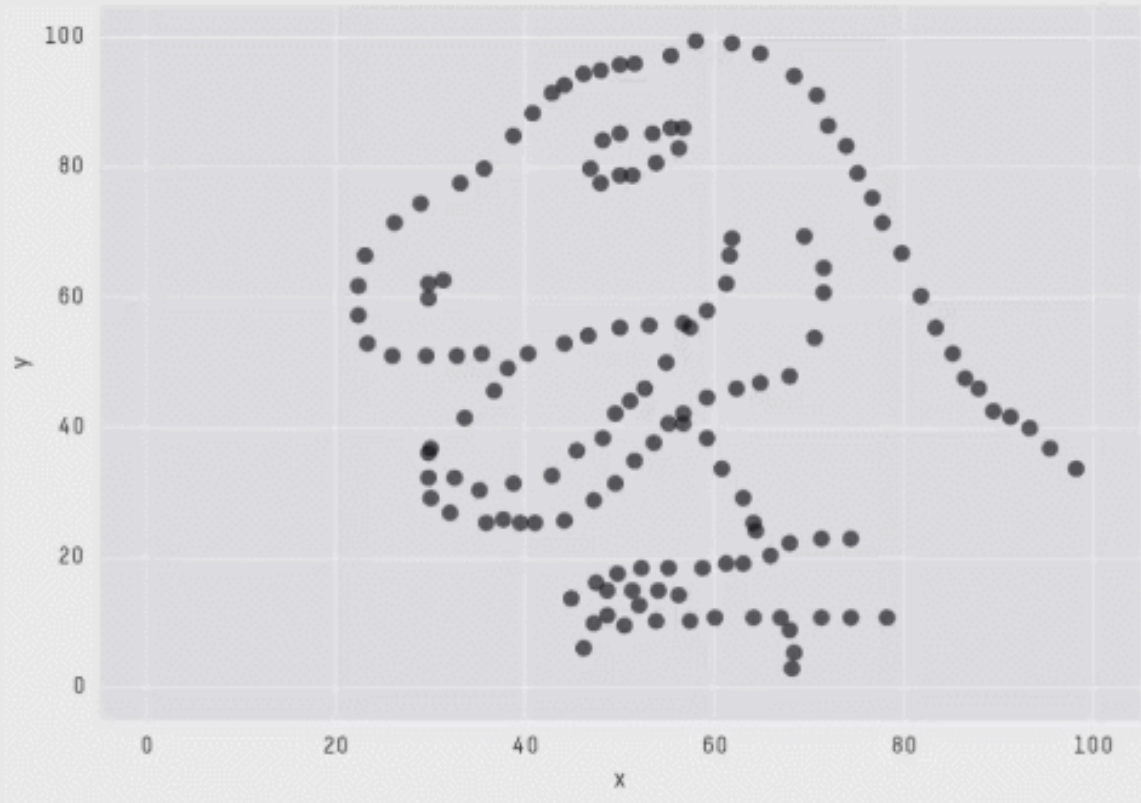
> *Normality is a myth; there never was, and never will be, a normal distribution.*
> "Testing for normality", R.C. Geary, 1947

Обманчивые среднее и дисперсия



Justin Matejka, George Fitzmaurice (2017), "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing", CHI 2017 Conference proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems

Обманчивые среднее и дисперсия



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Justin Matejka, George Fitzmaurice (2017), "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing", CHI 2017 Conference proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems

```
int N = 1000; // Количество итераций
var measurements = new long[N];
byte[] data = new byte[64 * 1024 * 1024]; // 64MB
```

```
int N = 1000; // Количество итераций
var measurements = new long[N];
byte[] data = new byte[64 * 1024 * 1024]; // 64MB

for (int i = 0; i < N; i++)
{
  var stopwatch = Stopwatch.StartNew();
  var fileName = Path.GetTempFileName();
  File.WriteAllBytes(fileName, data);
  File.Delete(fileName);
  stopwatch.Stop();
  measurements[i] = stopwatch.ElapsedMilliseconds;
}
```

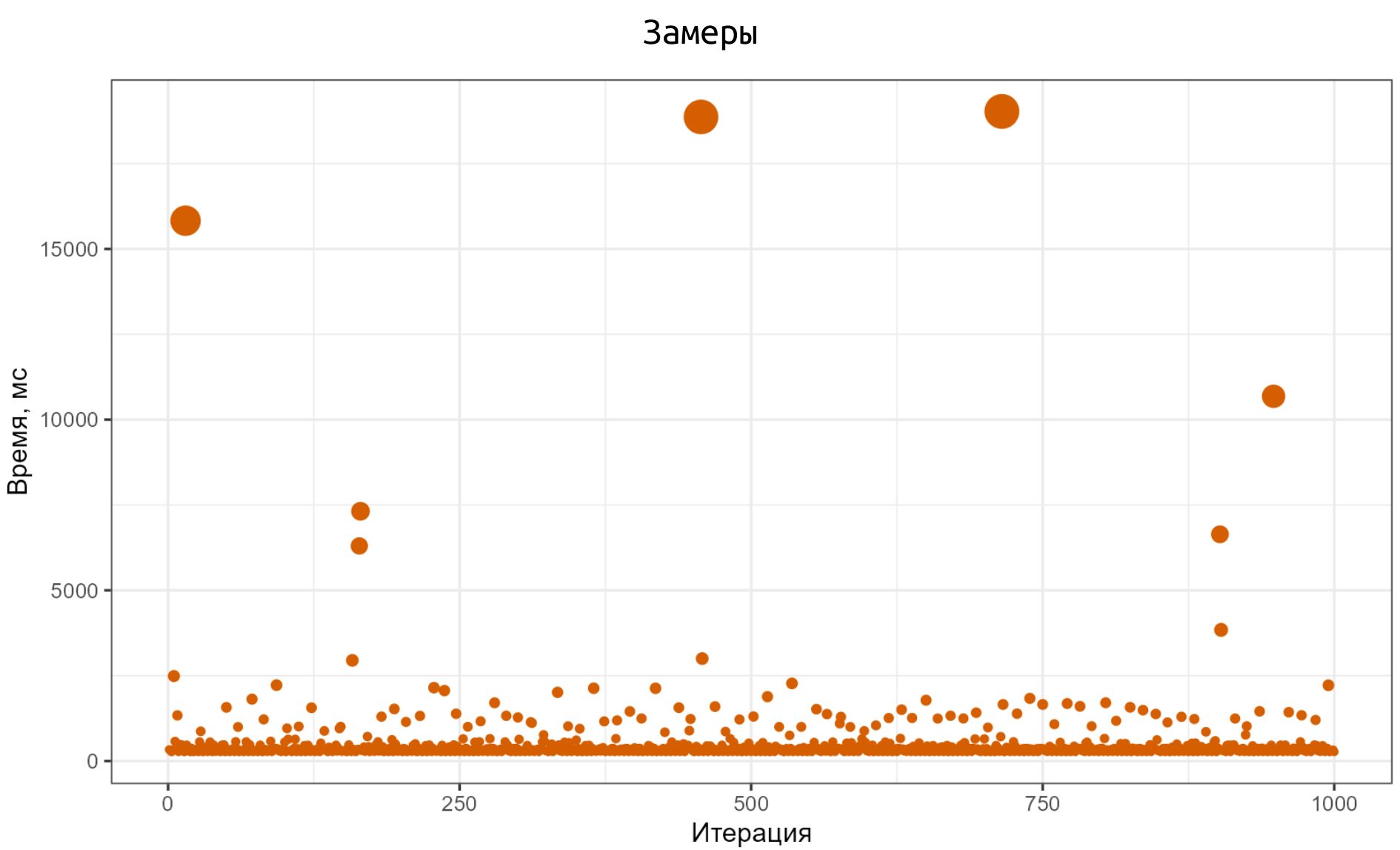
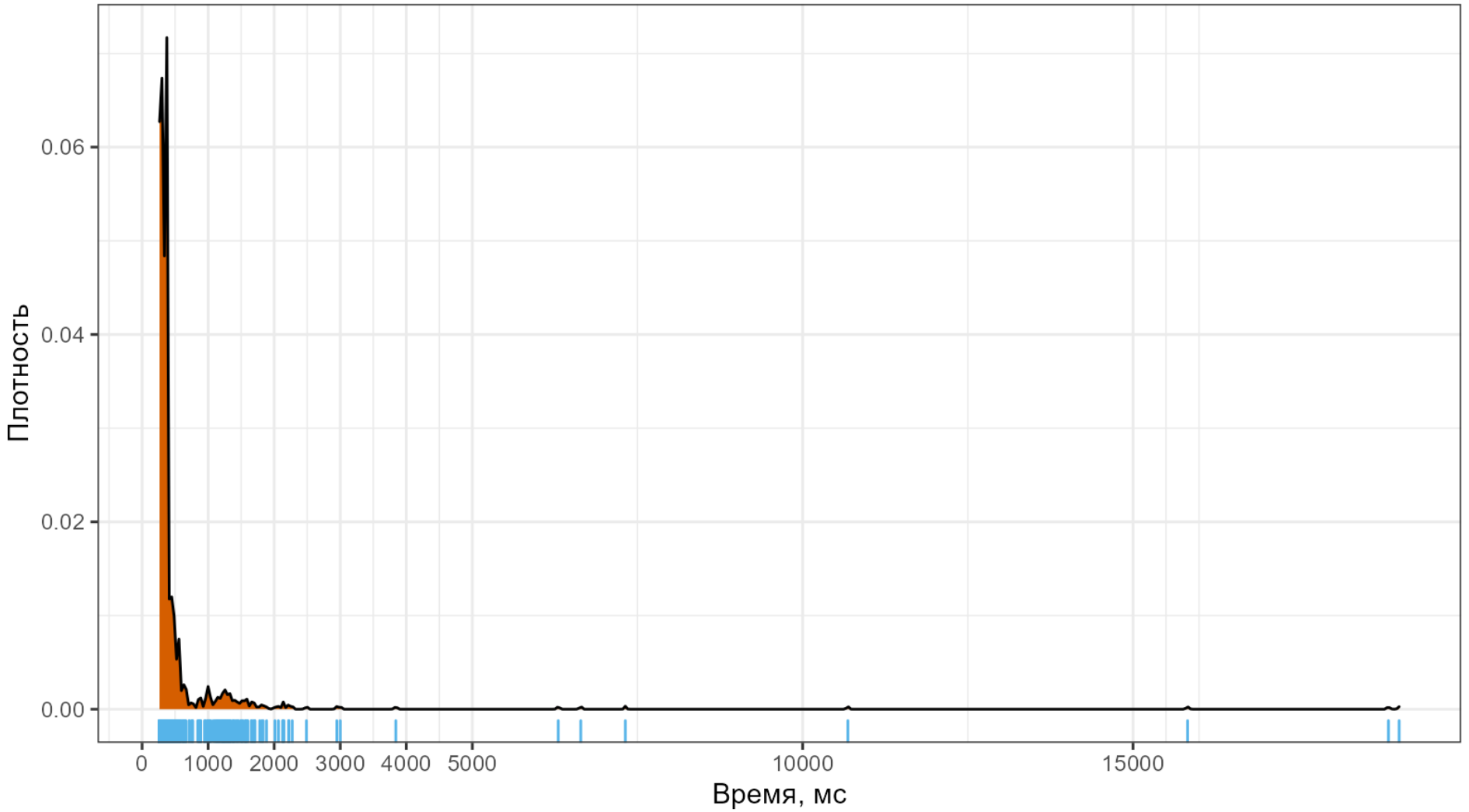

Замеры
Время, мс
15000
10000
5000
0
0
250
500
750
1000
Итерация

График плотности распределения



- Большой разброс значений

- Большой разброс значений
- Тяжелые хвосты и экстремальные выбросы

- Большой разброс значений
- Тяжелые хвосты и экстремальные выбросы
- Мультимодальность

- Большой разброс значений
- Тяжелые хвосты и экстремальные выбросы
- Мультимодальность
- Дискретизация

- Большой разброс значений
- Тяжелые хвосты и экстремальные выбросы
- Мультимодальность
- Дискретизация
- Асимметрия

Описательная статистика

Описательная статистика

- Центральная тенденция

Описательная статистика

- Центральная тенденция
- Квантильные оценки

Описательная статистика

- Центральная тенденция
- Квантильные оценки
- Вариация

Описательная статистика

- Центральная тенденция
- Квантильные оценки
- Вариация
- Плотность распределения

Описательная статистика

- Центральная тенденция
- Квантильные оценки
- Вариация
- Плотность распределения
- Мультимодальность

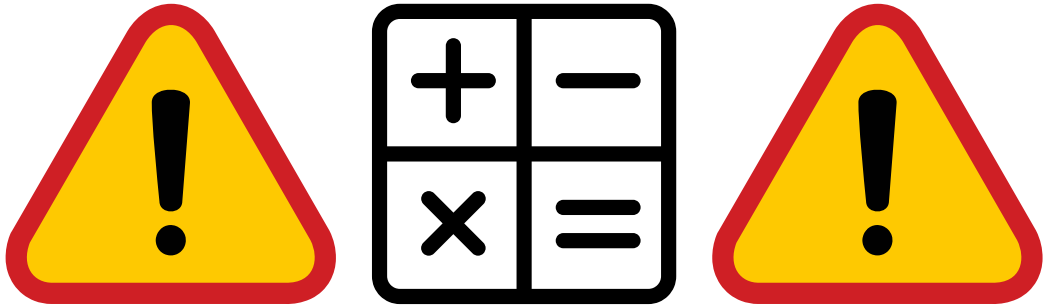
Описательная статистика

- Центральная тенденция
- Квантильные оценки
- Вариация
- Плотность распределения
- Мультимодальность
- Теория экстремальных значений

Описательная статистика

- Центральная тенденция
- Квантильные оценки
- Вариация
- Плотность распределения
- Мультимодальность
- Теория экстремальных значений

Математический доклад!



Зачем?

stars 8209 used by 12356 downloads 14359971

BenchmarkDotNet

Powerful .NET library for benchmarking

stars 8209 used by 12356 downloads 14359971

BenchmarkDotNet

Powerful .NET library for benchmarking

Perfolizer

Performance analysis toolkit

Центральная тенденция

Среднее арифметическое

$$x = \{x_1, x_2, \ldots, x_n\}; \quad \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$x = \{x_1, x_2, \ldots, x_n\}; \quad \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$x = \{1, 2, 3, 4, 5, 6, 7\}; \quad \overline{x} = 4$$

$$x = \{x_1, x_2, \ldots, x_n\}; \quad \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$x = \{1, 2, 3, 4, 5, 6, 7\}; \quad \overline{x} = 4$$

$$x = \{1, 2, 3, 4, 5, 6, 273\}; \quad \overline{x} = 42$$

$$x = \{x_1, x_2, \ldots, x_n\}; \quad \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$x = \{1, 2, 3, 4, 5, 6, 7\}; \quad \overline{x} = 4$$

$$x = \{1, 2, 3, 4, 5, 6, 273\}; \quad \overline{x} = 42$$

Среднее арифметическое не является робастным!

$$x = \{x_1, x_2, \ldots, x_n\}; \quad \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$x = \{1, 2, 3, 4, 5, 6, 7\}; \quad \overline{x} = 4$$

$$x = \{1, 2, 3, 4, 5, 6, 273\}; \quad \overline{x} = 42$$

Среднее арифметическое не является робастным!

Может попробуем медиану?

$$x = \{x_1, x_2, \ldots, x_n\}; \quad \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

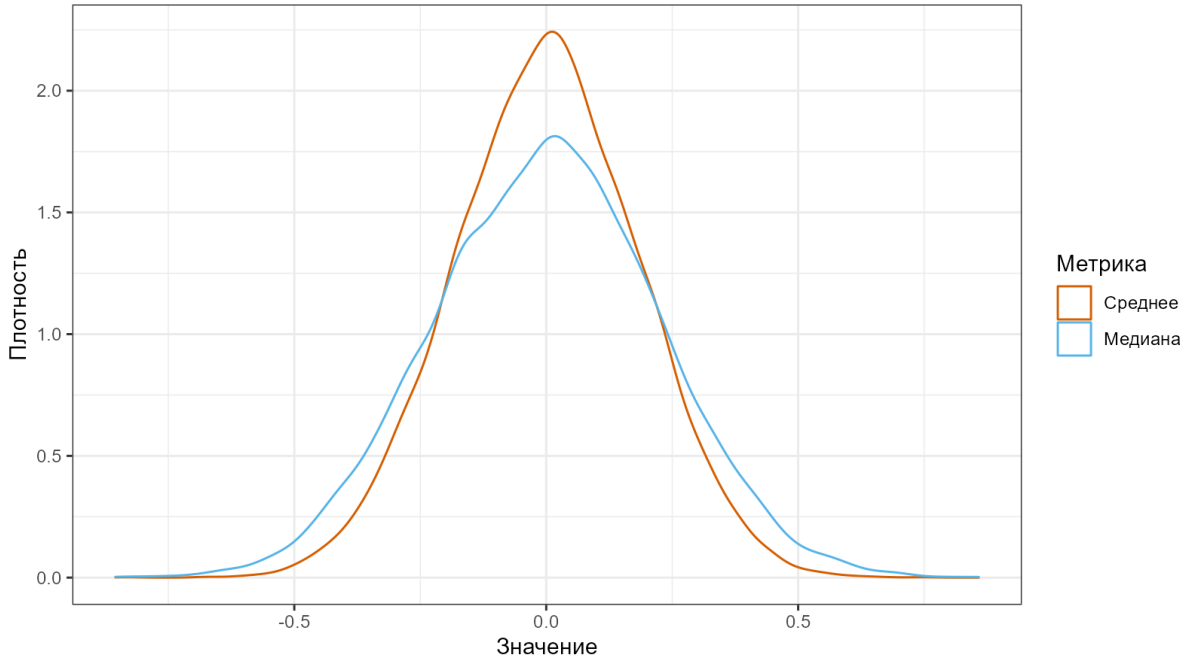
$$x = \{1, 2, 3, 4, 5, 6, 7\}; \quad \overline{x} = 4$$

$$x = \{1, 2, 3, 4, 5, 6, 273\}; \quad \overline{x} = 42$$

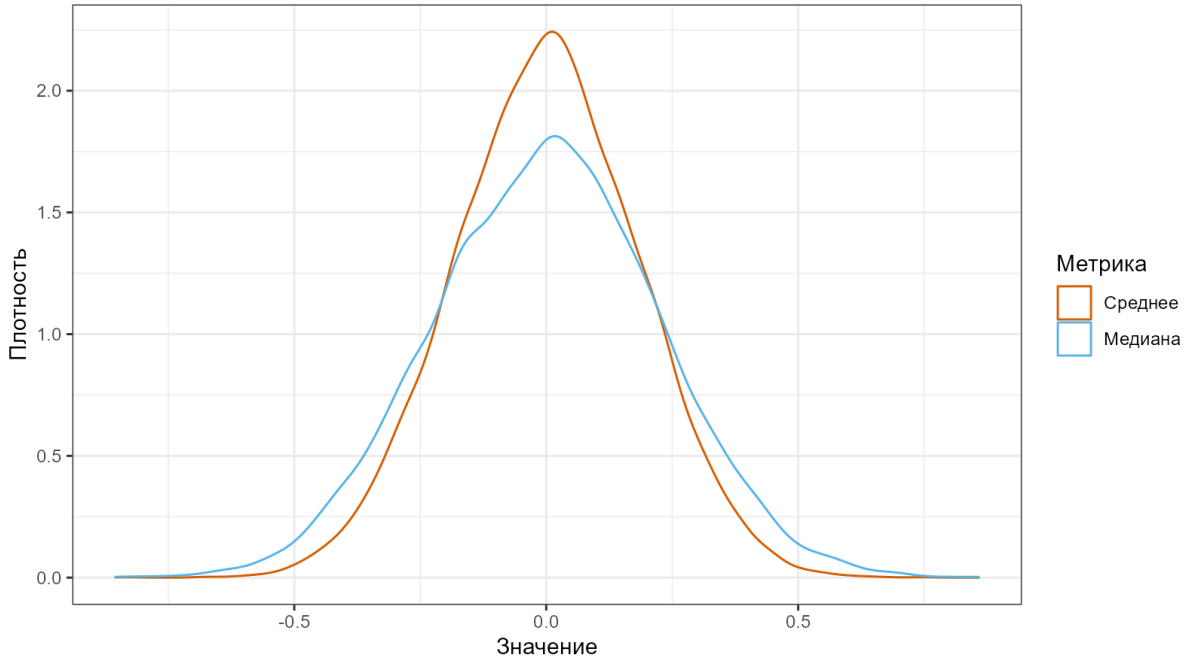
$$x = \{1, 2, 3, 4, 5, 6, 7\}; \quad \tilde{x} = 4$$

$$x = \{1, 2, 3, 4, 5, 6, 273\}; \quad \tilde{x} = 4$$

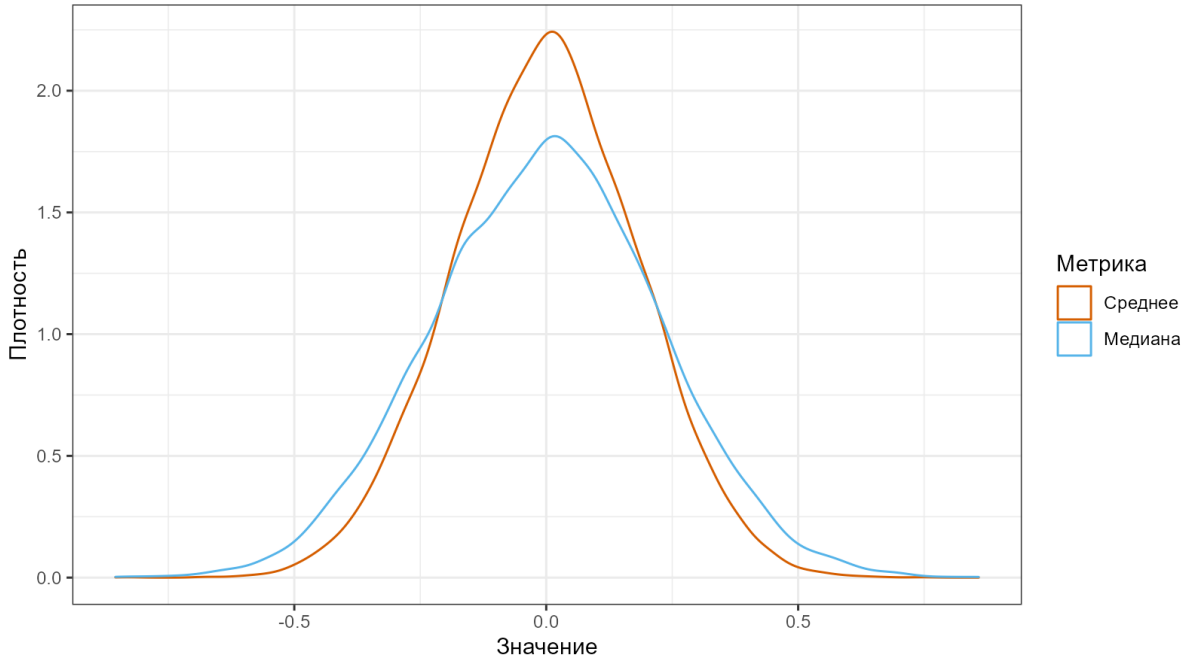
Гауссова эффективность



Гауссова эффективность



$$\mathbf{GaussianEfficiency(T)} = \frac{\mathbb{V}(\mathbf{mean})}{\mathbb{V}(\mathbf{T})}$$



$$\mathbf{GaussianEfficiency(T)} = \frac{\mathbb{V}(\mathbf{mean})}{\mathbb{V}(\mathbf{T})}$$

	Среднее	**Медиана**
Гауссова эффективность	100%	64%

Усечённое и винзоризованное среднее

$$x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}, \quad x_i \leq x_{i+1}$$

$$x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}, \quad x_i \leq x_{i+1}$$

Усечённое среднее:

$$x_{\text{trimmed}} = \{x_3, x_4, x_5, x_6\}$$

$$x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}, \quad x_i \leq x_{i+1}$$

Усечённое среднее:

$$x_{\text{trimmed}} = \{x_3, x_4, x_5, x_6\}$$

$$\overline{x_{\text{trimmed}}} = \frac{x_3 + x_4 + x_5 + x_6}{4}$$

$$x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}, \quad x_i \leq x_{i+1}$$

Усечённое среднее:

$$x_{\text{trimmed}} = \{x_3, x_4, x_5, x_6\}$$

$$\overline{x_{\text{trimmed}}} = \frac{x_3 + x_4 + x_5 + x_6}{4}$$

Винзоризованное среднее:

$$x_{\text{winsorized}} = \{x_3, x_3, x_3, x_4, x_5, x_6, x_6, x_6\}$$

$$x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}, \quad x_i \leq x_{i+1}$$

Усечённое среднее:

$$x_{\text{trimmed}} = \{x_3, x_4, x_5, x_6\}$$

$$\overline{x_{\text{trimmed}}} = \frac{x_3 + x_4 + x_5 + x_6}{4}$$

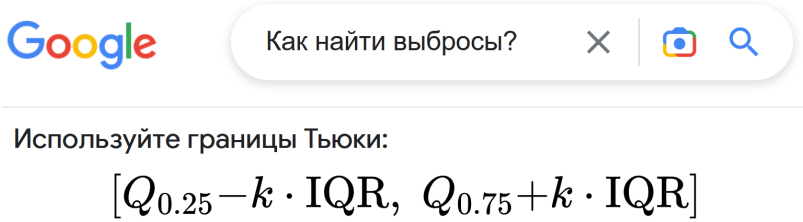
Винзоризованное среднее:

$$x_{\text{winsorized}} = \{x_3, x_3, x_3, x_4, x_5, x_6, x_6, x_6\}$$

$$\overline{x_{\text{winsorized}}} = \frac{x_3 + x_3 + x_3 + x_4 + x_5 + x_6 + x_6 + x_6}{8}$$

Выбрасываем выбросы

Выбрасываем выбросы



Google Как найти выбросы?

Используйте границы Тьюки:

$$[Q_{0.25} - k \cdot \mathrm{IQR},\ Q_{0.75} + k \cdot \mathrm{IQR}]$$

Есть и другие способы:

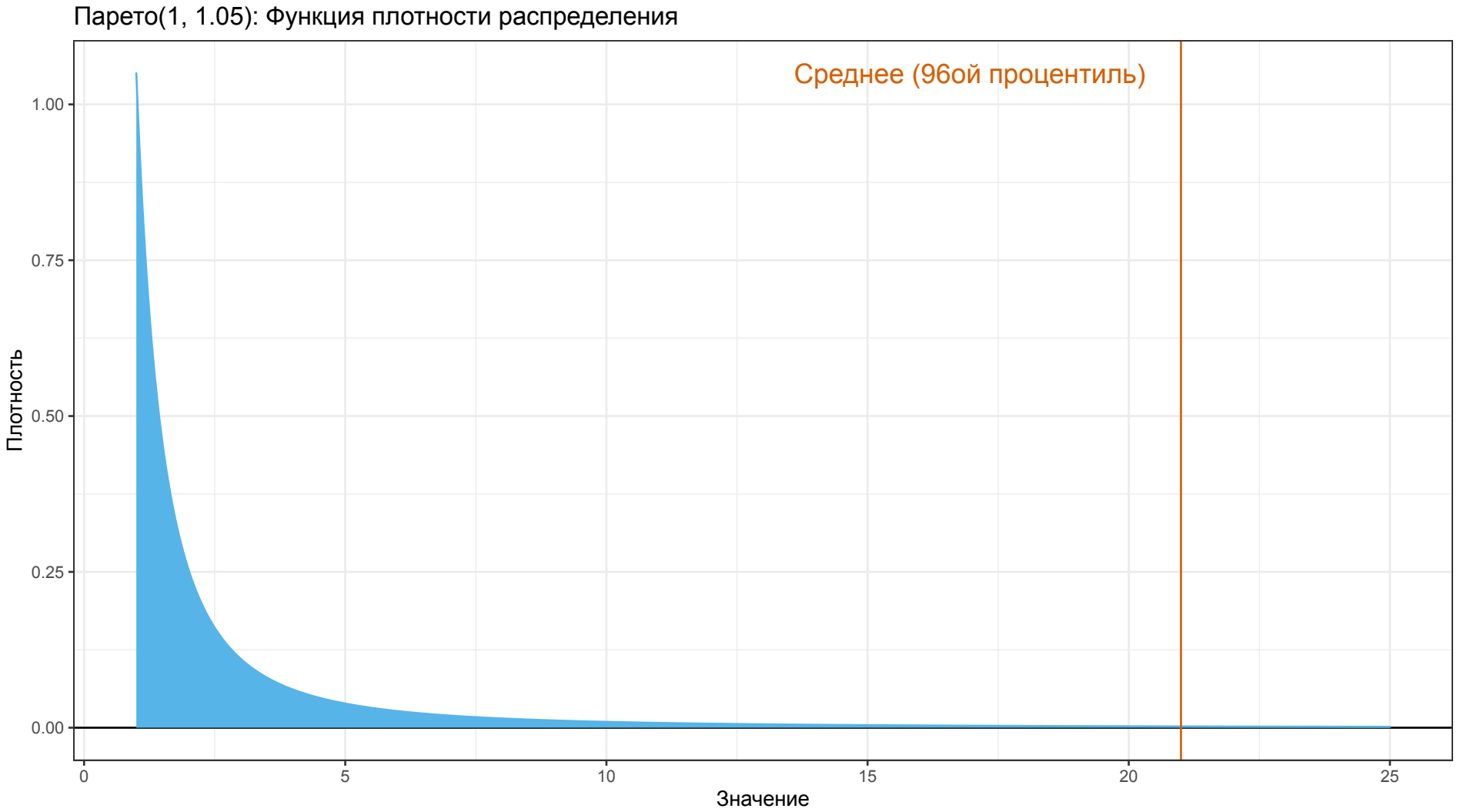
TABLE 1. **The different categories covered by our survey and other related survey.**

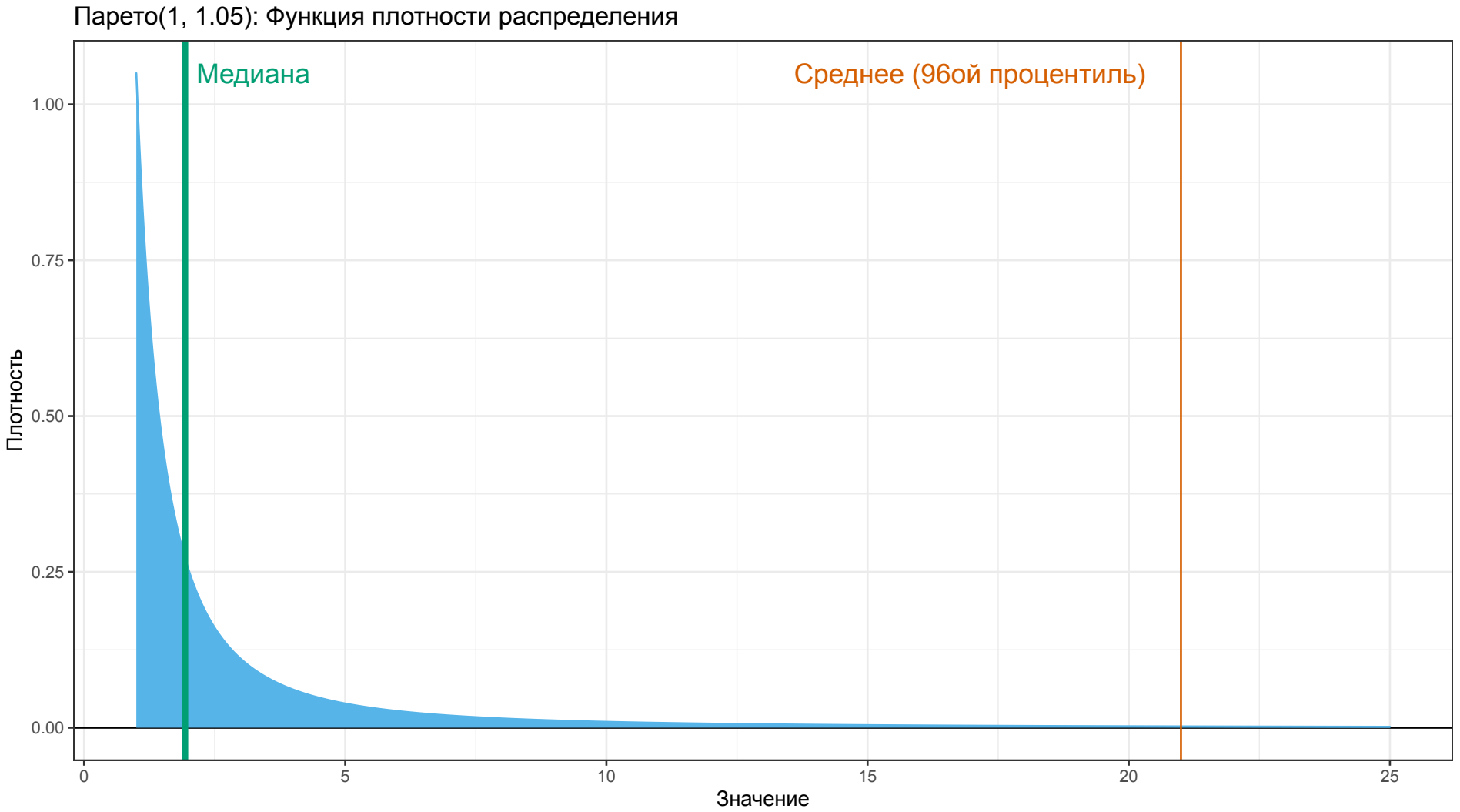
Paper & Year	1	2	3	4	5	6	7	8	9	10	11	12	13
Barnett et al. [39] 1994	X												
Hodge et al. [5] 2004	X	X	X	X		X						X	
Walfish et al. [40] 2006	X												
Patcha et al. [41] 2007	X	X		X		X							
Chandola et al. [22] 2009	X	X	X	X									
Hadi et al. [29] 2009	X	X	X										
Gogoi et al. [33] 2011	X	X	X	X		X		X					
Zhang [26] 2013	X	X	X	X					X				
Gupta et al. [31] 2014									X				
Akoglu et al. [34] 2014							X				X		
Ranshous et al. [23] 2015								X					
Aggarwal [28] 2016					X	X			X		X	X	
Kwon et al. [30] 2017						X							
Chalapathy et al. [32] 2019						X							
Ours	X	X	X	X	X	X	X	X	X	X	X	X	X

1. Statistics-based 2. Distance-based 3. Density-based 4. Clustering-Based 5. Ensemble-Based 6. Learning-Based 7. Graph-based 8. Network 9. Data Streams 10. Tools 11. Datasets 12. Applications 13. References later than 2016.

Wang, Hongzhi, Mohamed Jaward Bah, and Mohamed Hammad. "Progress in outlier detection techniques: A survey." Ieee Access 7 (2019): 107964-108000.

> **The discovery of the ozone hole was announced in 1985** by a British team working on the ground with “conventional” instruments and examining its observations in detail. Only later, after reexamining the data transmitted by the TOMS instrument on NASA’s Nimbus 7 satellite, was it found that **the hole had been forming for several years**. Why had nobody noticed it? The reason was simple: **the systems** processing the TOMS data, designed in accordance with predictions derived from models, which in turn were established on the basis of what was thought to be “reasonable”, **had rejected the very (“excessively”) low values** observed above the Antarctic during the Southern spring. As far as the program was concerned, there must have been an operating defect in the instrument.
>
> --- R. Kandel, Our Changing Climate (1991)


Парето(1, 1.05): Функция плотности распределения
Среднее (96ой процентиль)
1.00
0.75
0.50
0.25
0.00
Плотность
0
5
10
15
20
25
Значение


Парето(1, 1.05): Функция плотности распределения
Медиана
Среднее (96ой процентиль)
1.00
0.75
0.50
0.25
0.00
Плотность
0
5
10
15
20
25
Значение

Оценка Ходжеса-Леманна

$$\mathrm{HL} = \underset{i<j}{\mathrm{median}} \left(\frac{x_i + x_j}{2} \right)$$

Hodges Jr, Joseph L., and Erich L. Lehmann. "Estimates of location based on rank tests." The Annals of Mathematical Statistics (1963): 598-611.

	Среднее	Медиана	Ходжес-Леманн
Гауссова эффективность	100%	64%	96%

	Среднее	**Медиана**	**Ходжес-Леманн**
Гауссова эффективность	100%	64%	96%
Точка перелома	0%	50%	29%

Другие метрики

Другие метрики

- Моды

- Моды
- Геометрическое среднее: $\sqrt[n]{x_1 x_2 \ldots x_n}$

- Моды
- Геометрическое среднее: $\sqrt[n]{x_1 x_2 \ldots x_n}$
- Гармоническое среднее: $n/(x_1^{-1} + x_2^{-1} + \ldots + x_n^{-1})$

- Моды
- Геометрическое среднее: $\sqrt[n]{x_1 x_2 \ldots x_n}$
- Гармоническое среднее: $n/(x_1^{-1} + x_2^{-1} + \ldots + x_n^{-1})$
- Midhinge: $(Q_{0.25}(x) + Q_{0.75}(x))/2$

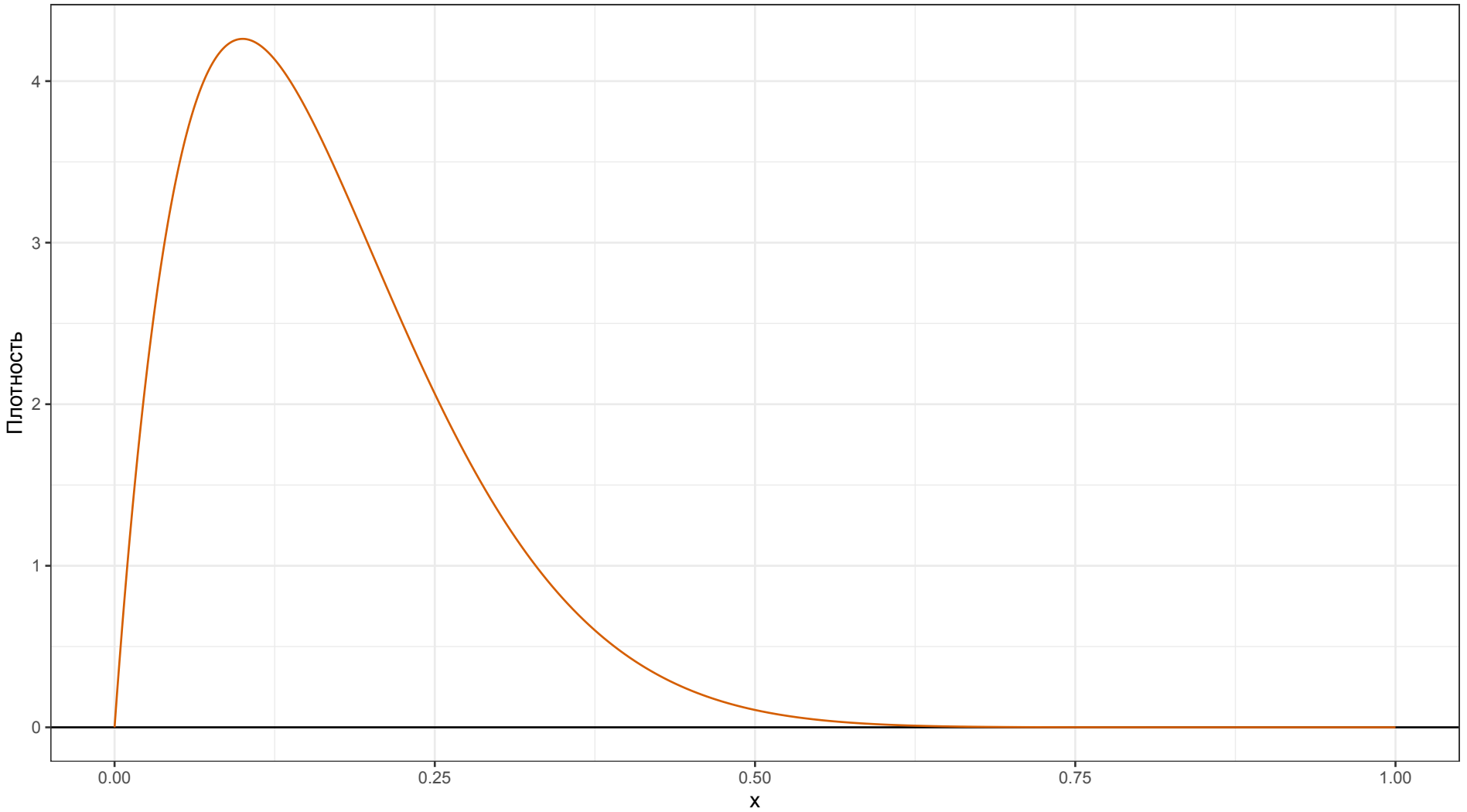
- Моды
- Геометрическое среднее: $\sqrt[n]{x_1 x_2 \ldots x_n}$
- Гармоническое среднее: $n/(x_1^{-1} + x_2^{-1} + \ldots + x_n^{-1})$
- Midhinge: $(Q_{0.25}(x) + Q_{0.75}(x))/2$
- Trimean: $(Q_{0.25}(x) + 2Q_{0.50}(x) + Q_{0.75}(x))/4$

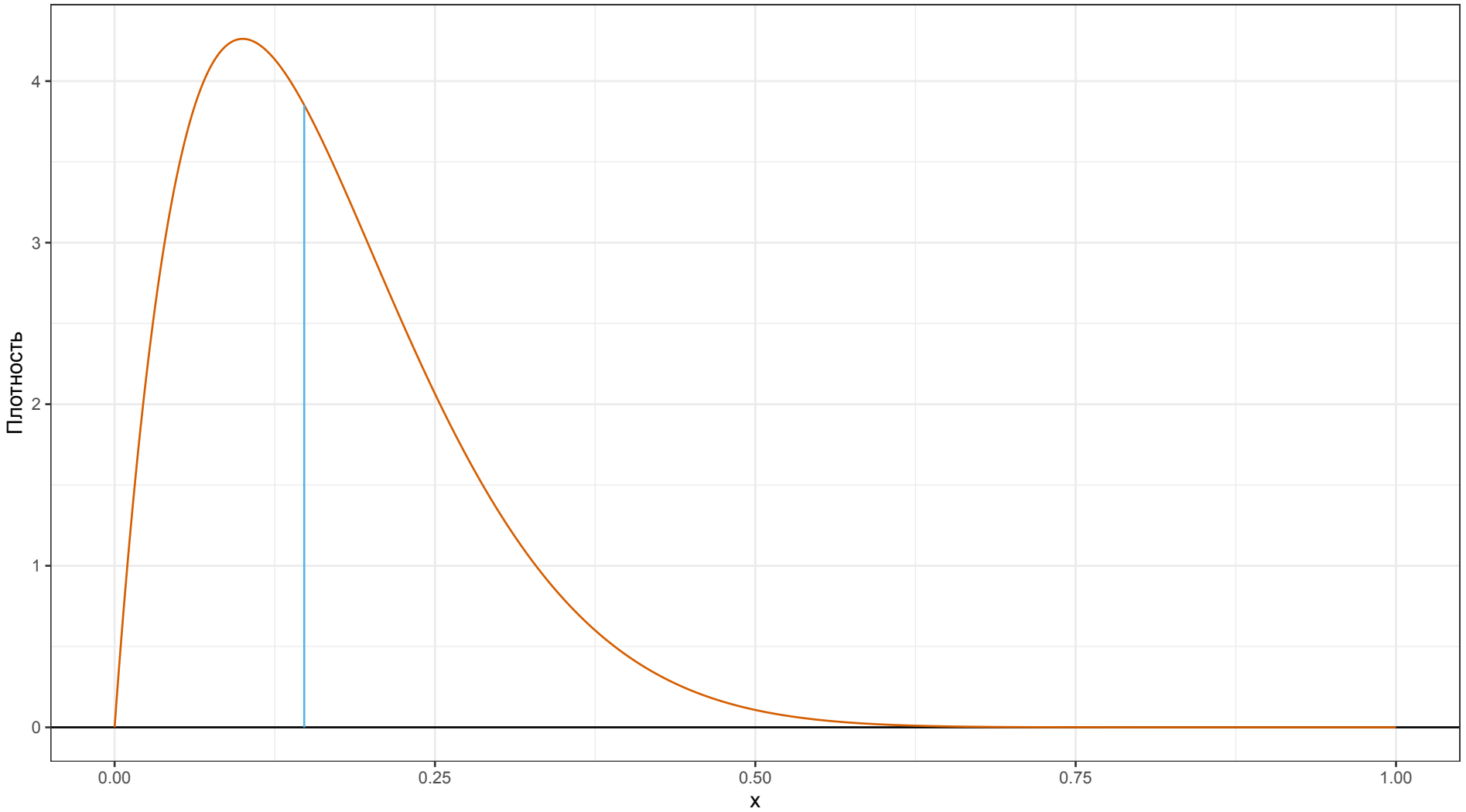
- Моды
- Геометрическое среднее: $\sqrt[n]{x_1 x_2 \ldots x_n}$
- Гармоническое среднее: $n/(x_1^{-1} + x_2^{-1} + \ldots + x_n^{-1})$
- Midhinge: $(Q_{0.25}(x) + Q_{0.75}(x))/2$
- Trimean: $(Q_{0.25}(x) + 2Q_{0.50}(x) + Q_{0.75}(x))/4$
- ...

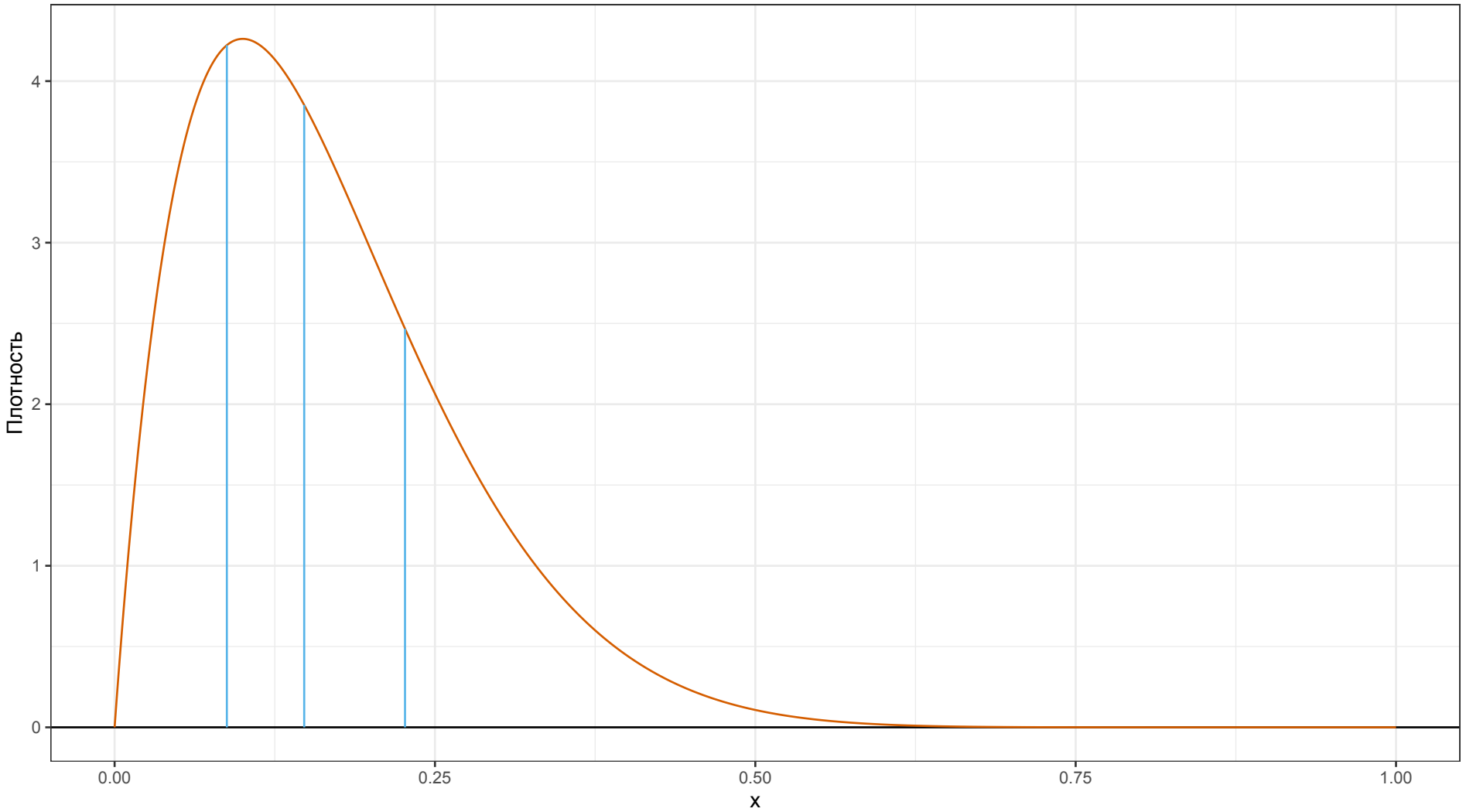
Для выбора центральной тенденции
нужно подумать

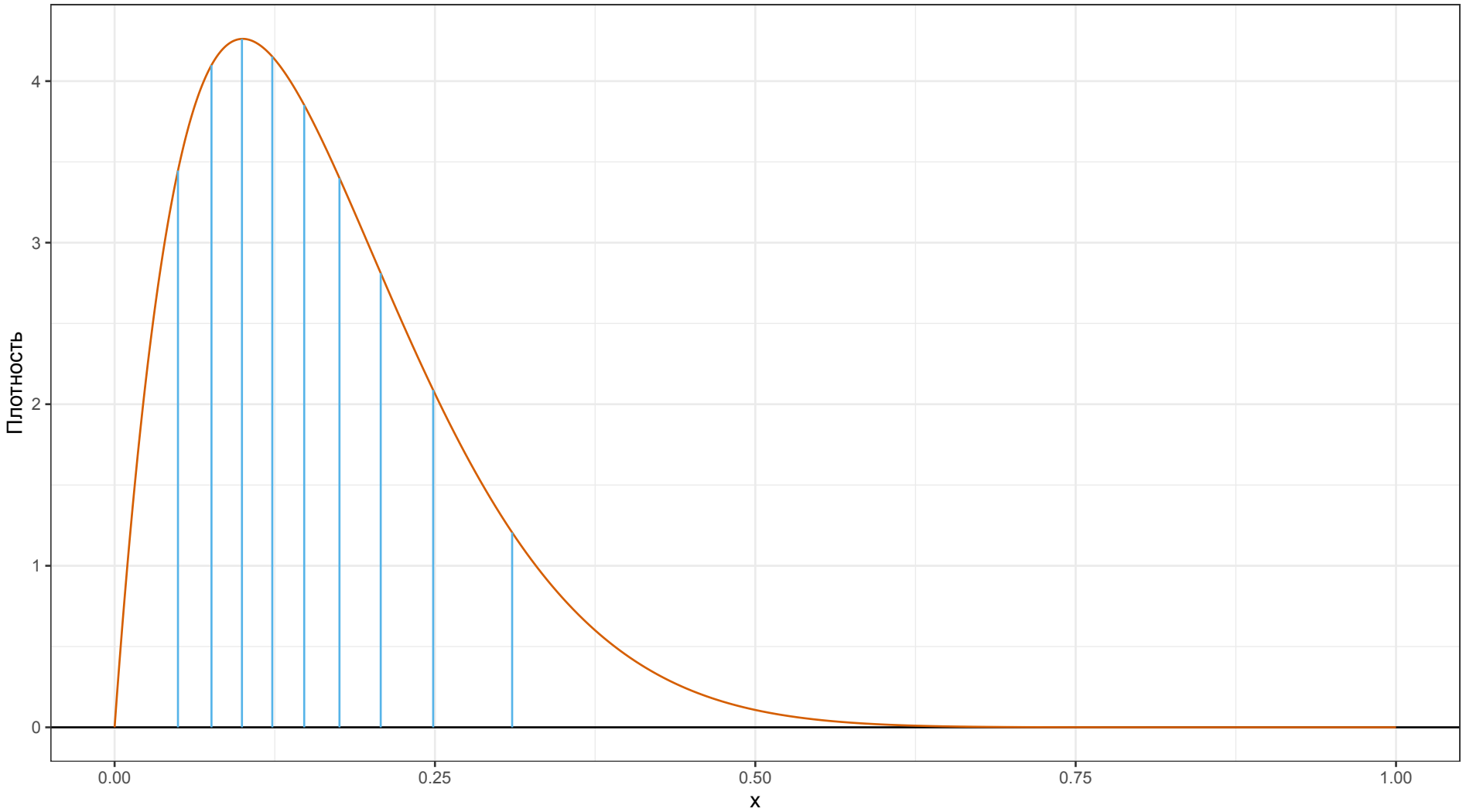
Квантильные оценки

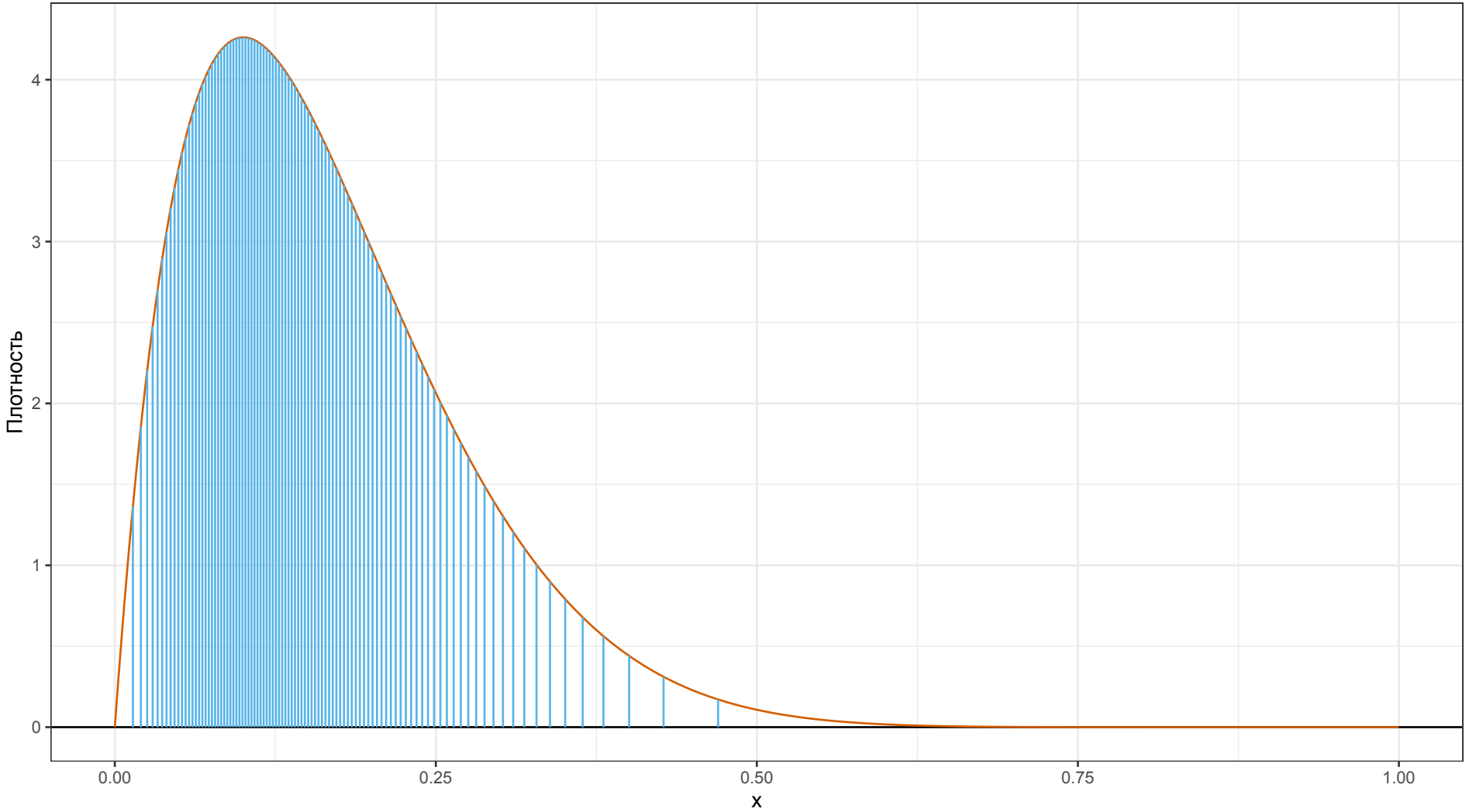
Функция плотность распределения




Плотность
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
x


Плотность
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
x


Плотность
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
x


Плотность
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
x

Квантили распределения и квантили выборки

Выборка из нормального распределения:

Выборка из нормального распределения:

$$x = \{-1.37, -0.19, 0.14, 0.16, 0.26, 0.39, 0.46, 0.74, 0.89\}$$

Выборка из нормального распределения:

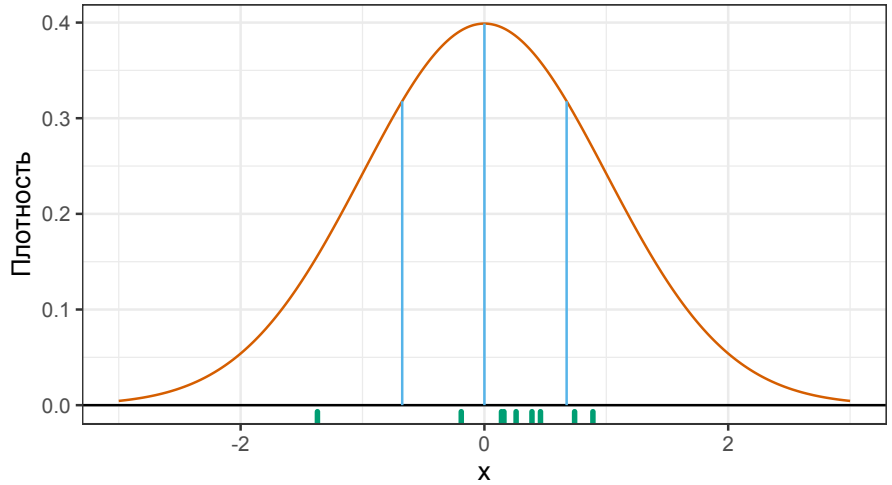
$$x = \{-1.37, -0.19, 0.14, 0.16, 0.26, 0.39, 0.46, 0.74, 0.89\}$$

	Выборка	
$Q_{0.00}$	-1.37	
$Q_{0.25}$	0.14	
$Q_{0.50}$	0.26	
$Q_{0.75}$	0.46	
$Q_{1.00}$	0.89	

Выборка из нормального распределения:

$$x = \{-1.37, -0.19, 0.14, 0.16, 0.26, 0.39, 0.46, 0.74, 0.89\}$$

	Выборка	Распределение
$Q_{0.00}$	-1.37	$-\infty$
$Q_{0.25}$	0.14	-0.67
$Q_{0.50}$	0.26	0.00
$Q_{0.75}$	0.46	0.67
$Q_{1.00}$	0.89	$+\infty$



Тип	h	Уравнение
1	$Np + 1/2$	$x_{\lceil h-1/2 \rceil}$
2	$Np + 1/2$	$(x_{\lceil h-1/2 \rceil} + x_{\lceil h+1/2 \rceil})/2$
3	Np	$x_{\lfloor h \rceil}$
4	Np	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor})$
5	$Np + 1/2$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor})$
6	$(N + 1)p$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor})$
7	$(N - 1)p + 1$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor})$
8	$(N + 1/3)p + 1/3$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor})$
9	$(N + 1/4)p + 3/8$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor})$

Hyndman, R. J. and Fan, Y. 1996. Sample quantiles in statistical packages, *American Statistician* 50, 361–365.

Квантильная оценка Харрела-Дэвиса

Harrell, F.E. and Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika, 69(3), pp.635-640.

Квантильная оценка Харрела-Дэвиса

$$Q_{HD}(p) = \sum_{i=1}^{n} W_i \cdot x_{(i)}$$

Harrell, F.E. and Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika, 69(3), pp.635-640.

$$Q_{HD}(p) = \sum_{i=1}^{n} W_i \cdot x_{(i)}$$

$$W_i = I_{i/n}(a, b) - I_{(i-1)/n}(a, b)$$

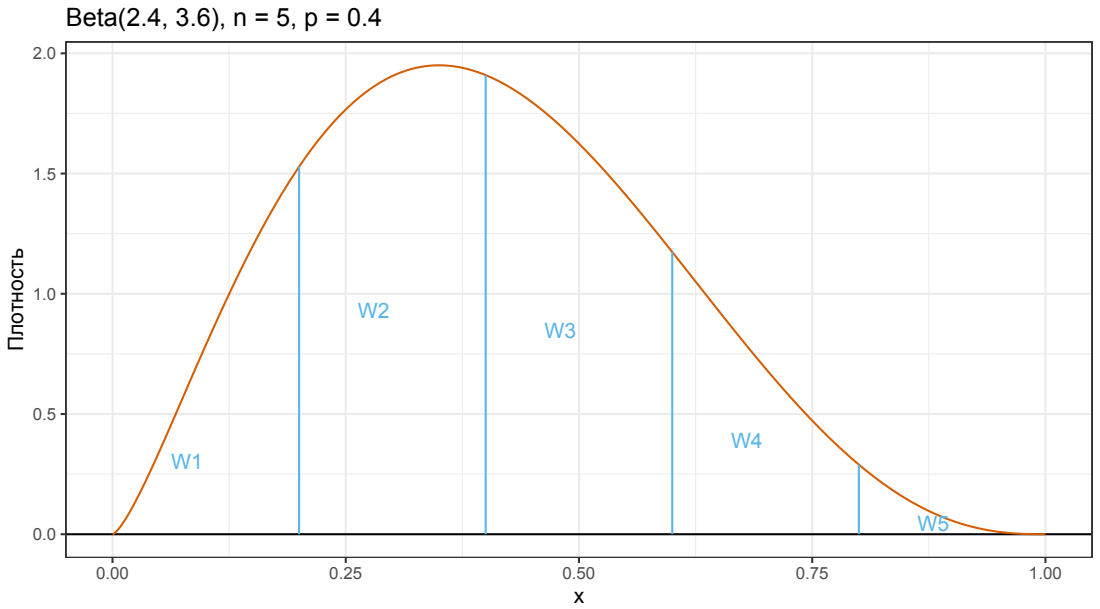
$$a = p(n+1),\ b = (1-p)(n+1)$$

Harrell, F.E. and Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika, 69(3), pp.635-640.

$$Q_{HD}(p) = \sum_{i=1}^{n} W_i \cdot x_{(i)}$$

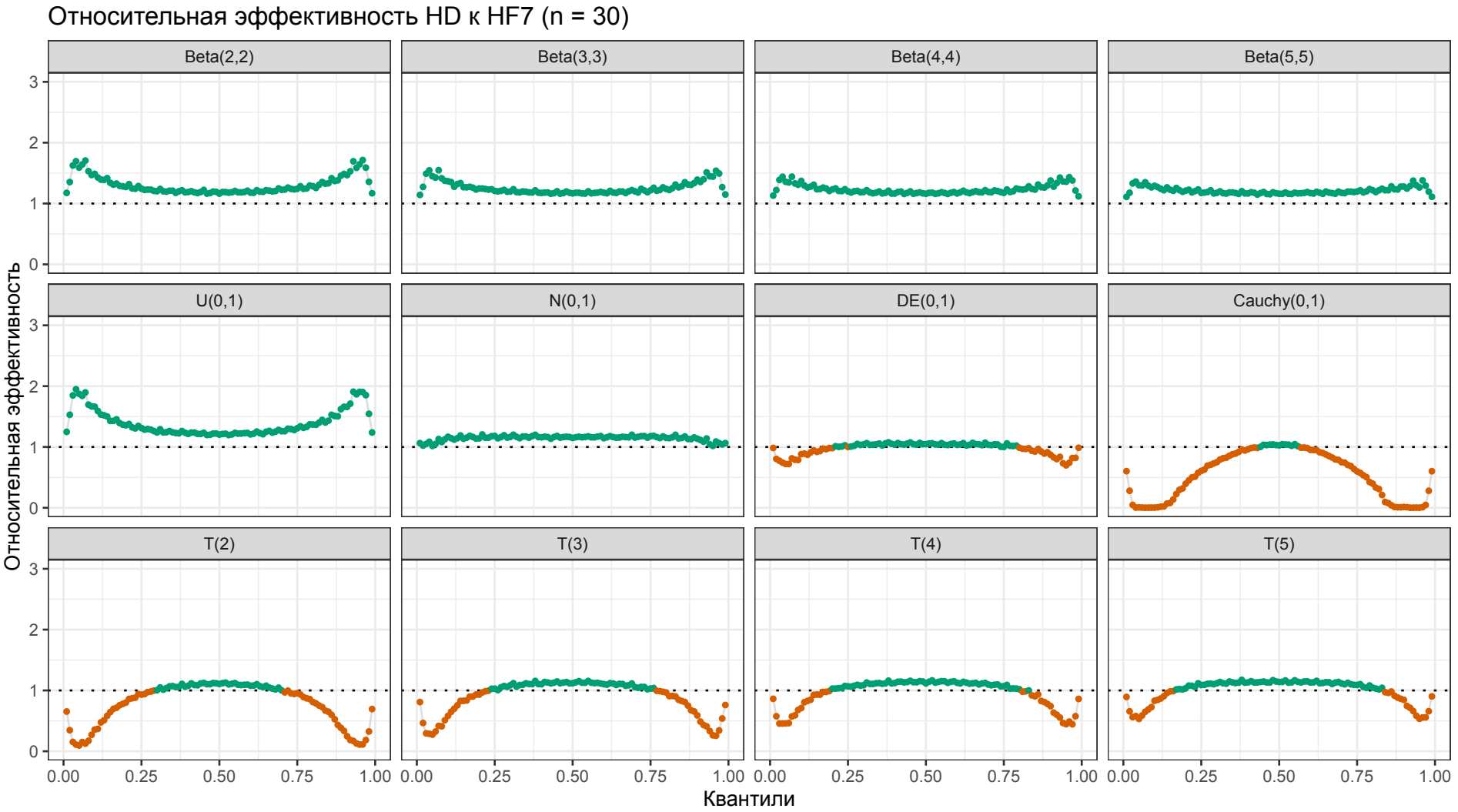
$$W_i = I_{i/n}(a, b) - I_{(i-1)/n}(a, b)$$

$$a = p(n+1),\ b = (1-p)(n+1)$$



Harrell, F.E. and Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika, 69(3), pp.635-640.

Эффективность квантильной оценки Харрела-Дэвиса



Andrey Akinshin (2022) Trimmed Harrell-Davis quantile estimator based on the highest density interval of the given width, Communications in Statistics - Simulation and Computation

Доверительные интервалы Мэритц-Джэррэт

k-ый момент:

$$C_k = \sum_{i=1}^{n} W_i \cdot x_{(i)}^k$$

k-ый момент:

$$C_k = \sum_{i=1}^{n} W_i \cdot x_{(i)}^k$$

Квантильная оценка Харрела-Дэвиса:

$$Q_{HD}(p) = C_1$$

k-ый момент:

$$C_k = \sum_{i=1}^{n} W_i \cdot x_{(i)}^k$$

Квантильная оценка Харрела-Дэвиса:

$$Q_{HD}(p) = C_1$$

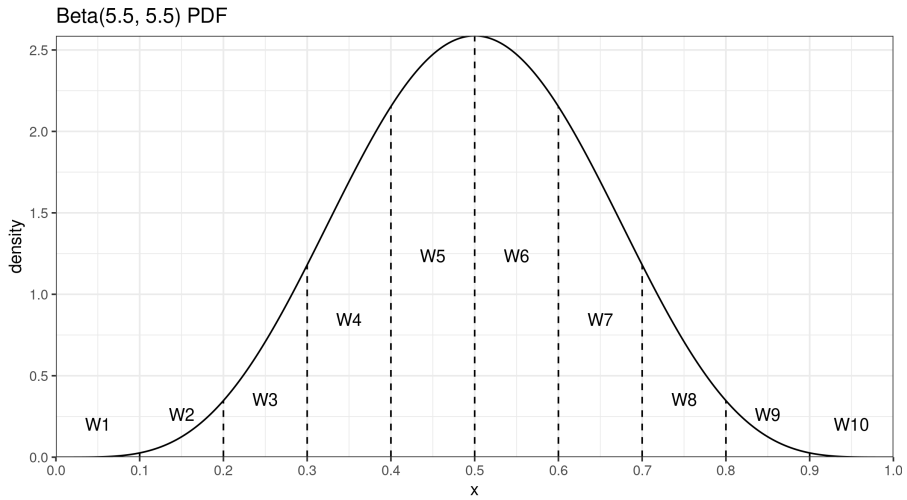
Оценка доверительного интервала Мэритц-Джэррэт:

$$s_{MJ}(p) = \sqrt{C_2 - C_1^2}$$

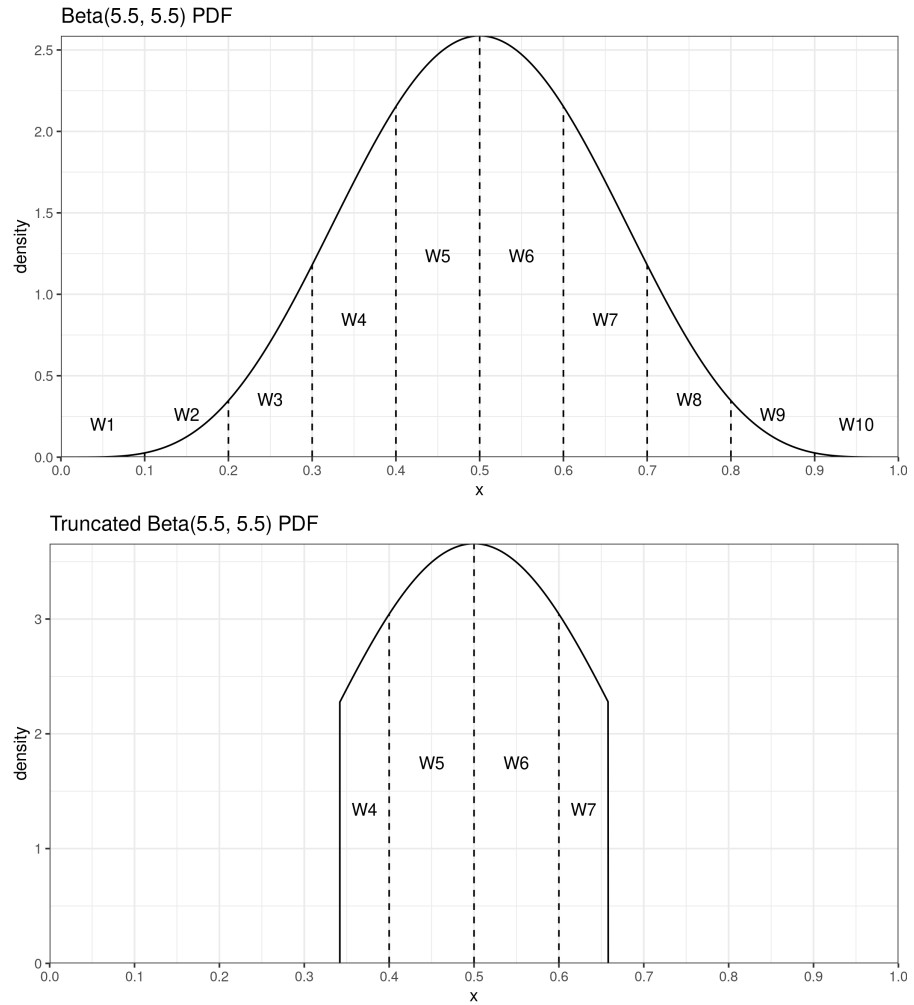
Maritz, J. S., and R. G. Jarrett. 1978. “A Note on Estimating the Variance of the Sample Median.” Journal of the American Statistical Association 73 (361): 194–196.

Усечённая квантильная оценка Харрела-Дэвиса

Усечённая квантильная оценка Харрела-Дэвиса



Усечённая квантильная оценка Харрела-Дэвиса



Andrey Akinshin (2022) Trimmed Harrell-Davis quantile estimator based on the highest density interval of the given width, Communications in Statistics - Simulation and Computation

Квантильная оценка Сфакианакиса-Виргиниса

$$
\begin{aligned}
\mathrm{SV1}_p = & \frac{B_0}{2}\left(X_{(1)} + X_{(2)} - X_{(3)}\right) + \\
& \sum_{i=1}^{n} \frac{B_i + B_{i-1}}{2} X_{(i)} + \\
& \frac{B_n}{2}\left(-X_{(n-2)} + X_{(n-1)} - X_{(n)}\right), \\
\mathrm{SV2}_p = & \sum_{i=1}^{n} B_{i-1} X_{(i)} + B_n \cdot \left(2X_{(n)} - X_{(n-1)}\right), \\
\mathrm{SV3}_p = & \sum_{i=1}^{n} B_i X_{(i)} + B_0 \cdot \left(2X_{(1)} - X_{(2)}\right).
\end{aligned}
$$

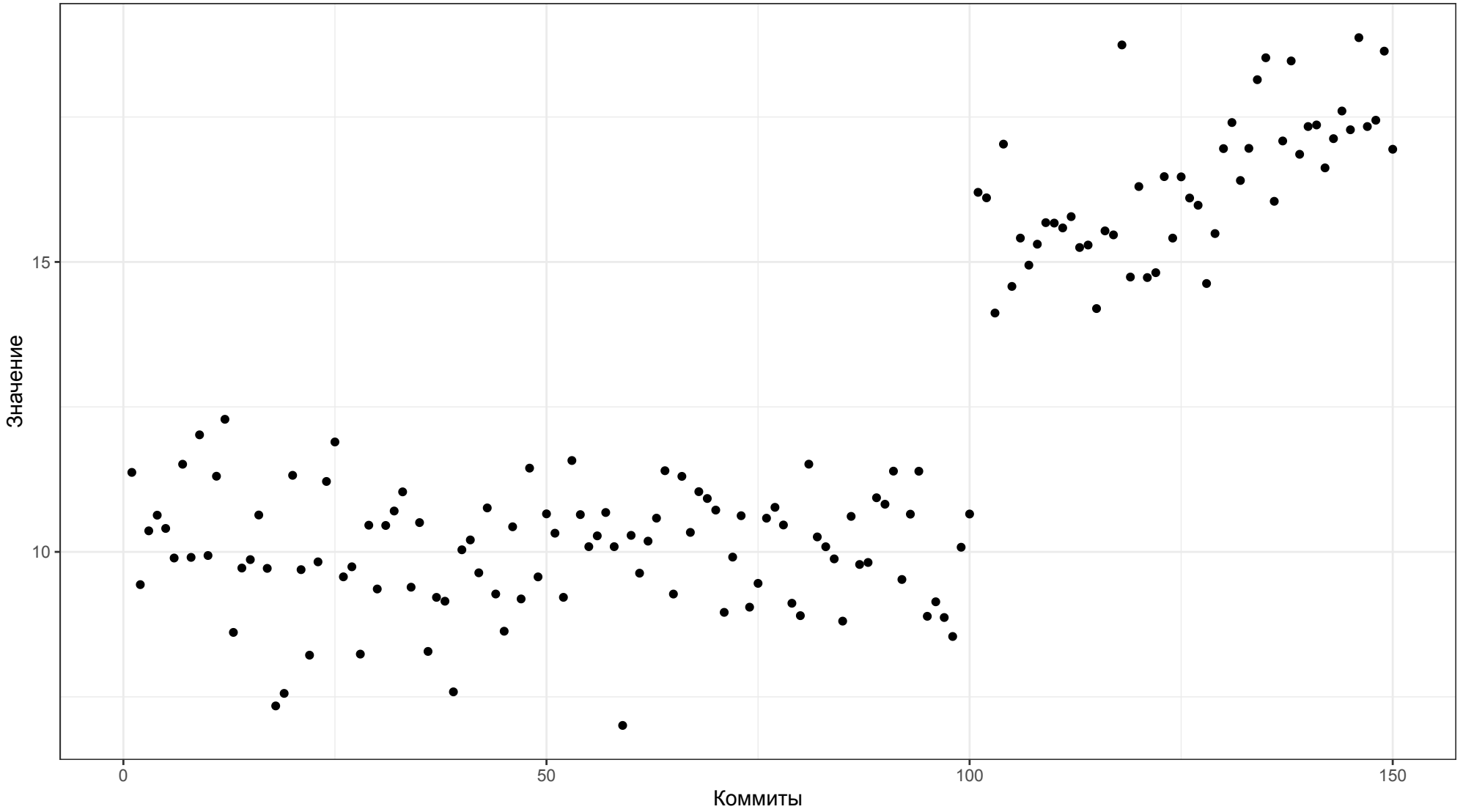
Sfakianakis, Michael E., and Dimitris G. Verginis. “A new family of nonparametric quantile estimators.” Communications in Statistics—Simulation and Computation® 37, no. 2 (2008): 337-345.
See also: https://aakinshin.net/posts/sfakianakis-verginis-quantile-estimator/

$$\begin{aligned}\mathrm{NO}_p = &\Big((3p-1)X_{(1)} + (2-3p)X_{(2)} - (1-p)X_{(3)}\Big)B_0 + \\ &+ \sum_{i=1}^{n}\Big((1-p)B_{i-1} + pB_i\Big)X_{(i)} + \\ &+ \Big(-pX_{(n-2)} + (3p-1)X_{(n-1)} + (2-3p)X_{(n)}\Big)B_n\end{aligned}$$

Navruz, Gözde, and A. Fırat Özdemir. "A new quantile estimator with weights based on a subsampling approach." British Journal of Mathematical and Statistical Psychology 73, no. 3 (2020): 506-521.
See also: https://aakinshin.net/posts/navruz-ozdemir-quantile-estimator/

Проблема: агрегация истории коммитов

Проблема: агрегация истории коммитов



$$x = \{x_1, x_2, \ldots, x_n\}$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$w = \{w_1, w_2, \ldots, w_n\}$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$w = \{w_1, w_2, \ldots, w_n\}$$

Экспоненциальный закон:

$$w_i = \alpha^{n-i}$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$w = \{w_1, w_2, \ldots, w_n\}$$

Экспоненциальный закон:

$$w_i = \alpha^{n-i}$$

	Общий вид	**Пример**
w_n	α^0	1
w_{n-1}	α^1	0.5
w_{n-2}	α^2	0.25
w_{n-3}	α^3	0.125
...	...	...

Экспоненциальное сглаживание среднего арифметического[1]

$$s_t = \alpha x_t + (1 - \alpha) s_{t-1}$$

[1] https://aakinshin.net/posts/quantile-exponential-smoothing/

Экспоненциальное сглаживание среднего арифметического[1]

$$s_t = \alpha x_t + (1 - \alpha) s_{t-1}$$

Взвешенная квантильная оценка Харрела-Дэвиса[2]

$$Q_{HD} = \sum_{i=1}^{n} W_i \cdot x_{(i)}, \quad W_i = I_{r_i}(a, b) - I_{l_i}(a, b)$$

$$l_i = \frac{s_{i-1}}{s_n}, \quad r_i = \frac{s_i}{s_n}, \quad s_i = w_1 + w_2 + \ldots + w_i$$

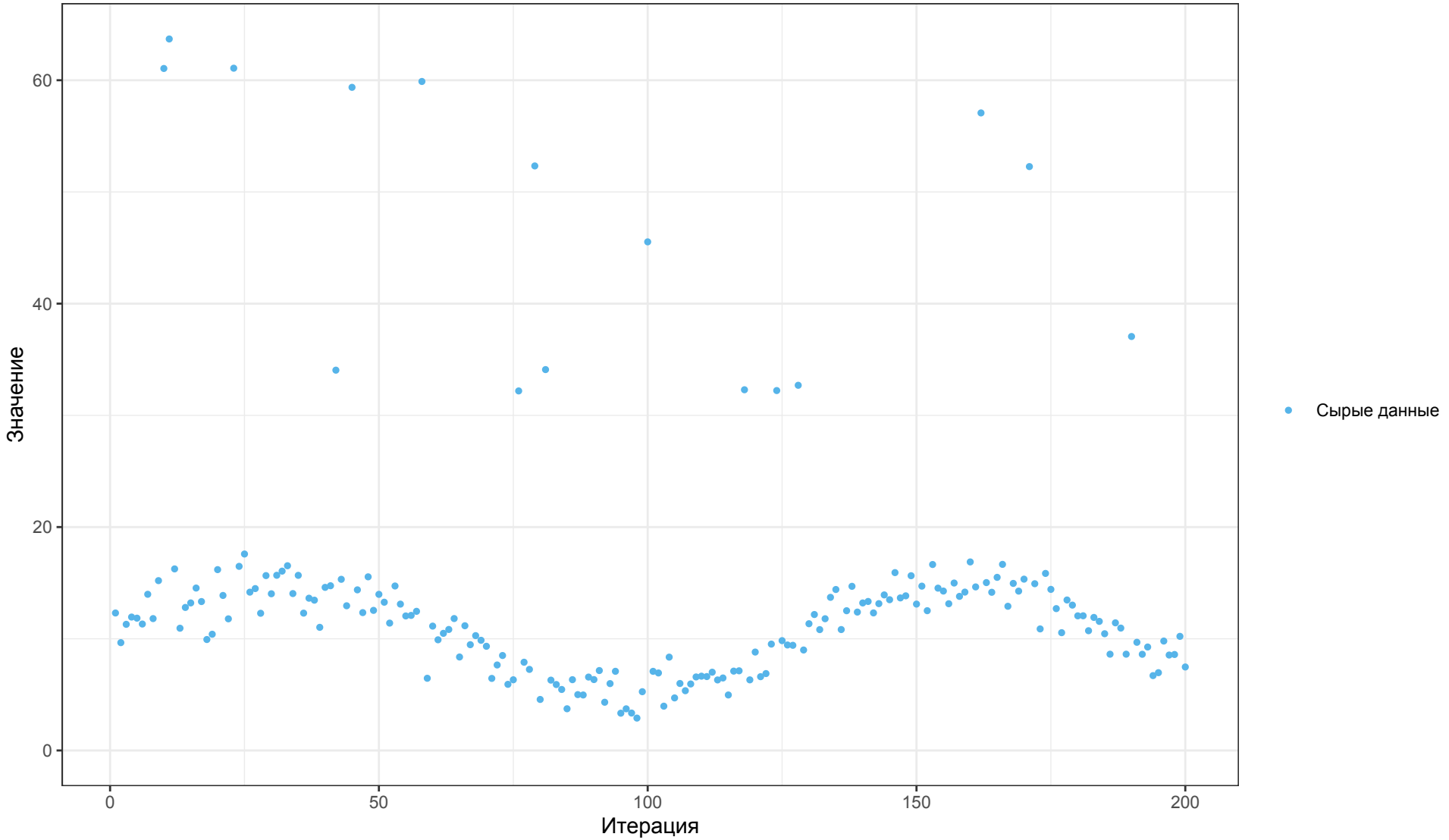
[1] https://aakinshin.net/posts/quantile-exponential-smoothing/

[2] https://aakinshin.net/posts/weighted-quantiles/
https://aakinshin.net/posts/weighted-quantiles-ci/

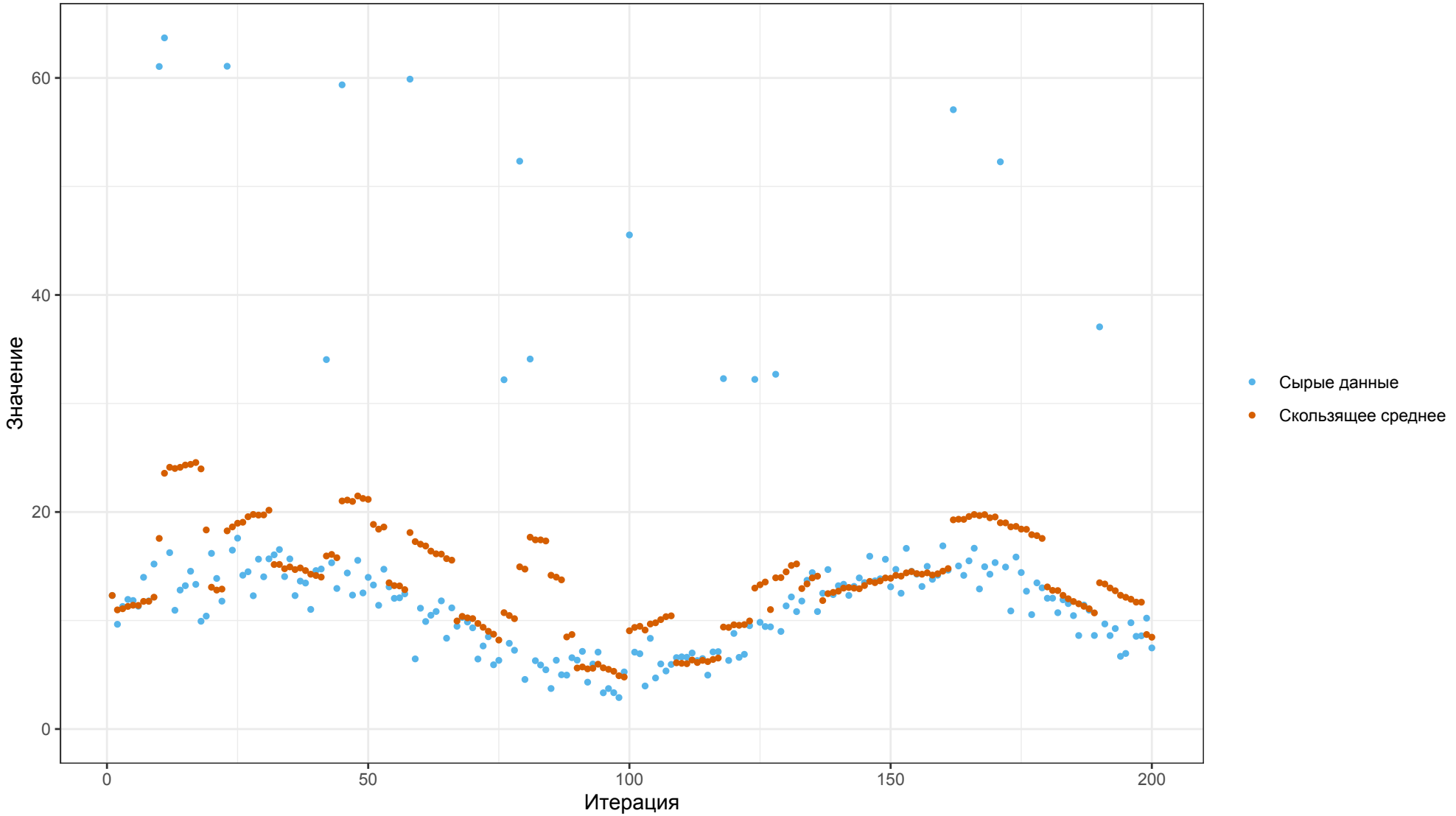
Для выбора квантильной оценки
нужно подумать

Скользящие квантильные оценки

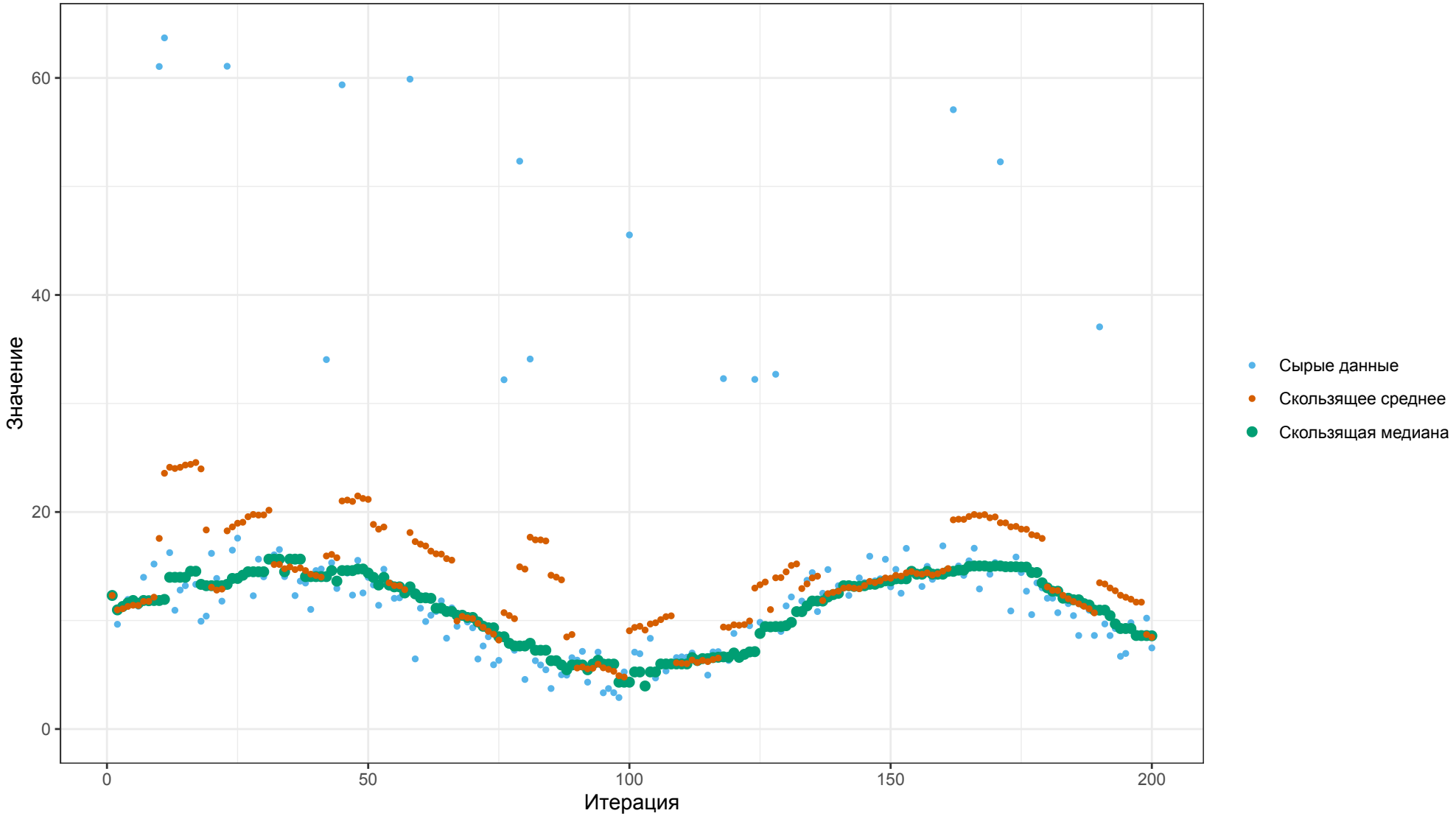
Скользящее среднее и скользящая медиана



Скользящее среднее и скользящая медиана

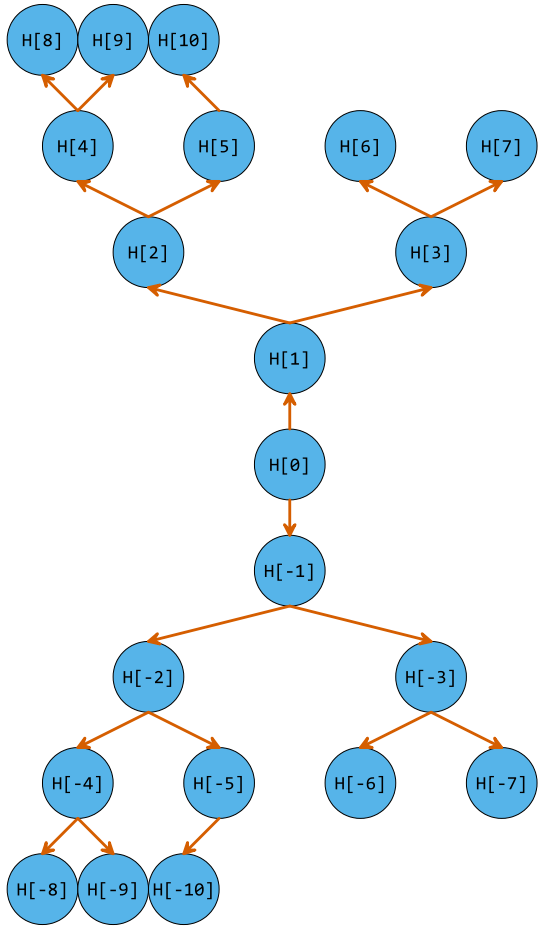


Скользящее среднее и скользящая медиана



Разделённые кучи (Partitioning heaps)

Разделённые кучи (Partitioning heaps)



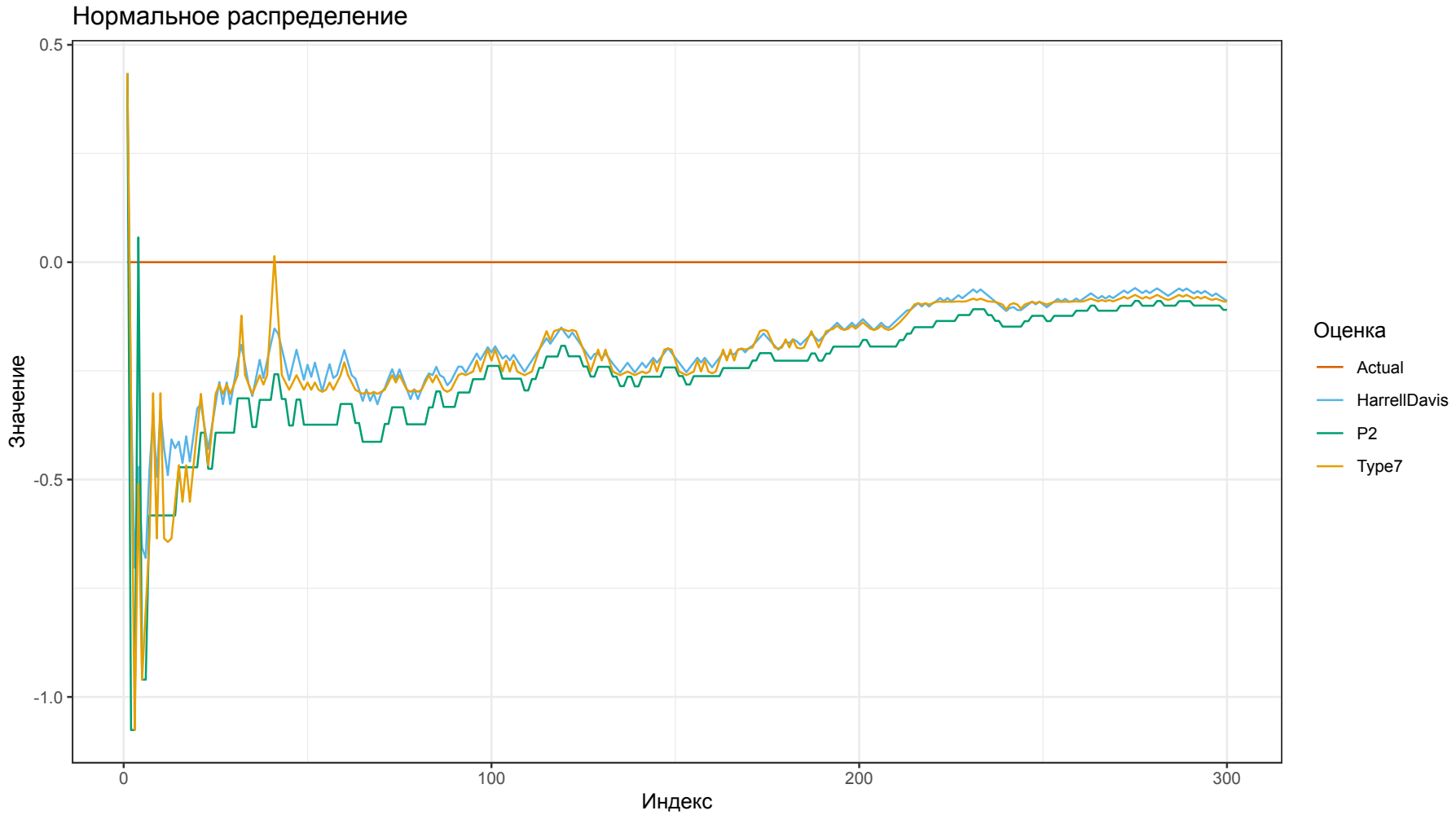
Hardle, W., and William Steiger. “Algorithm AS 296: Optimal median smoothing.” Journal of the Royal Statistical Society. Series C (Applied Statistics) 44, no. 2 (1995): 258-264.

See also: https://aakinshin.net/posts/partitioning-heaps-quantile-estimator2/

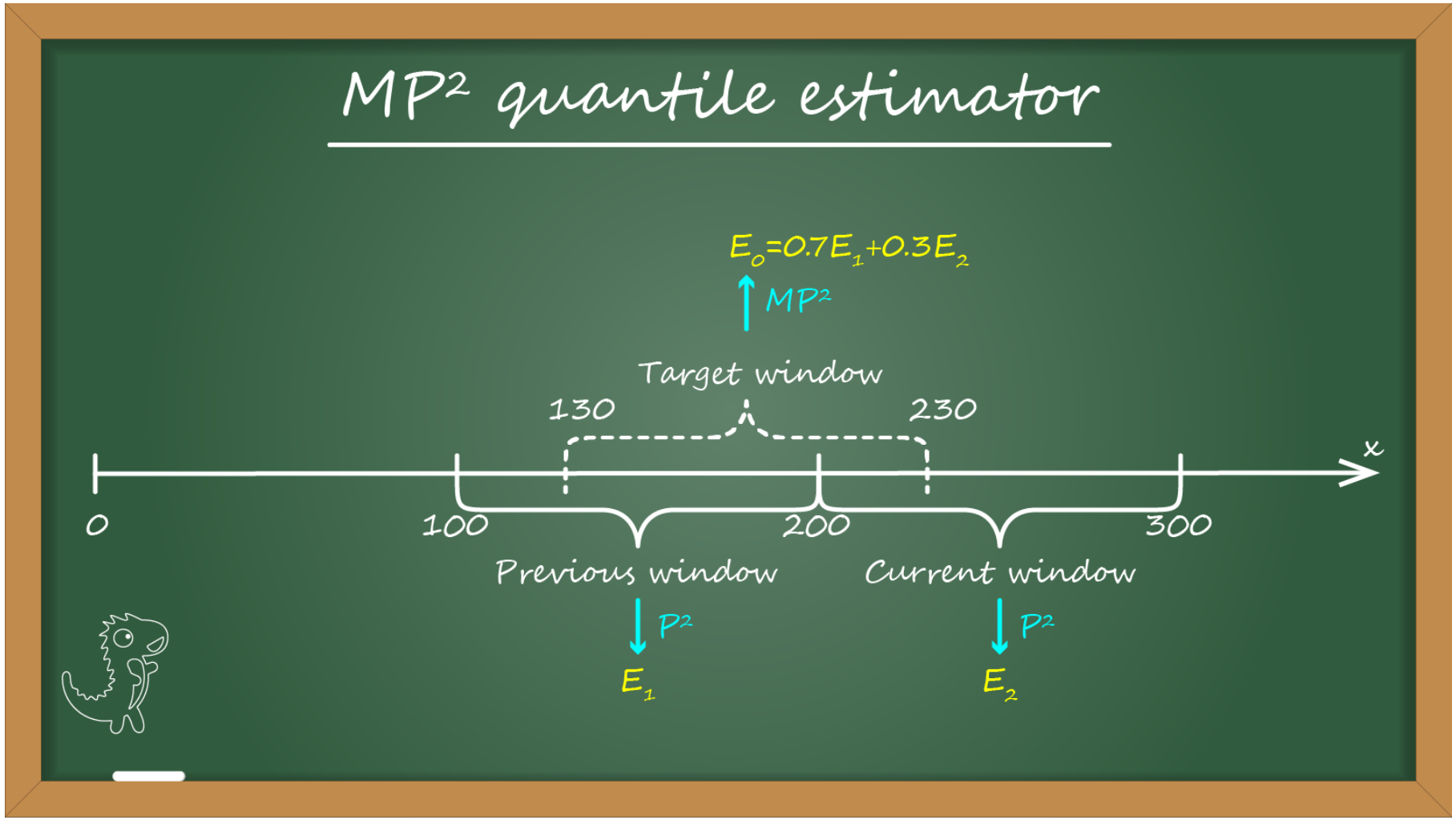
Квантильная оценка P²

- q_0: Минимум
- q_1: (p/2)-ый квантиль
- q_2: p-ый квантиль
- q_3: ((1+p)/2)-ый квантиль
- q_4: Максимум

Jain, Raj, and Imrich Chlamtac. "The P2 algorithm for dynamic calculation of quantiles and histograms without storing observations." Communications of the ACM 28, no. 10 (1985): 1076-1085
See also: https://aakinshin.net/posts/p2-quantile-estimator/

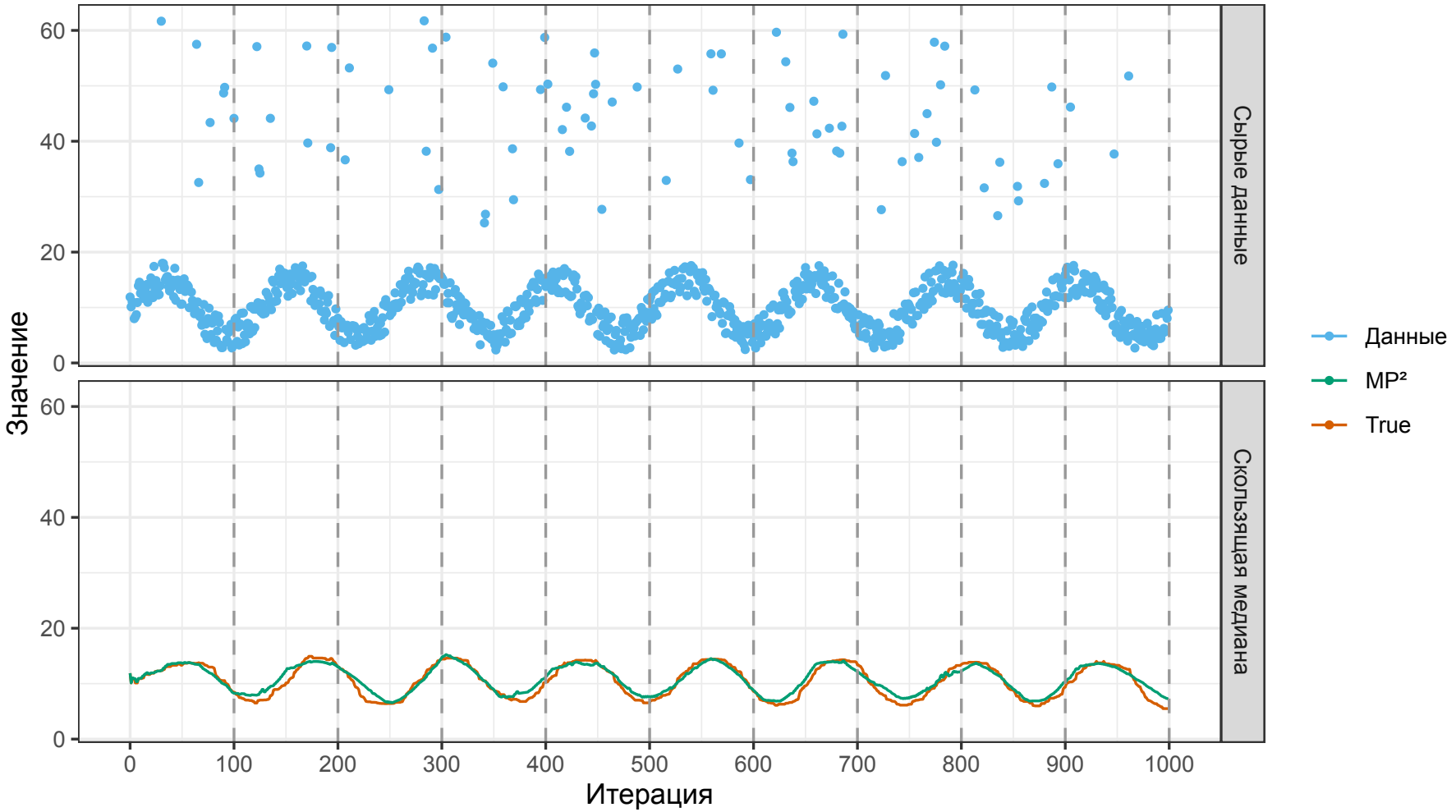


Jain, Raj, and Imrich Chlamtac. "The P2 algorithm for dynamic calculation of quantiles and histograms without storing observations." Communications of the ACM 28, no. 10 (1985): 1076-1085
See also: https://aakinshin.net/posts/p2-quantile-estimator/



https://aakinshin.net/posts/mp2-quantile-estimator/

Квантильная оценка MP²



https://aakinshin.net/posts/mp2-quantile-estimator/

tdunning / t-digest Public

Fork 219 Star 1.7k

- has smaller summaries when serialized
- works on double precision floating point as well as integers.
- provides part per million accuracy for extreme quantiles and typically <1000 ppm accuracy for middle quantiles
- is very fast (~ 140 ns per add)
- is very simple (~ 5000 lines of code total, <1000 for the most advanced implementation alone)
- has a reference implementation that has > 90% test coverage
- can be used with map-reduce very easily because digests can be merged
- requires no dynamic allocation after initial creation (`MergingDigest` only)
- has no runtime dependencies

https://github.com/tdunning/t-digest

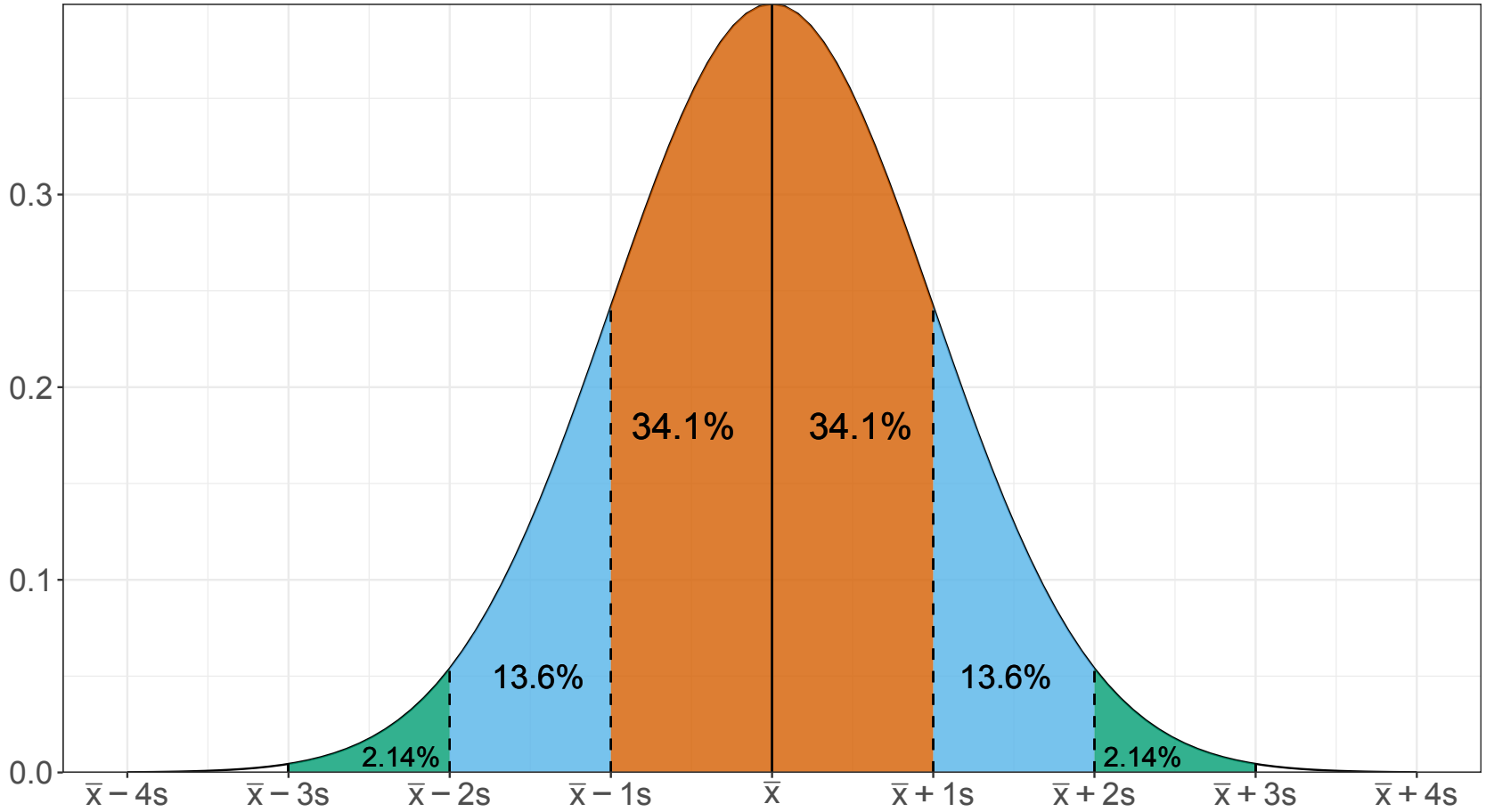
Для выбора квантильной оценки
нужно подумать

Вариация

Русский	English
Вариация	Dispersion
Дисперсия	Variance

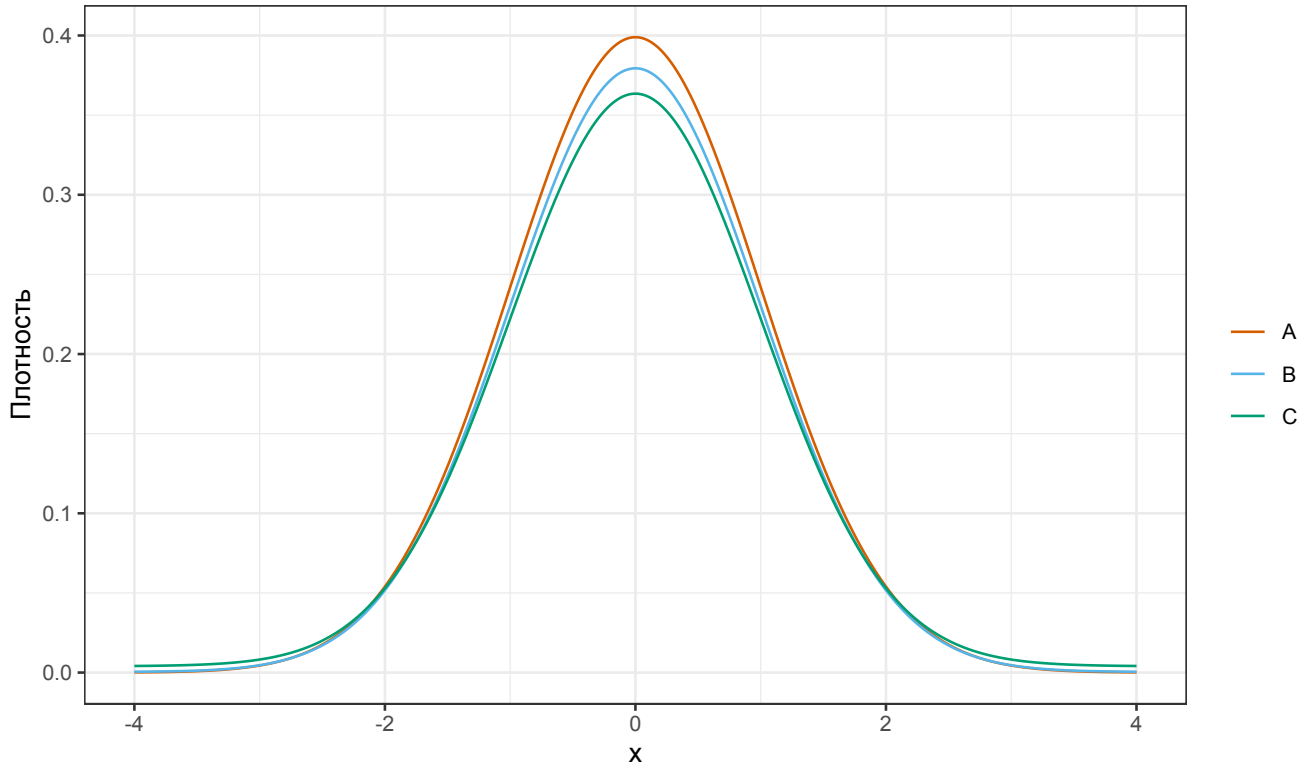
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

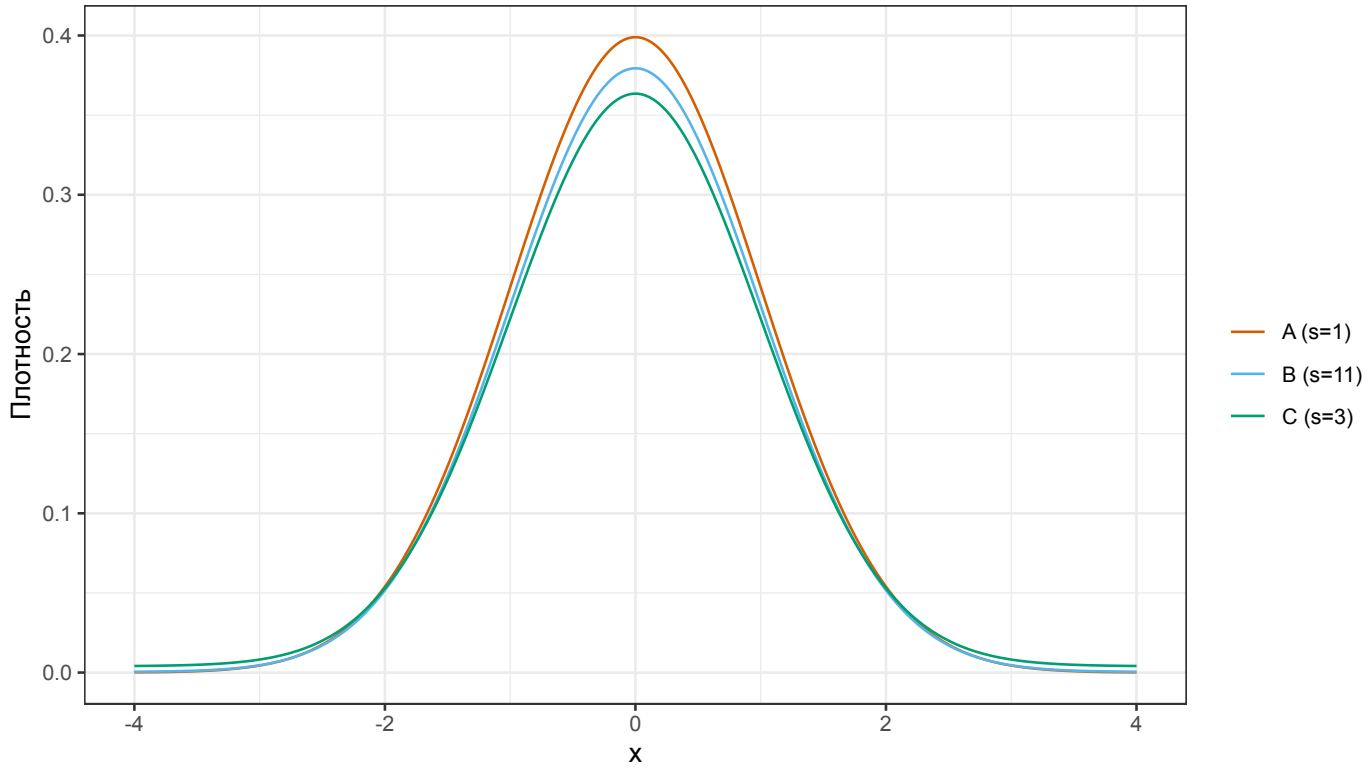


Обманчивое стандартное отклонение

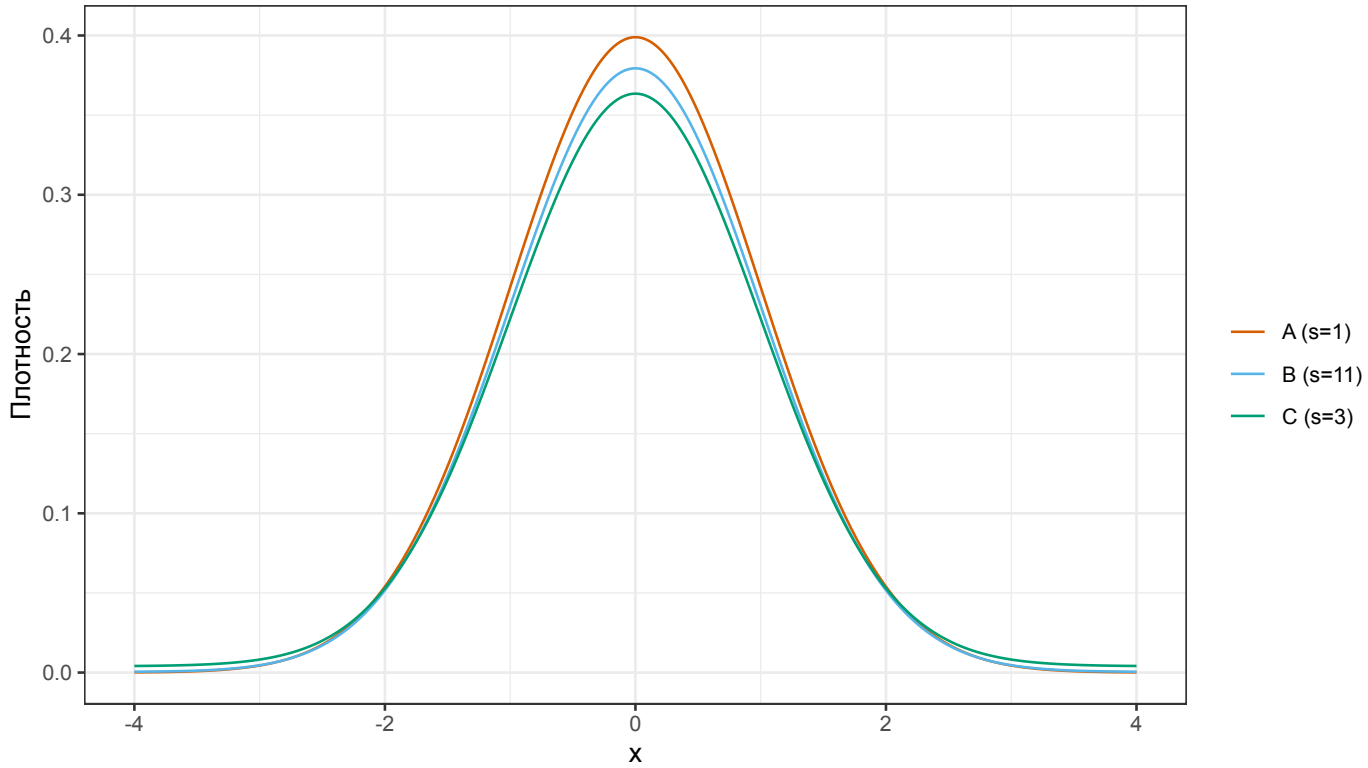
Обманчивое стандартное отклонение



Обманчивое стандартное отклонение

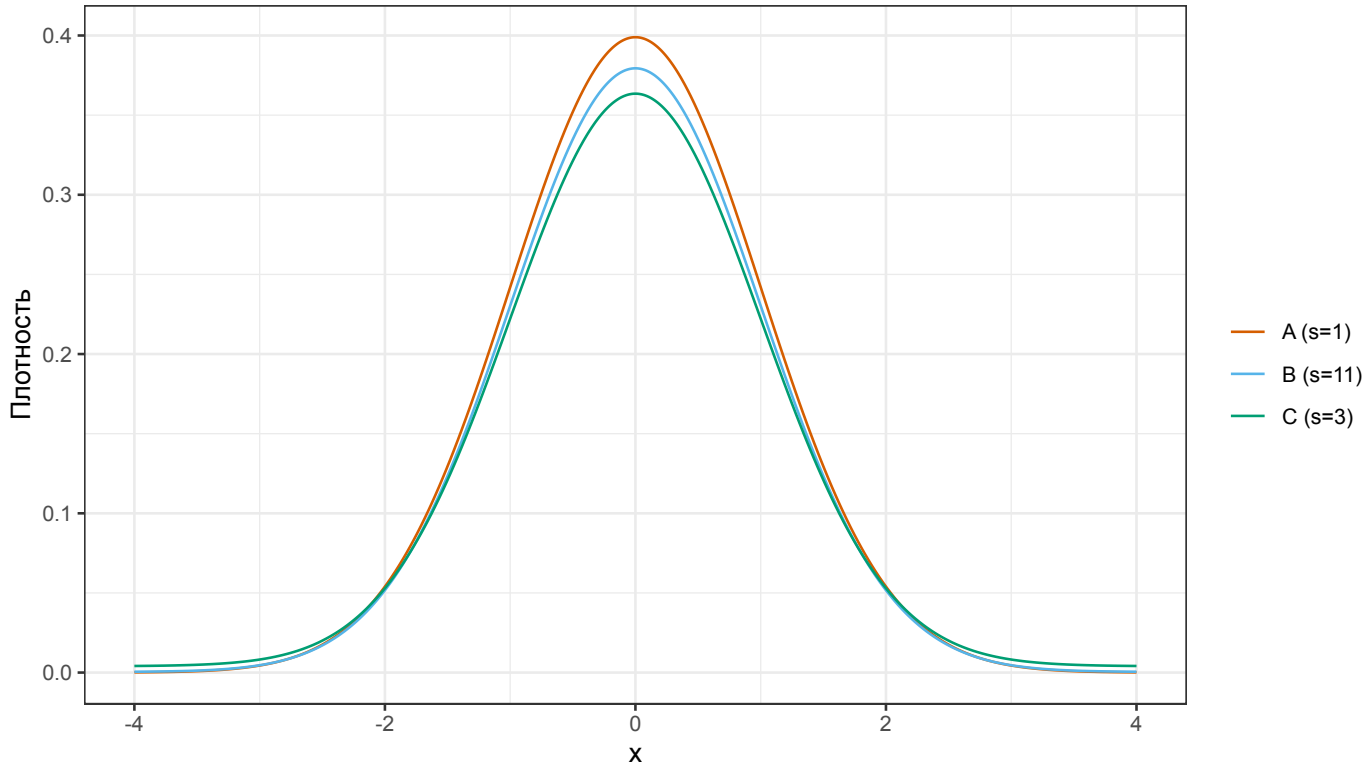


Обманчивое стандартное отклонение



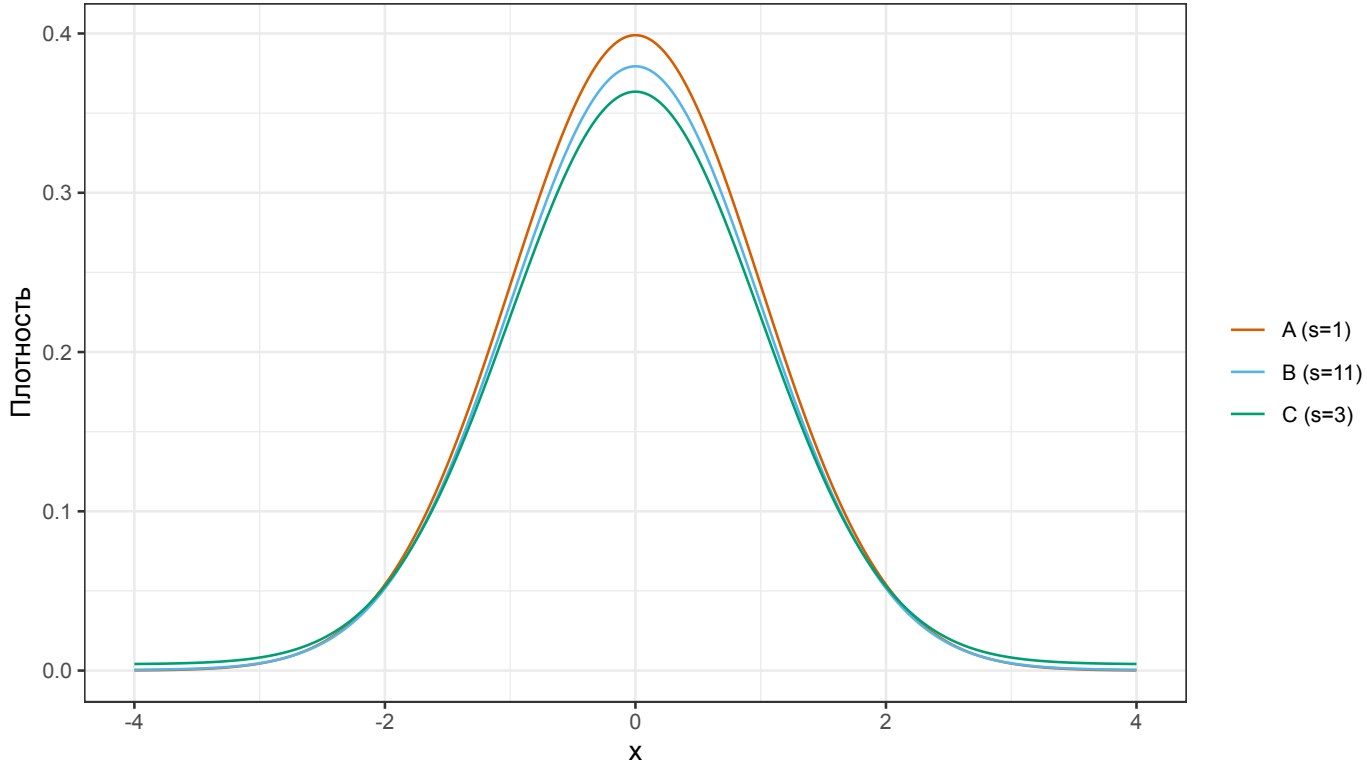
- $\mathrm{A} = \mathcal{N}(0,1^2); \sigma_A = 1$

Обманчивое стандартное отклонение



- $\mathrm{A} = \mathcal{N}(0,1^2); \sigma_A = 1$
- $\mathrm{B} = 0.95\mathcal{N}(0,1^2) + 0.05\mathcal{N}(0,49^2); \sigma_B = \sqrt{0.95 \cdot 1^2 + 0.05 \cdot 49^2} = 11$

Обманчивое стандартное отклонение



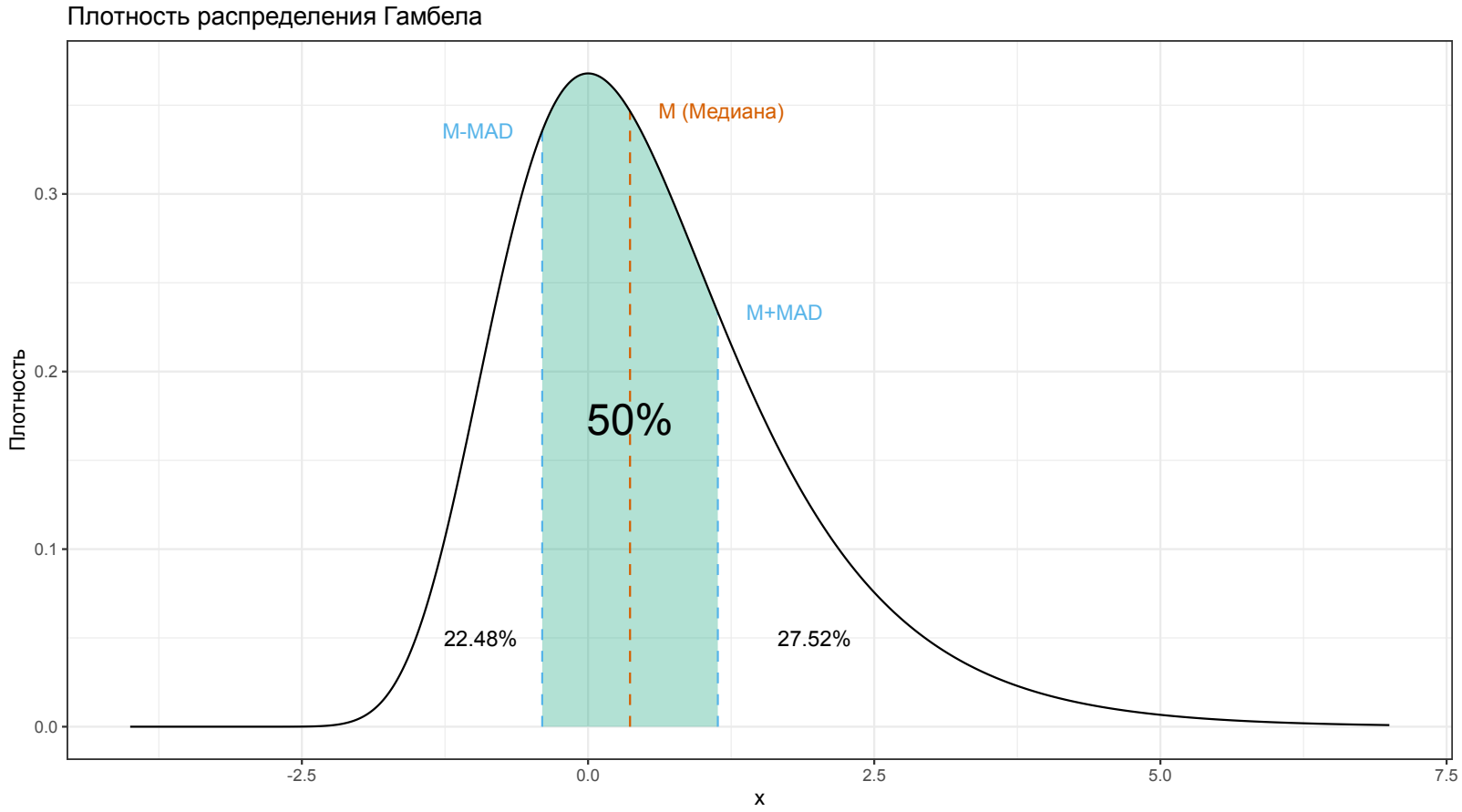
- $\mathrm{A} = \mathcal{N}(0,1^2); \sigma_A = 1$
- $\mathrm{B} = 0.95\mathcal{N}(0,1^2) + 0.05\mathcal{N}(0,49^2); \sigma_B = \sqrt{0.95 \cdot 1^2 + 0.05 \cdot 49^2} = 11$
- $\mathrm{C} = 0.9\mathcal{N}(0,1^2) + 0.1\mathcal{N}(0,9^2); \sigma_C = \sqrt{0.9 \cdot 1^2 + 0.1 \cdot 9^2} = 3$

Медианное абсолютное отклонение (MAD)

$$\mathbf{MAD} = C_n \cdot \mathbf{median}\Big(|x_i - \mathbf{median}(x)|\Big)$$

Медианное абсолютное отклонение (MAD)

$$\mathbf{MAD} = C_n \cdot \mathbf{median}\Big(|x_i - \mathbf{median}(x)|\Big)$$



https://aakinshin.net/posts/gumbel-mad/

$$\mathbf{MAD} = C_n \cdot \mathbf{median}\Big(|x_i - \mathbf{median}(x)|\Big)$$

$$\mathrm{MAD} = C_n \cdot \mathrm{median}\Big(|x_i - \mathrm{median}(x)|\Big)$$

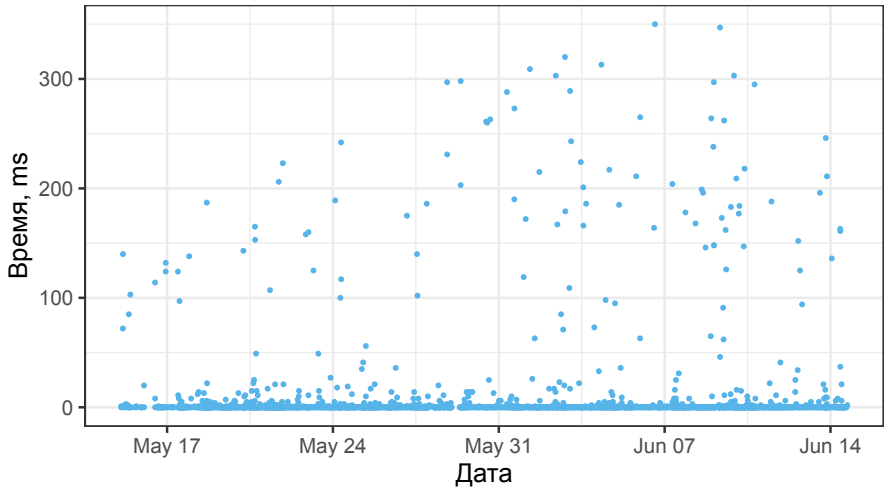
$$C_\infty = \frac{1}{\Phi^{-1}(3/4)} \approx 1.4826022185056$$

$$\mathrm{MAD} = C_n \cdot \mathrm{median}\Big(|x_i - \mathrm{median}(x)|\Big)$$

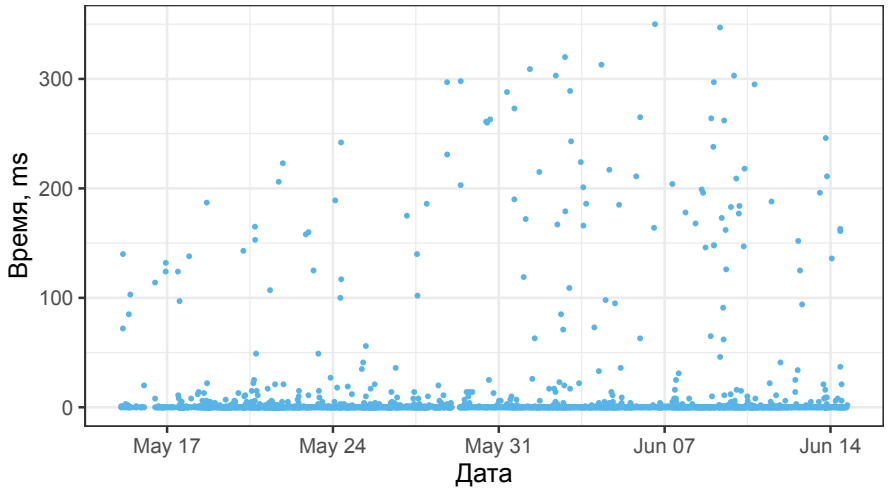
$$C_\infty = \frac{1}{\Phi^{-1}(3/4)} \approx 1.4826022185056$$

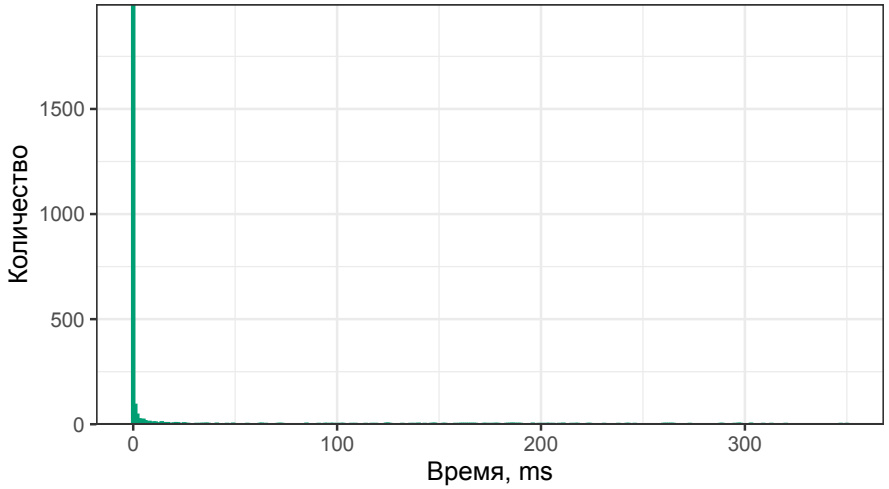
n	C_n (HF7)	C_n (HD)
2	1.77215	1.77215
3	2.20490	1.56816
4	2.01667	1.59589
5	1.80392	1.56611
6	1.76378	1.56656
...	...	...

Andrey Akinshin (2022) “Finite-sample bias-correction factors for the median absolute deviation based on the Harrell-Davis quantile estimator and its trimmed modification,” arXiv:2207.12005


300
200
100
0
Время, ms
May 17
May 24
May 31
Jun 07
Jun 14
Дата

Нулевые медиана/MAD





https://aakinshin.net/posts/discrete-performance-distributions/

```
// Непрерывные распределения
x = { 2174, 1984, 2746, 1596, 1874, 2812, 2245,... } ms
y = { 1874, 2207, 1984, 1729, 2301, 1781, 2022,... } ms
```

```
// Непрерывные распределения
x = { 2174, 1984, 2746, 1596, 1874, 2812, 2245,... } ms
y = { 1874, 2207, 1984, 1729, 2301, 1781, 2022,... } ms
```

```
// Дискретные распределения
y = { 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6 } ms
z = { 6, 7, 7, 6, 6, 6, 7, 6, 6, 7, 7 } ms
```

Квантильное абсолютное отклонение (QAD)

Медианное абсолютное отклонение около медианы:

$$\mathbf{MAD}(x) = C_n \cdot \mathbf{median}\Big(|x_i - \mathbf{median}(x)|\Big)$$

Медианное абсолютное отклонение около медианы:

$$\mathrm{MAD}(x) = C_n \cdot \mathrm{median}\Big(|x_i - \mathrm{median}(x)|\Big)$$

Квантильное абсолютное отклонение около медианы:

$$\mathrm{QAD}(x, p) = C_n \cdot Q\Big(|x_i - \mathrm{median}(x)|, p\Big)$$

Медианное абсолютное отклонение около медианы:

$$\mathrm{MAD}(x) = C_n \cdot \mathrm{median}\Big(|x_i - \mathrm{median}(x)|\Big)$$

Квантильное абсолютное отклонение около медианы:

$$\mathrm{QAD}(x, p) = C_n \cdot Q\Big(|x_i - \mathrm{median}(x)|, p\Big)$$

Оценка	p		
MAD	0.50		
Standard QAD (SQAD)	0.68		
Optimal QAD (OQAD)	0.86		

Andrey Akinshin (2022) “Quantile absolute deviation,” arXiv:2208.13459

Квантильное абсолютное отклонение (QAD)

Медианное абсолютное отклонение около медианы:

$$\mathrm{MAD}(x) = C_n \cdot \mathrm{median}\Big(|x_i - \mathrm{median}(x)|\Big)$$

Квантильное абсолютное отклонение около медианы:

$$\mathrm{QAD}(x, p) = C_n \cdot Q\Big(|x_i - \mathrm{median}(x)|, p\Big)$$

Оценка	p	Точка перелома	
MAD	0.50	50%	
Sandard QAD (SQAD)	0.68	32%	
Optimal QAD (OQAD)	0.86	14%	

Andrey Akinshin (2022) “Quantile absolute deviation,” arXiv:2208.13459

Квантильное абсолютное отклонение (QAD)

Медианное абсолютное отклонение около медианы:

$$\mathrm{MAD}(x) = C_n \cdot \mathrm{median}\Big(|x_i - \mathrm{median}(x)|\Big)$$

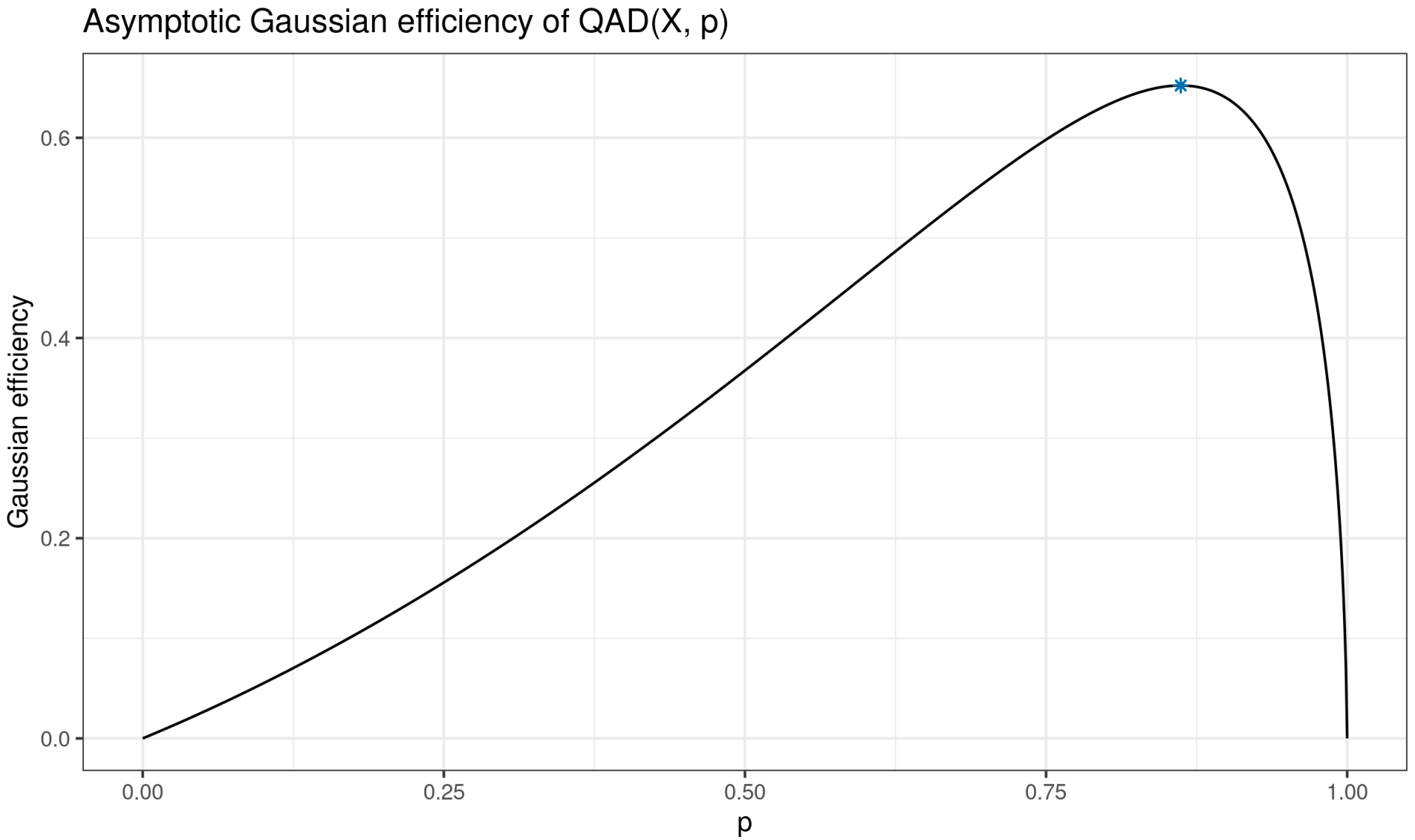
Квантильное абсолютное отклонение около медианы:

$$\mathrm{QAD}(x, p) = C_n \cdot Q\Big(|x_i - \mathrm{median}(x)|, p\Big)$$

Оценка	p	Точка перелома	Гауссова эффективность
MAD	0.50	50%	37%
Sandard QAD (SQAD)	0.68	32%	54%
Optimal QAD (OQAD)	0.86	14%	65%

Andrey Akinshin (2022) “Quantile absolute deviation,” arXiv:2208.13459

Квантильное абсолютное отклонение (QAD)



Andrey Akinshin (2022) “Quantile absolute deviation,” arXiv:2208.13459

Прочие оценки стандартного отклонения

Оценка Шамоса

$$\mathrm{Shamos} = C_n \cdot \mathrm{median}(|x_i - x_j|_{i<j}); \quad C_\infty \approx 1.048358$$

Shamos, M. I. "Geometry and statistics: Problems at the interface. V: JF Traub (Ed.): Algorithms and complexity: New directions and recent results." (1976).

Оценка Шамоса

$$\mathrm{Shamos} = C_n \cdot \mathrm{median}(|x_i - x_j|_{i<j}); \quad C_\infty \approx 1.048358$$

Оценка Руссо-Крукса

$$S_n = C_n \cdot \mathrm{median}_i\Big(\mathrm{median}_j\big(|x_i - x_j|\big)\Big); \quad C_\infty \approx 1.1926$$

$$Q_n = C_n \cdot Q(|x_i - x_j|_{i<j}, 0.25); \quad C_\infty \approx 2.2191$$

Shamos, M. I. "Geometry and statistics: Problems at the interface. V: JF Traub (Ed.): Algorithms and complexity: New directions and recent results." (1976).

Rousseeuw, P., and C. Croux. 1993. Alternatives to the median absolute deviation. Journal of the American Statistical Association 88 (424):1273–83. doi:10.2307/2291267.

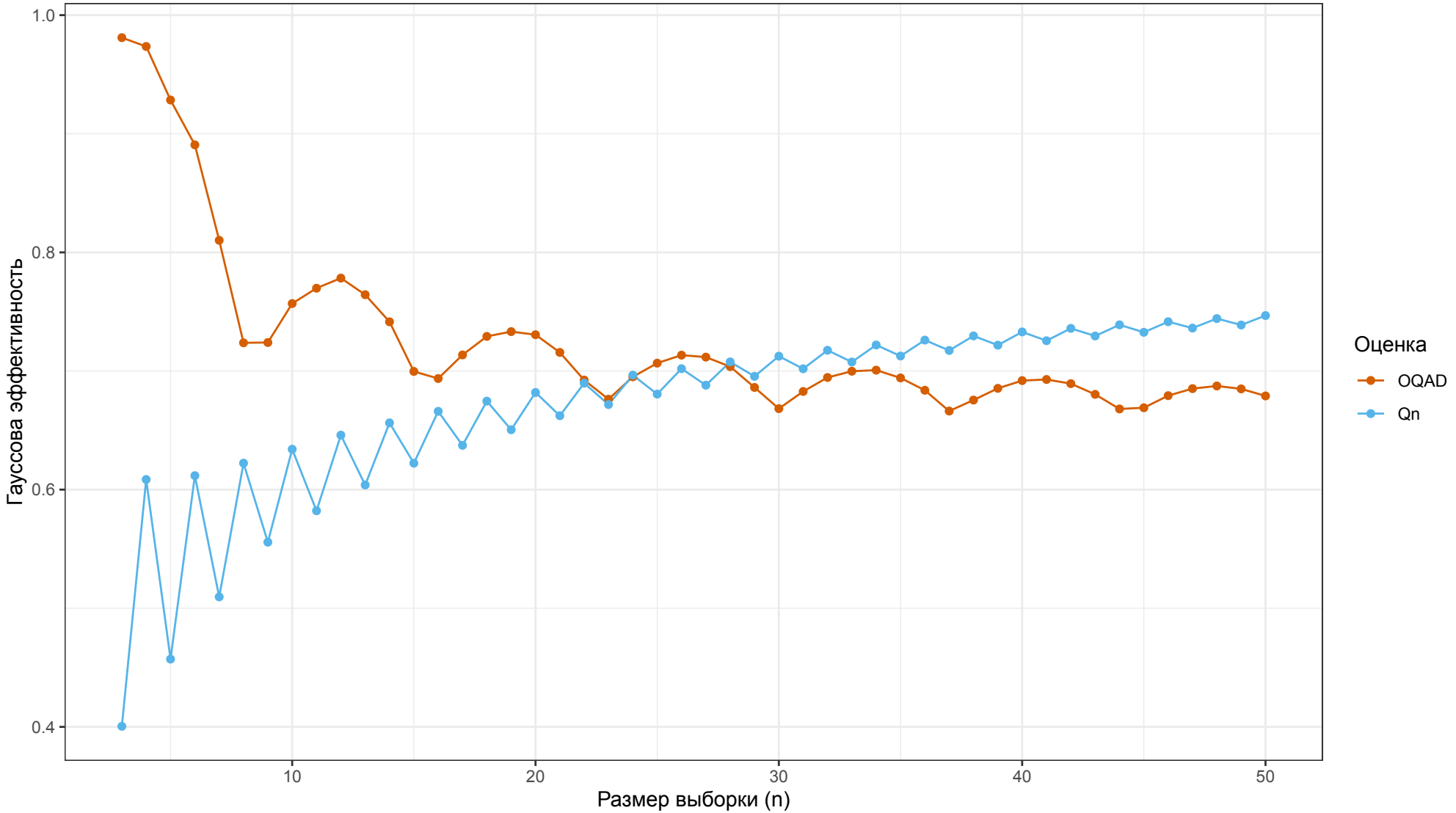
Сравнение асимптотических оценок стандартных отклонений

Сравнение асимптотических оценок стандартных отклонений

	SD	**MAD**	**SQAD**	**QQAD**	**Shamos**	S_n	Q_n
Гауссова эффективность	100%	37%	54%	65%	86%	58%	**82%**
Точка перелома	0%	50%	32%	14%	29%	50%	**50%**

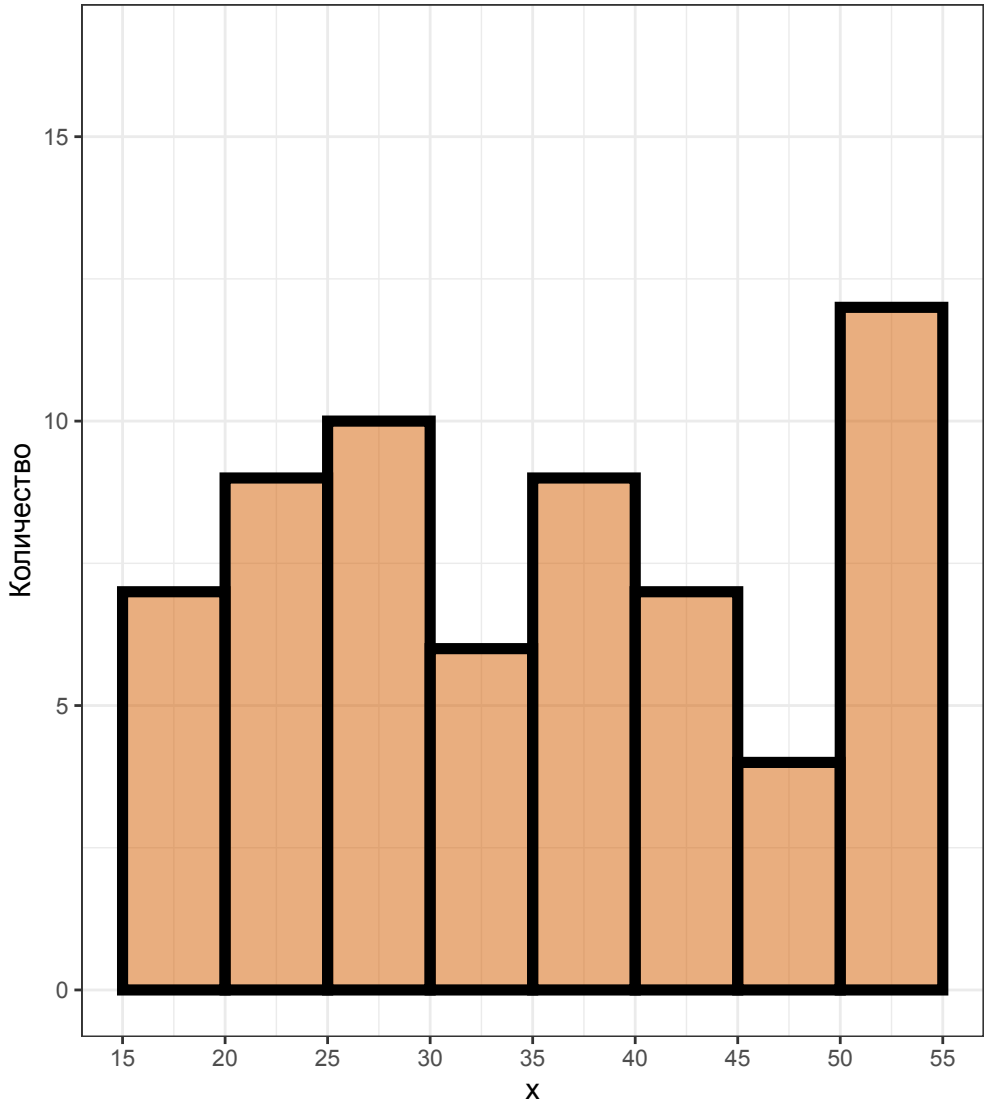
Сравнение **асимптотических** оценок стандартных отклонений

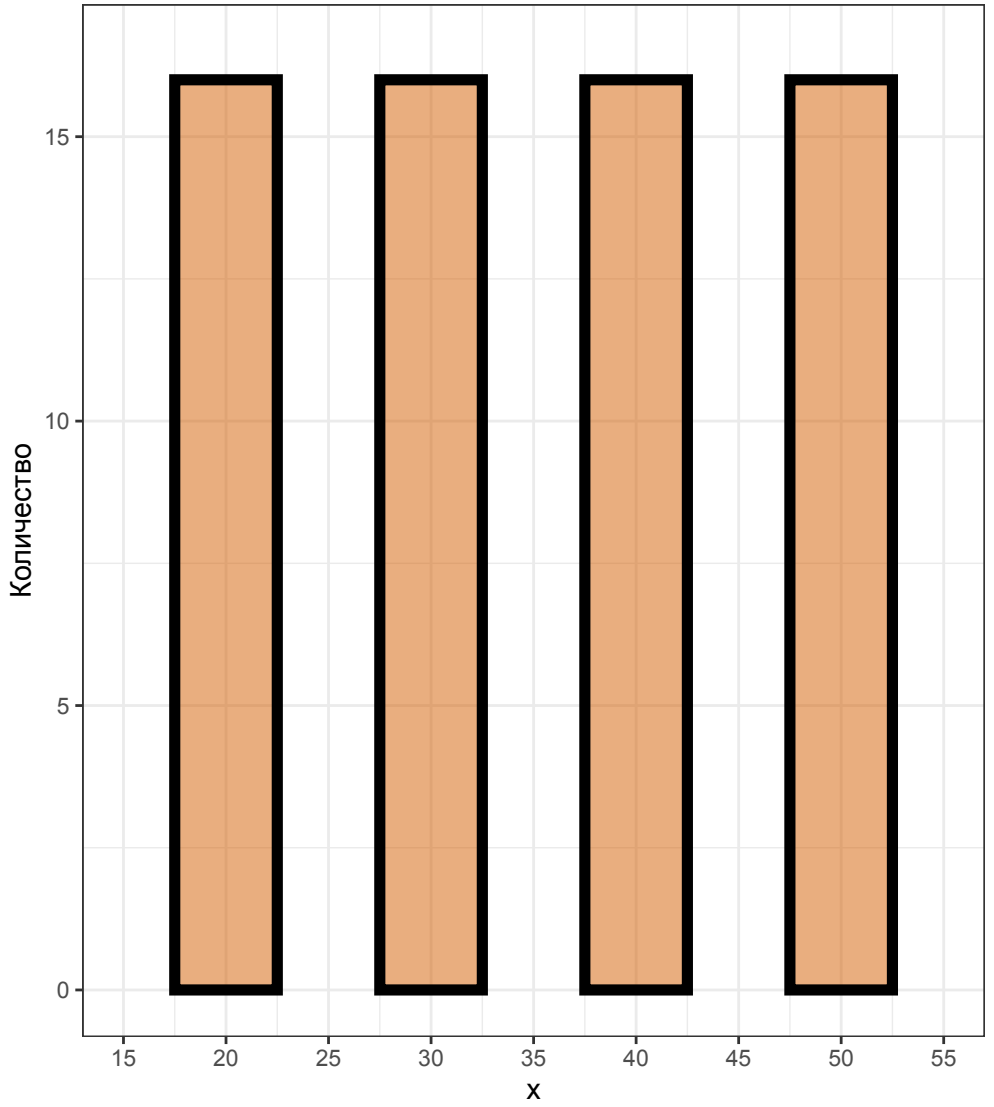
	SD	**MAD**	**SQAD**	**QQAD**	**Shamos**	S_n	Q_n
Гауссова эффективность	100%	37%	54%	65%	86%	58%	**82%**
Точка перелома	0%	50%	32%	14%	29%	50%	**50%**

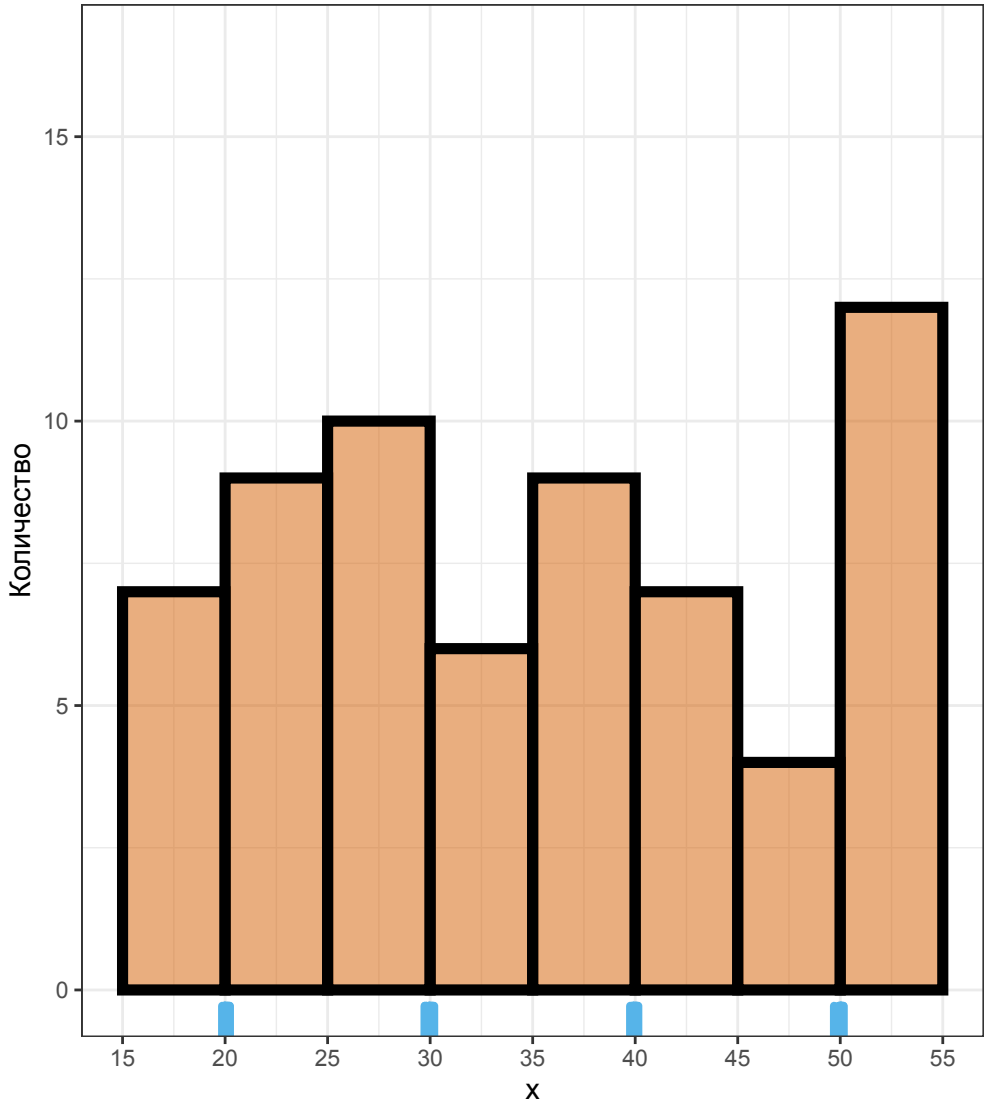

1.0
0.8
0.6
0.4
Гауссова эффективность
10
20
30
40
50
Размер выборки (n)
Оценка
OQAD
Qn

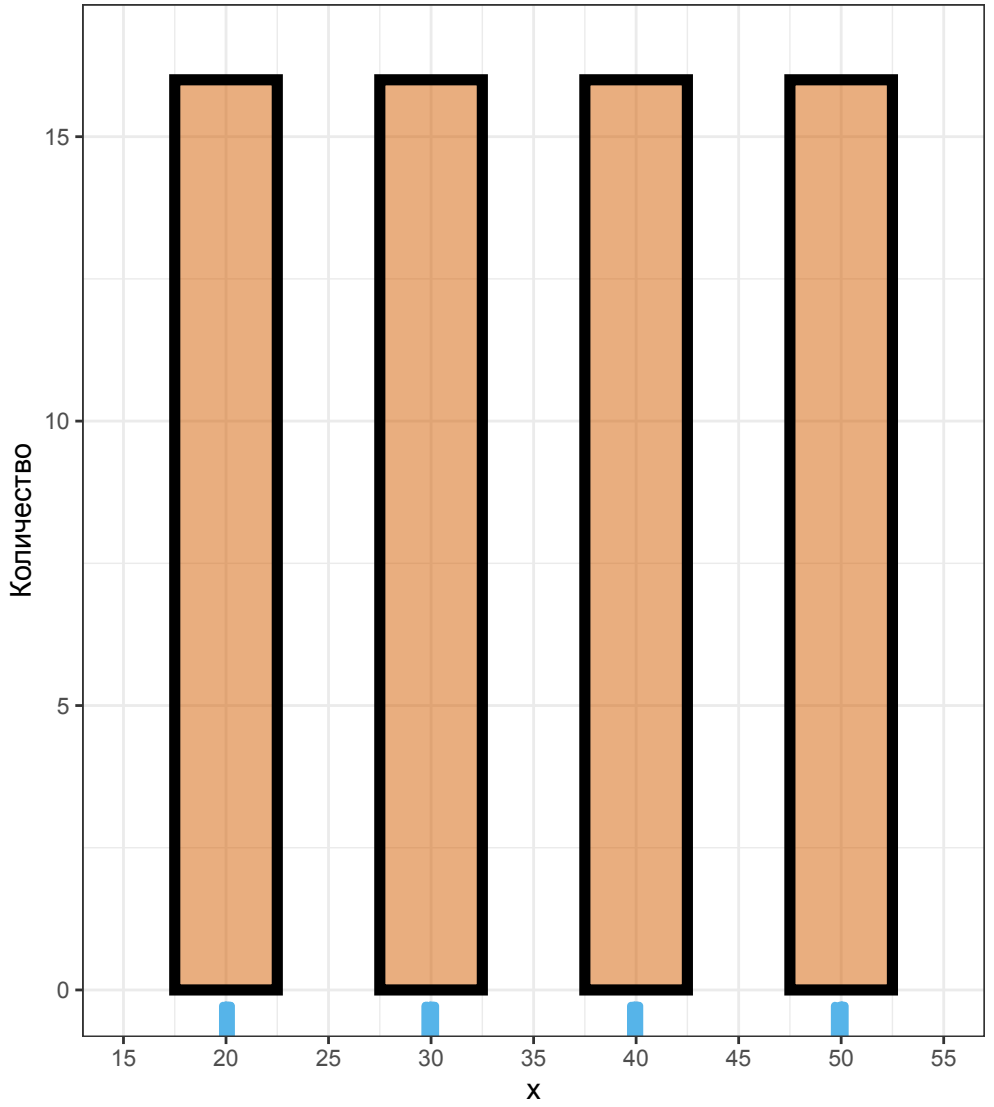
Для выбора метрики вариации
нужно подумать

Плотность распределения

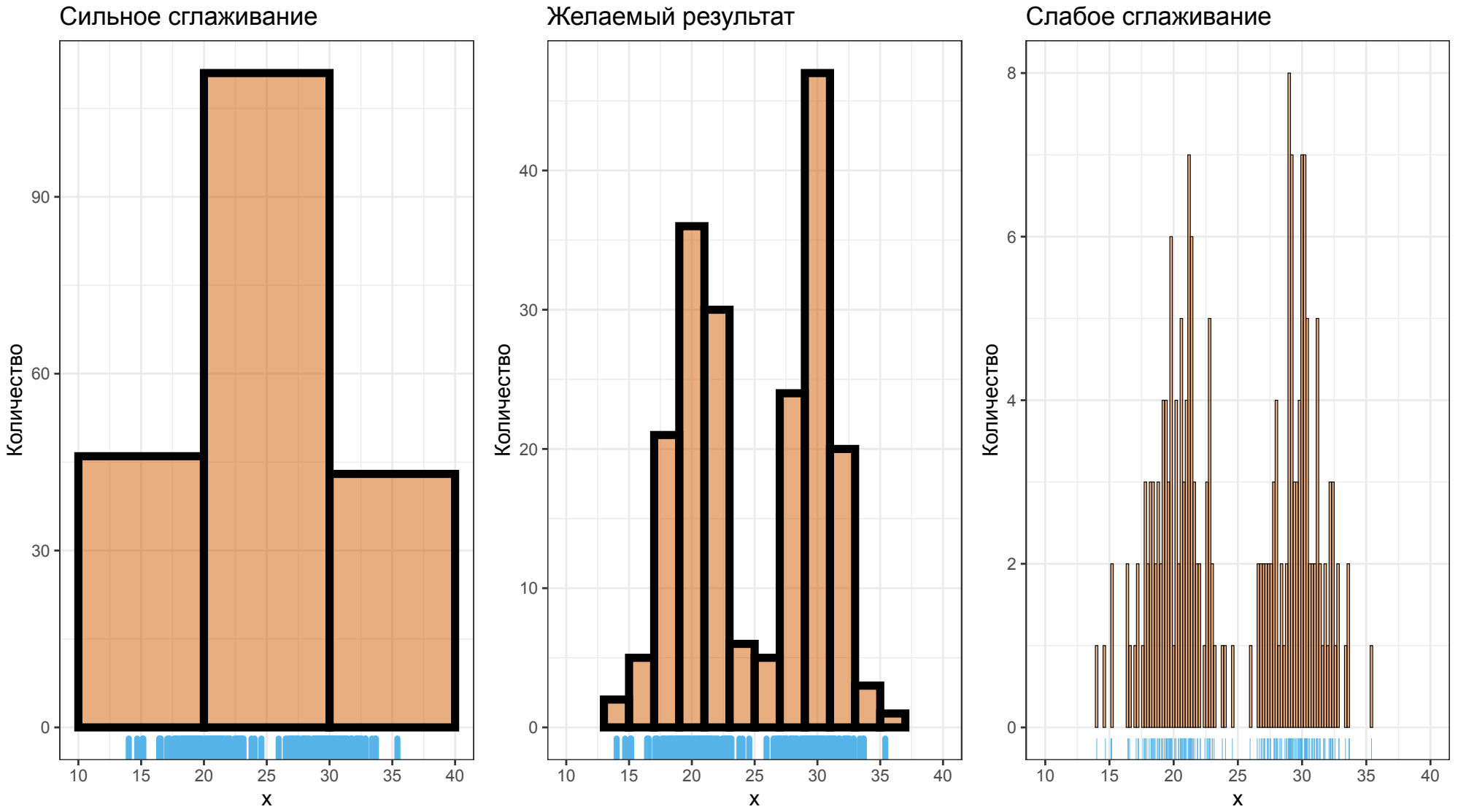

Количество
15
10
5
0
15
20
25
30
35
40
45
50
55
x


Количество
15
10
5
0
15
20
25
30
35
40
45
50
55
x



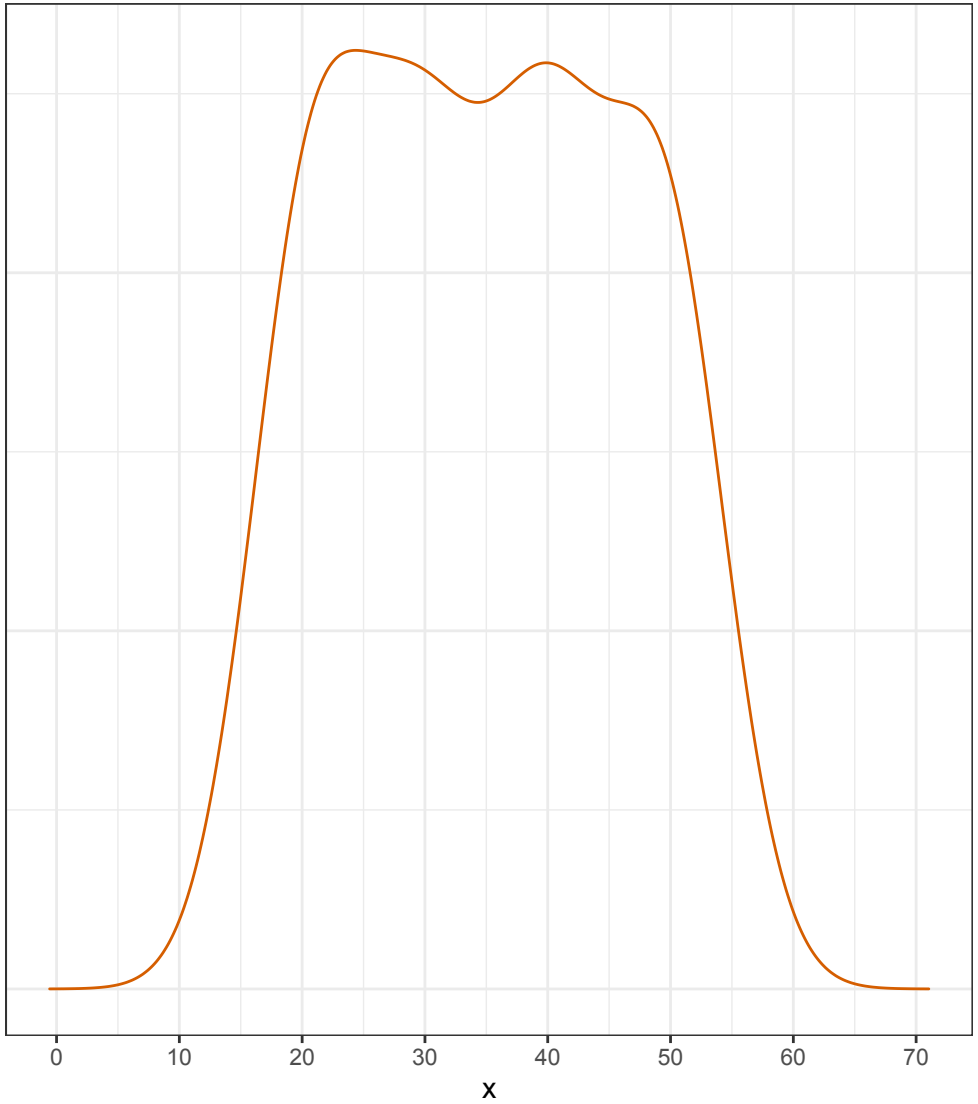


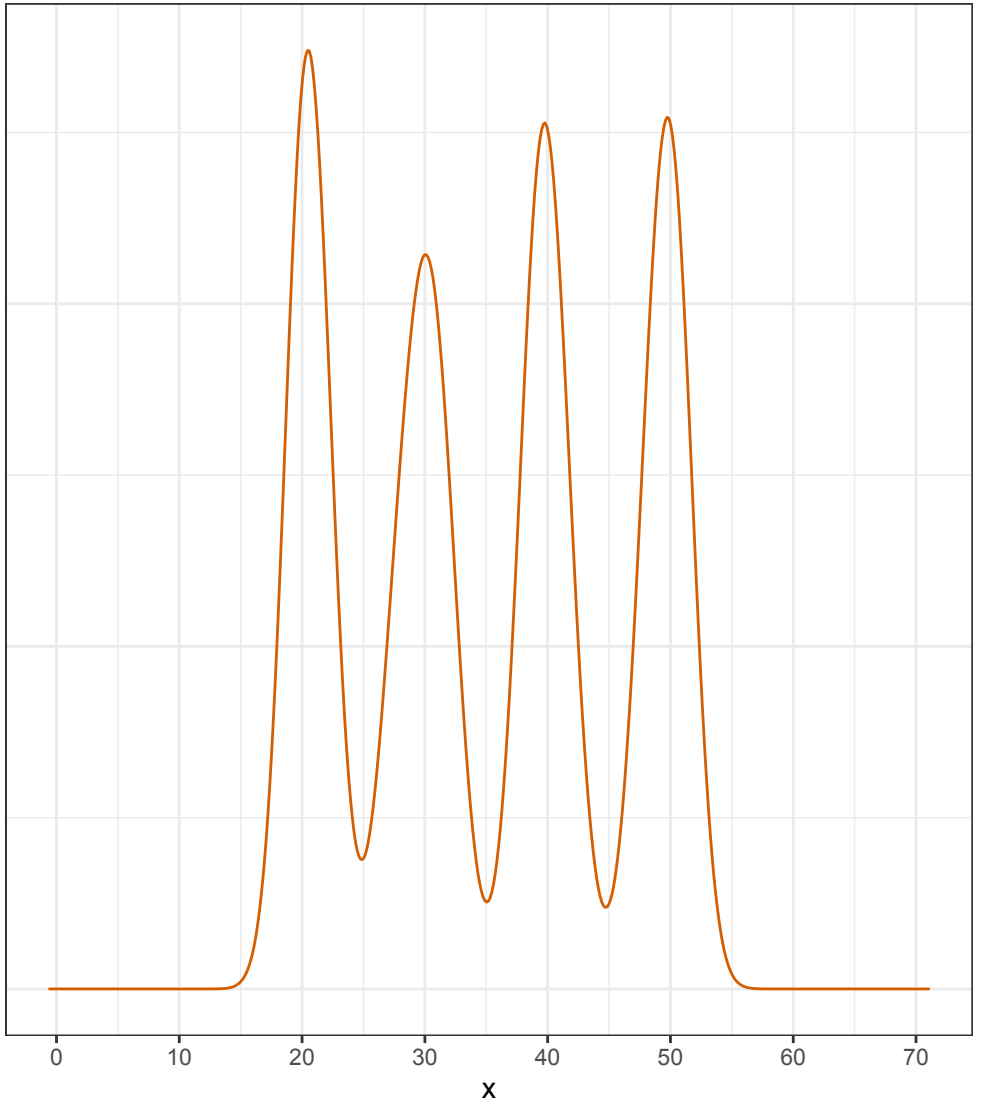
https://aakinshin.net/posts/misleading-histograms/



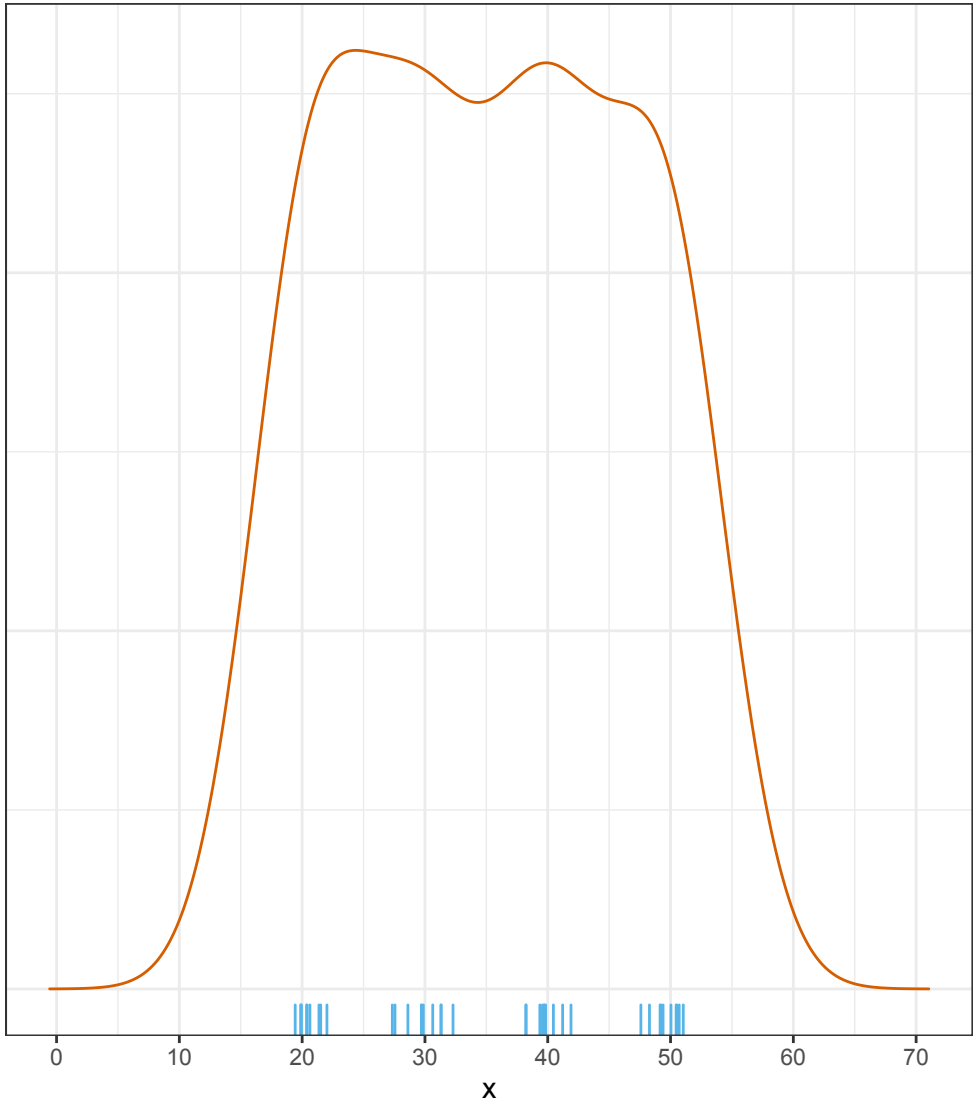
https://aakinshin.net/posts/misleading-histograms/

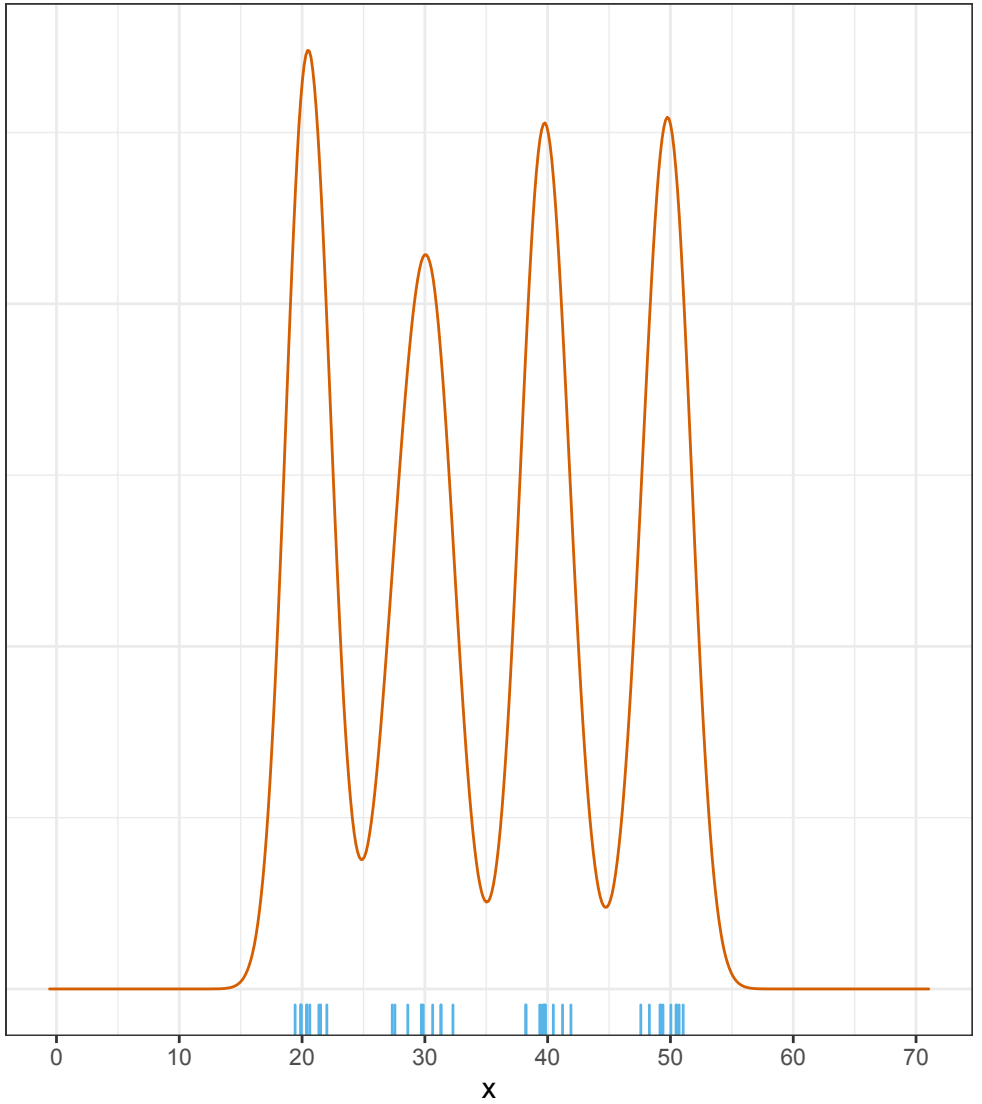
Ядерная оценка плотности (KDE)





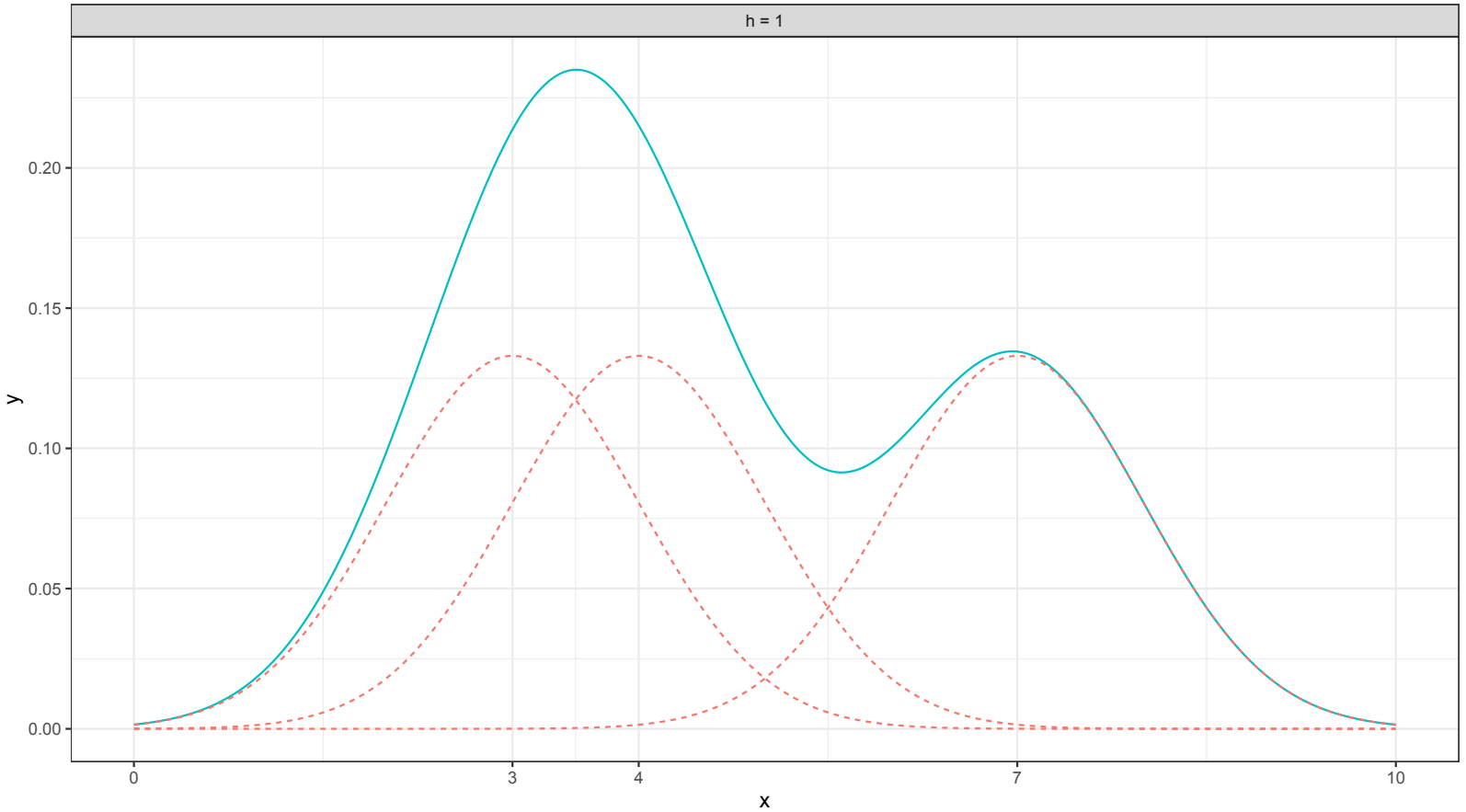
Ядерная оценка плотности (KDE)





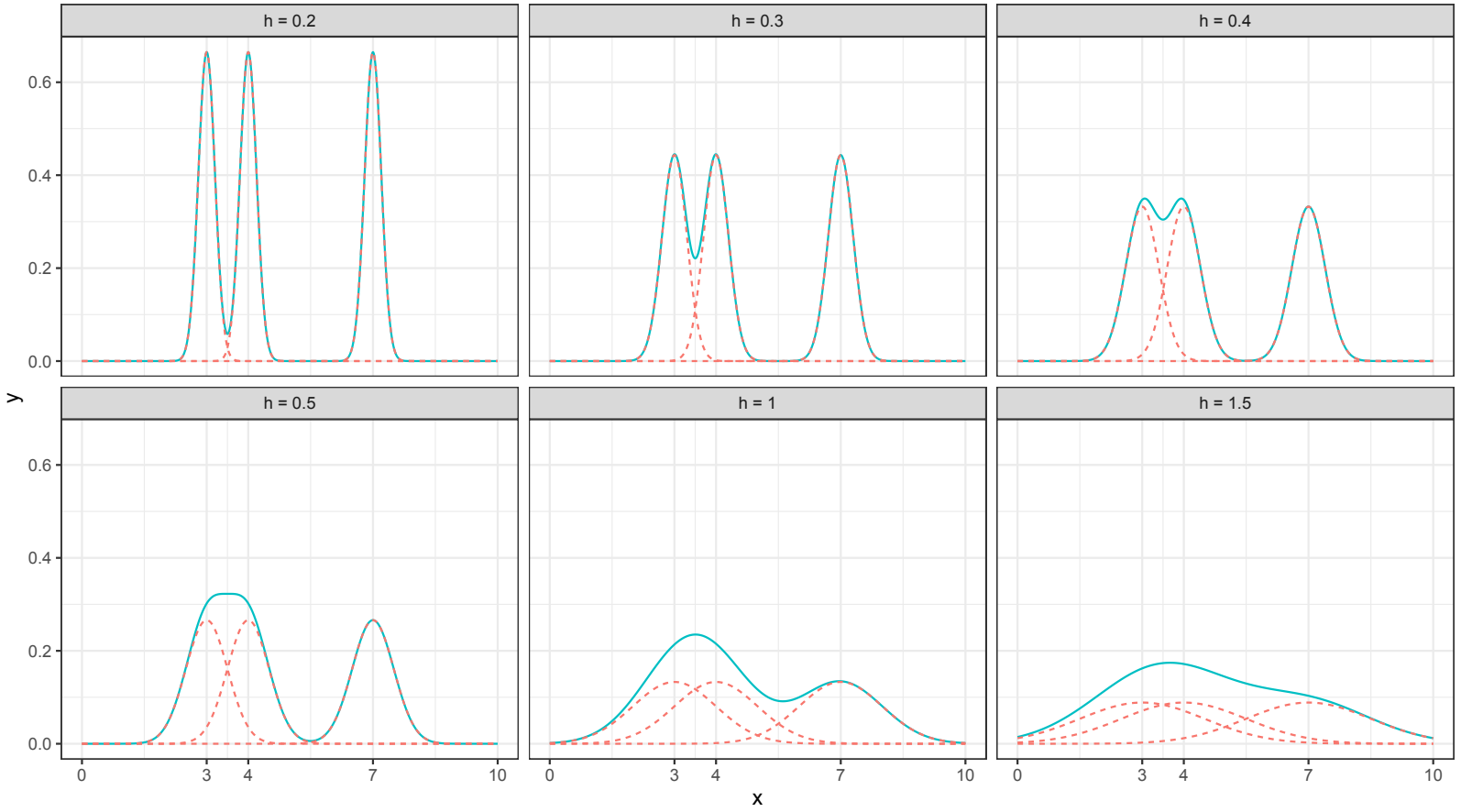
$$\hat{f}_{h(x)} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

$$\hat{f}_{h(x)} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$



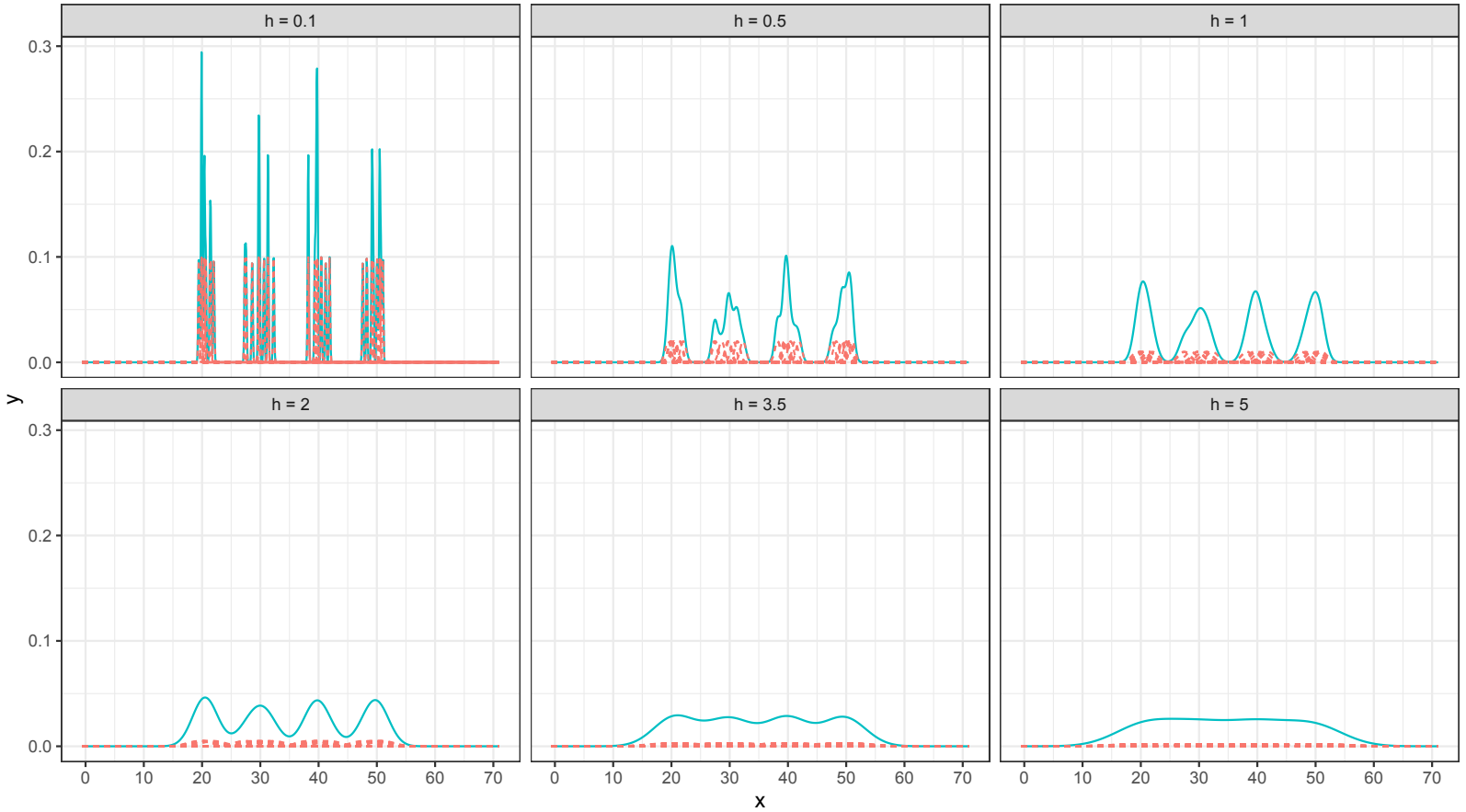
https://aakinshin.net/posts/kde-bw/

$$\hat{f}_{h(x)} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$



https://aakinshin.net/posts/kde-bw/

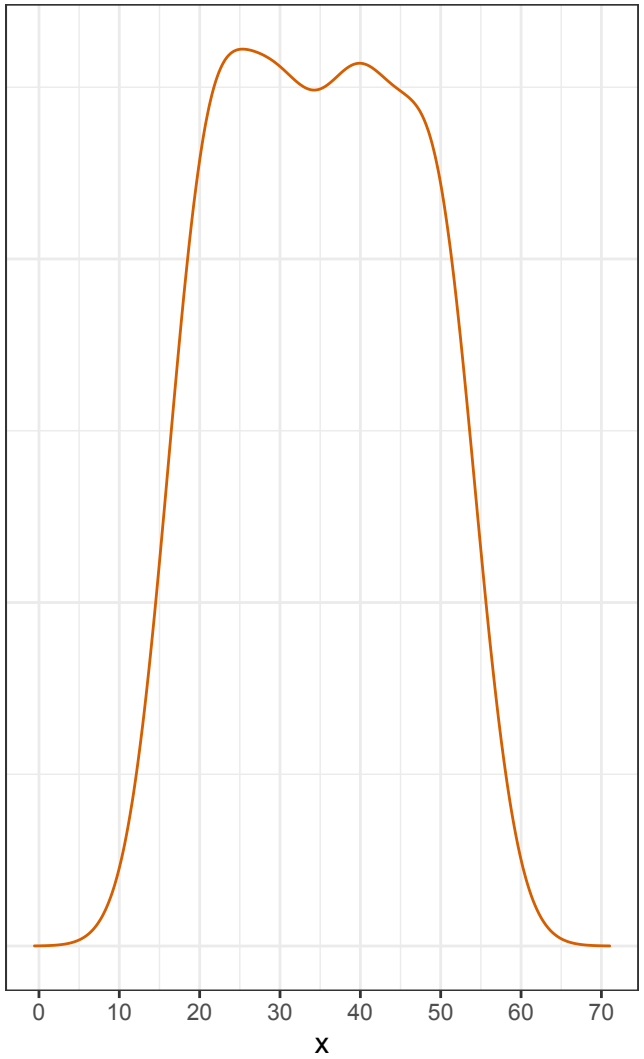
$$\hat{f}_{h(x)} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$



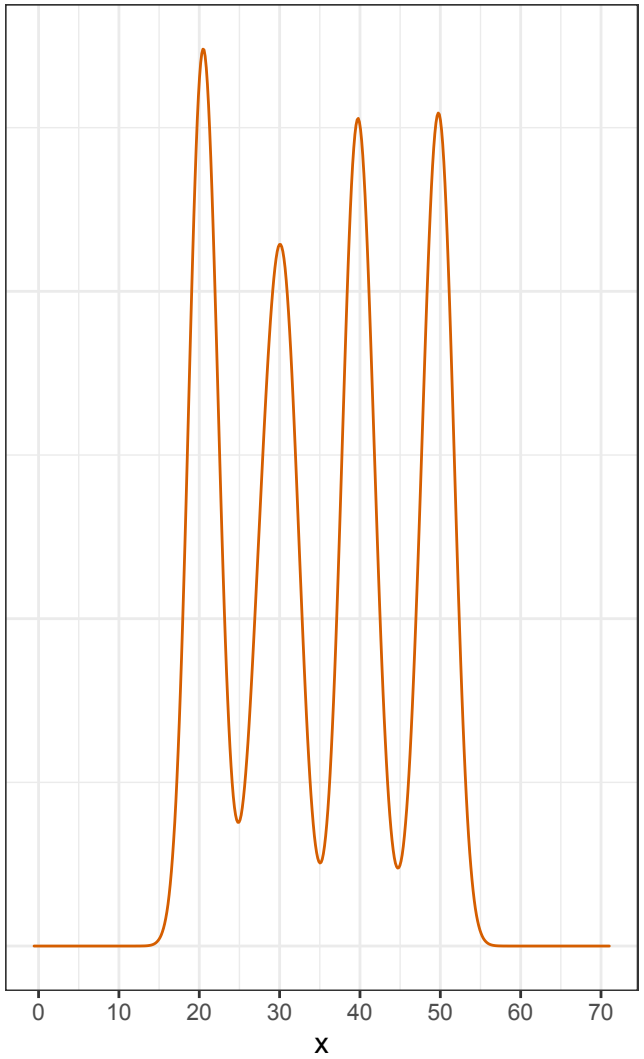
https://aakinshin.net/posts/kde-bw/

Ядерная оценка плотности (KDE)

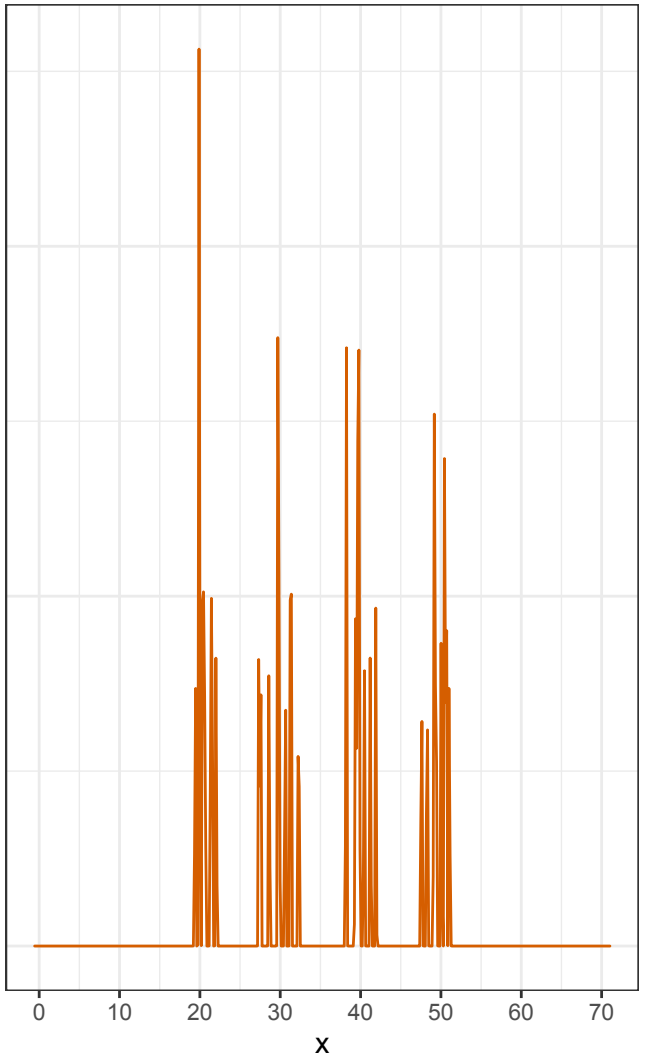
Сильное сглаживание



Желаемый результат

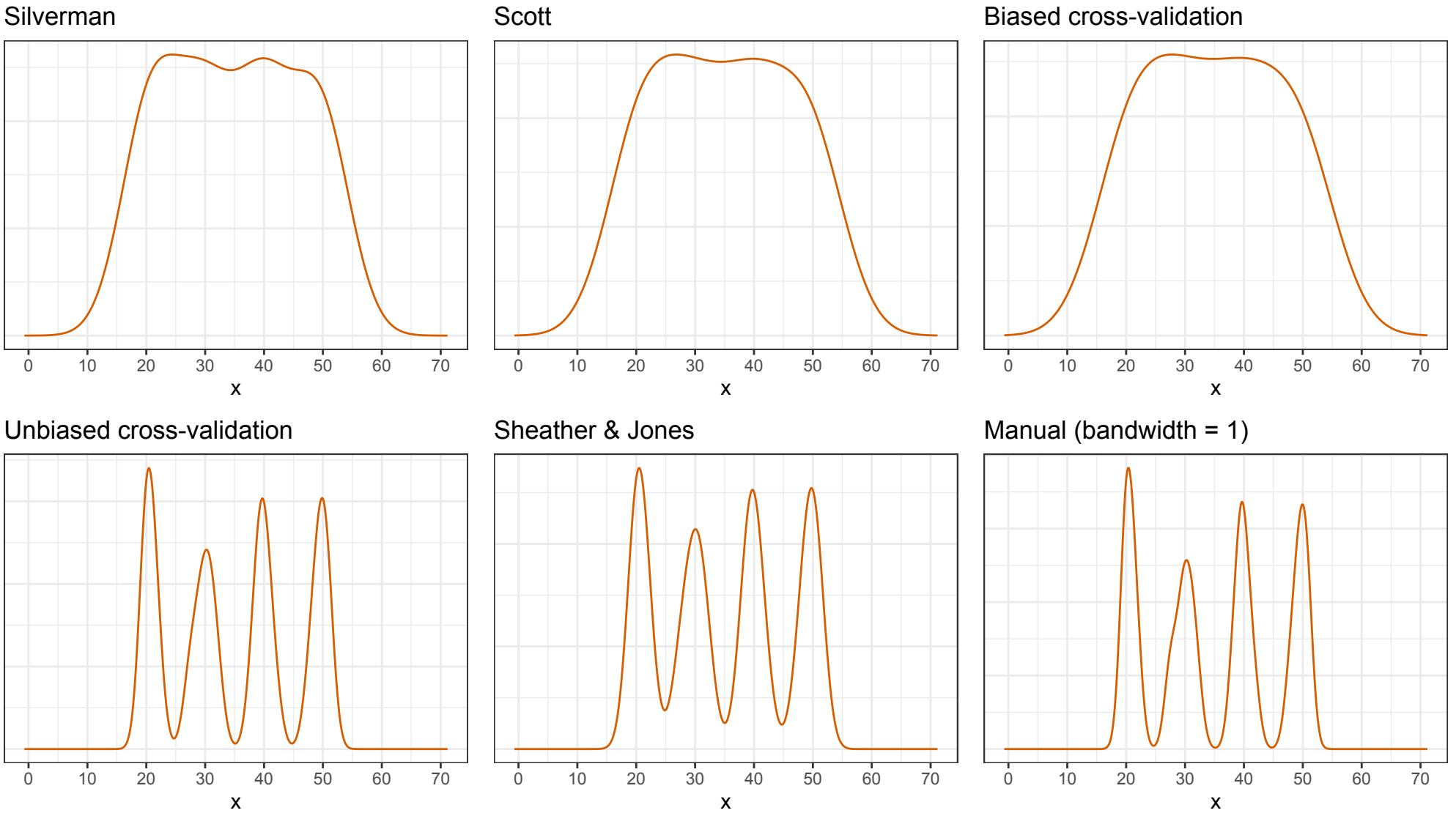


Слабое сглаживание

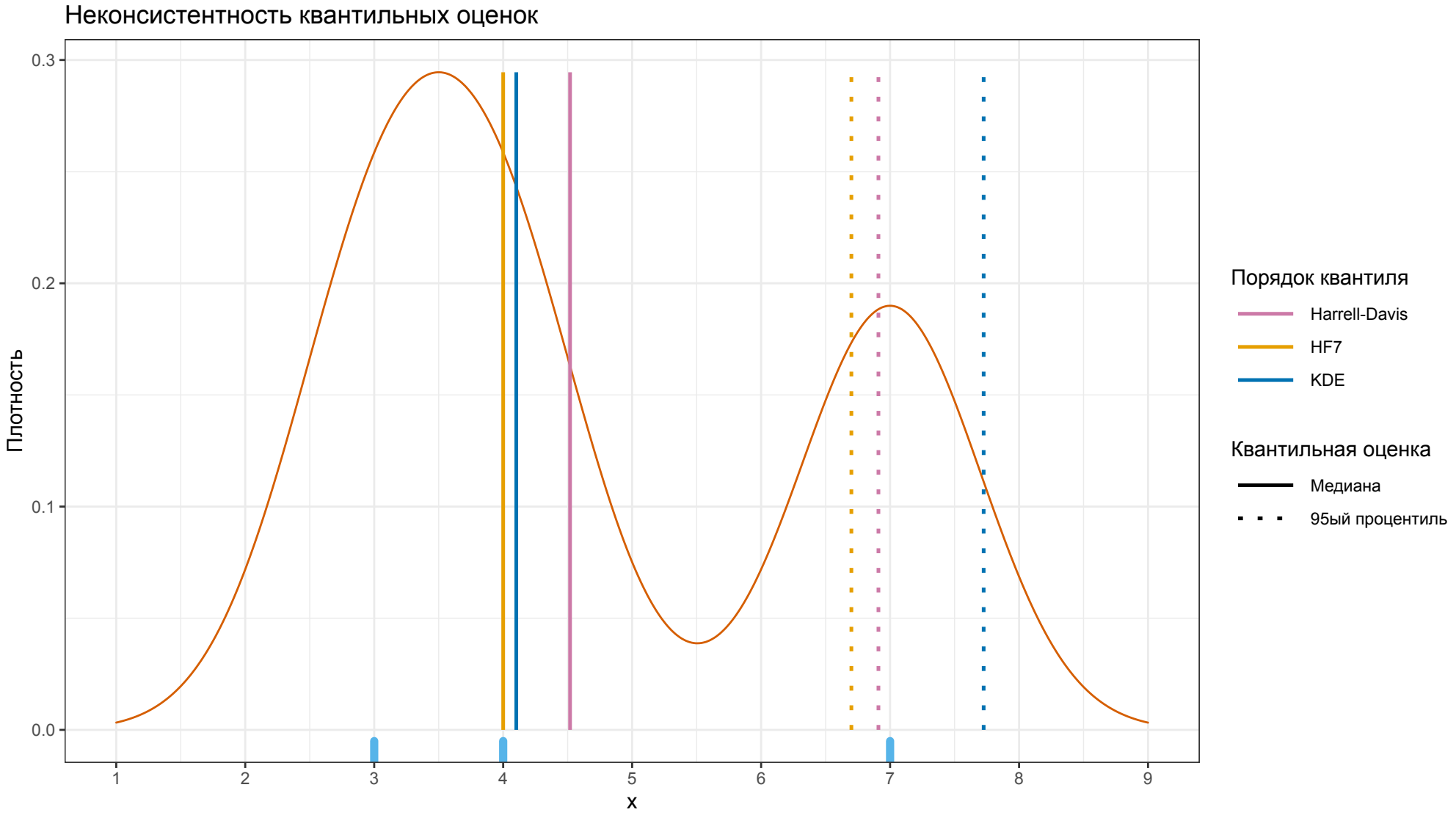


https://aakinshin.net/posts/kde-bw/

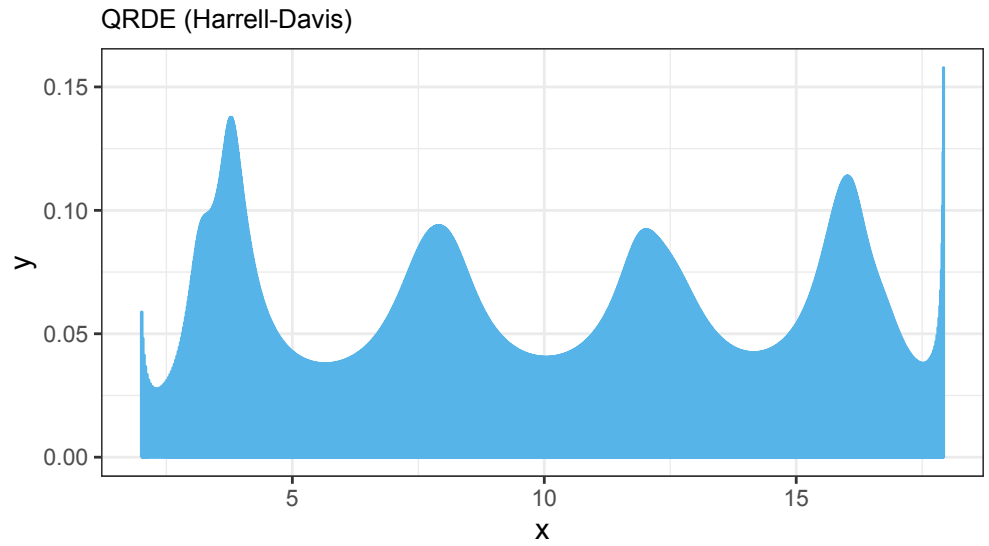
Ядерная оценка плотности (KDE)

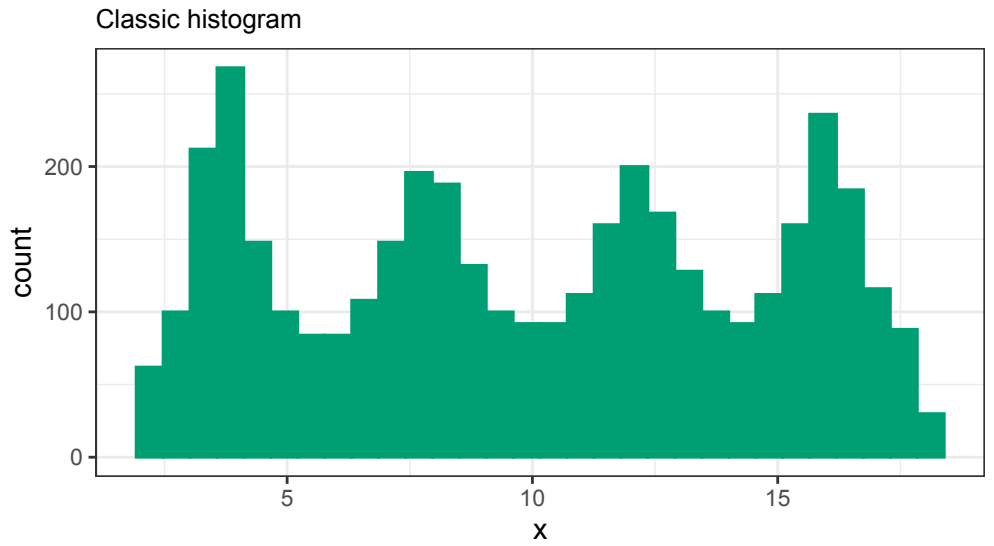


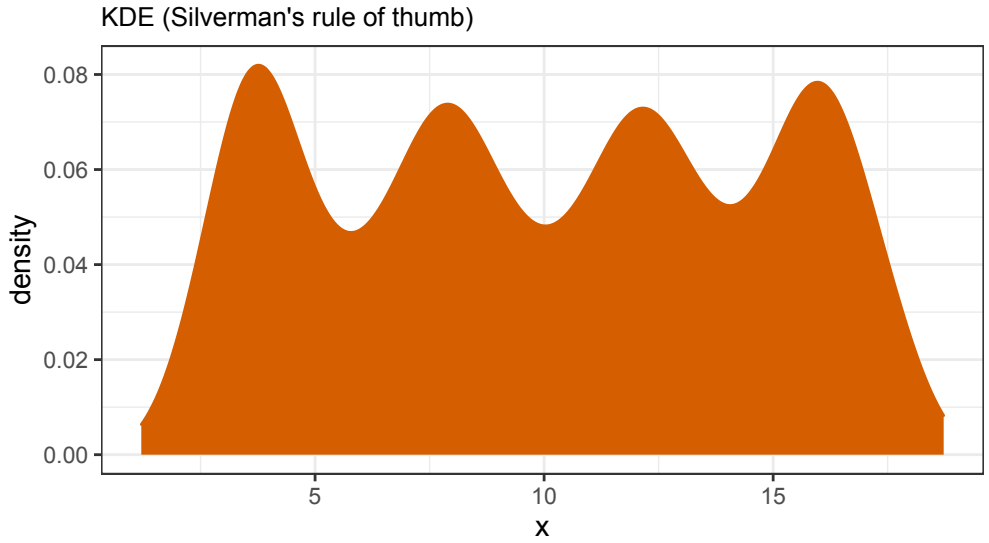
https://aakinshin.net/posts/kde-bw/

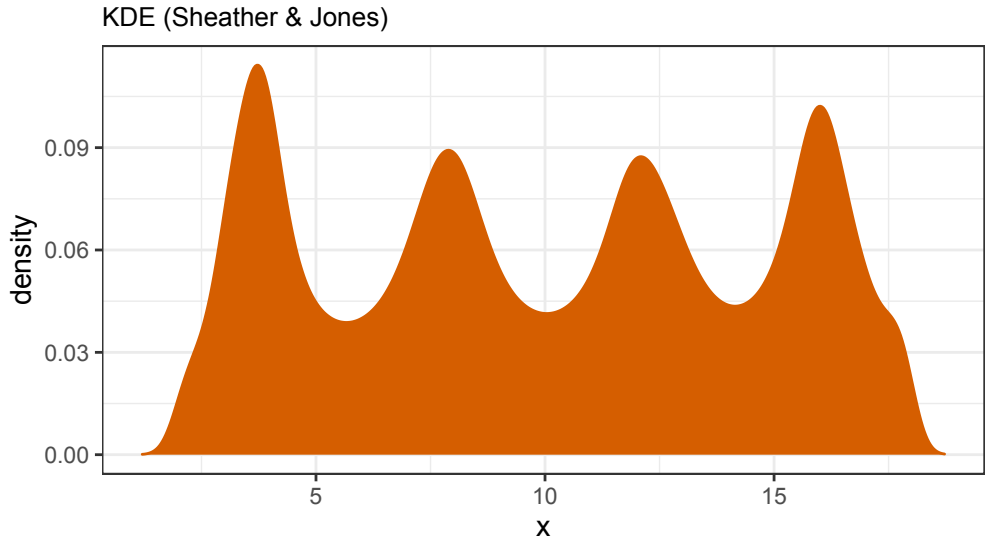

Неконсистентность квантильных оценок
Плотность
0.0
0.1
0.2
0.3
1
2
3
4
5
6
7
8
9
x
Порядок квантиля
Harrell-Davis
HF7
KDE
Квантильная оценка
Медиана
95ый процентиль

Квантильная оценка плотности (QRDE)



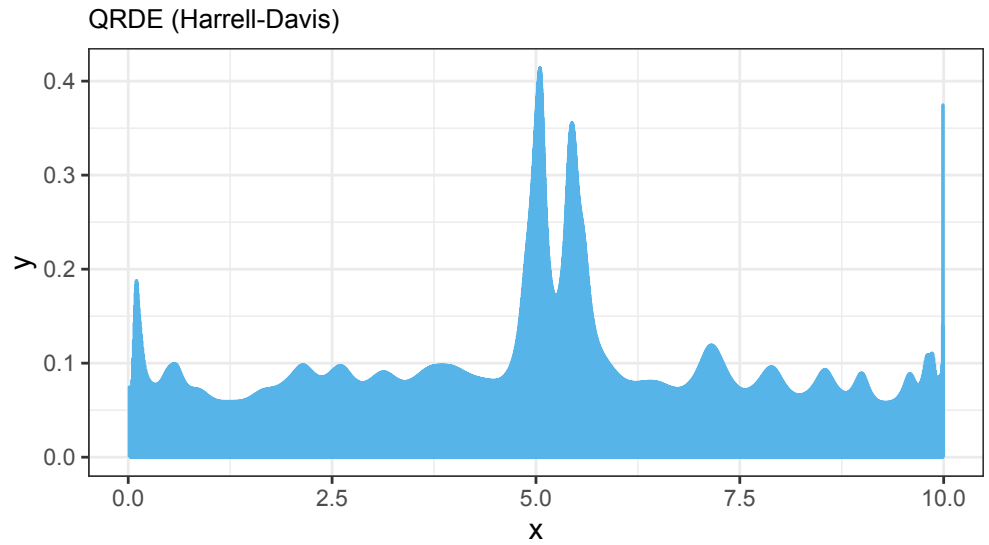


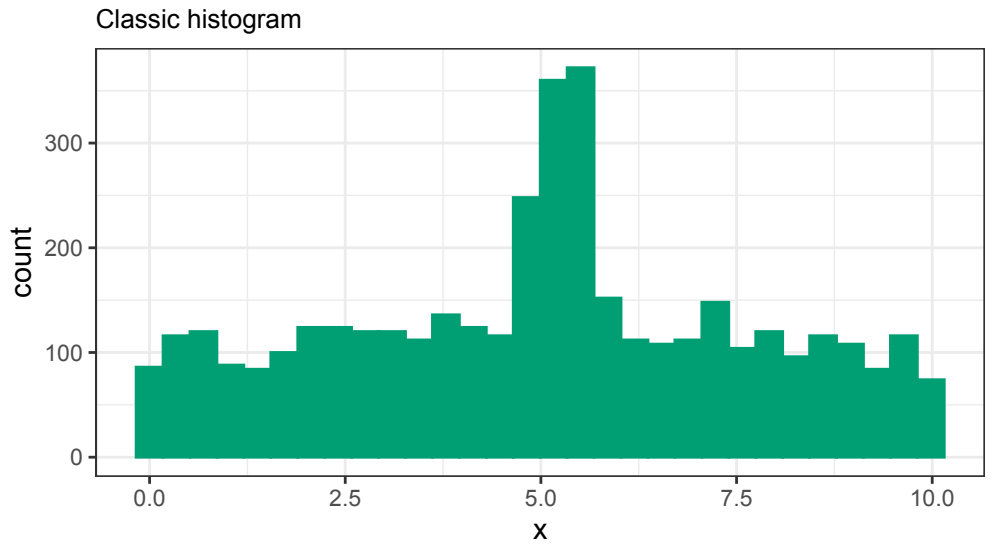


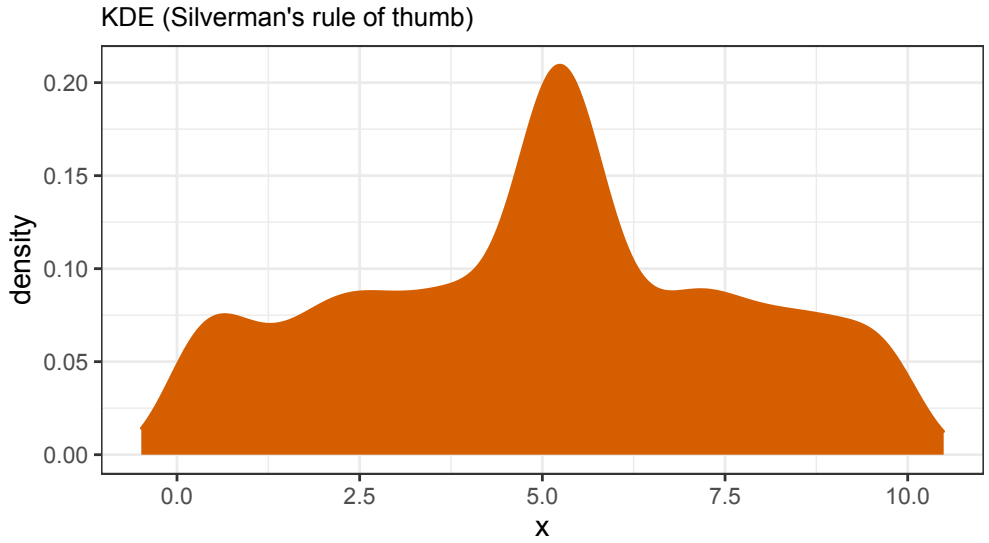


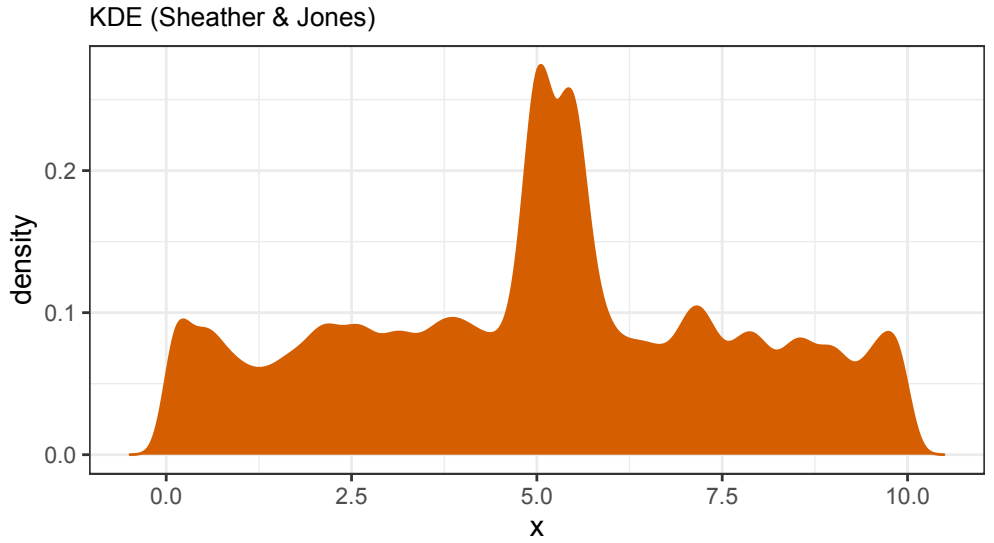
https://aakinshin.net/posts/qrde-hd/

Квантильная оценка плотности (QRDE)



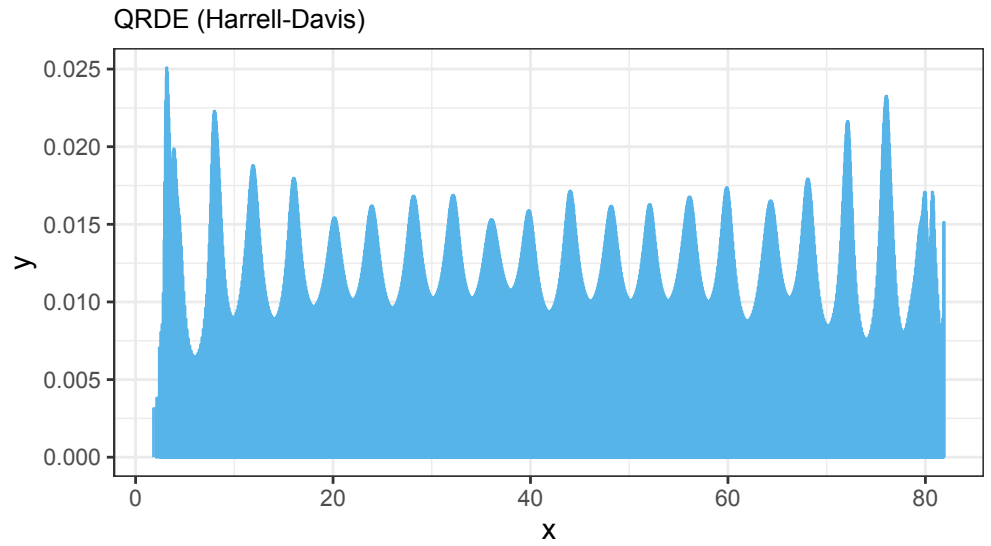


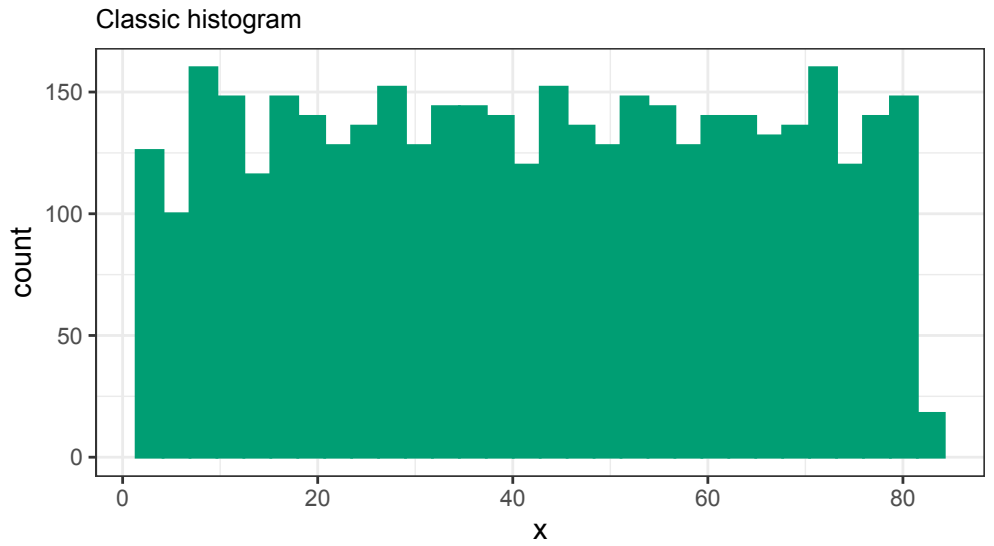


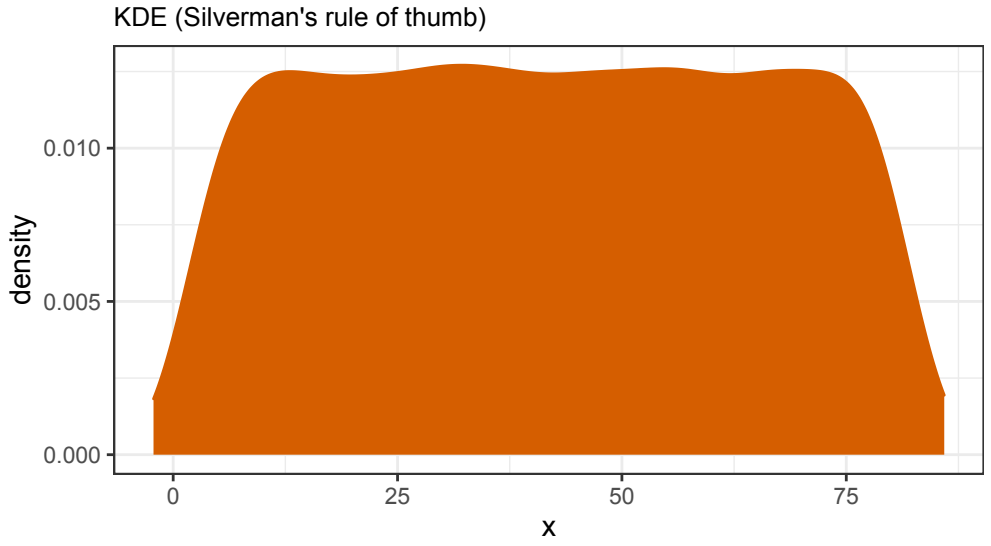


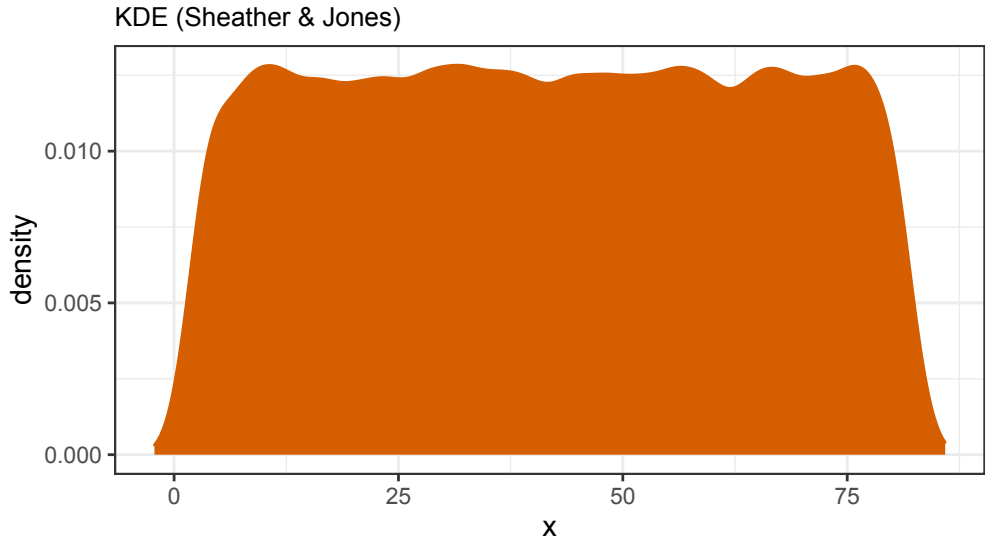
https://aakinshin.net/posts/qrde-hd/

Квантильная оценка плотности (QRDE)





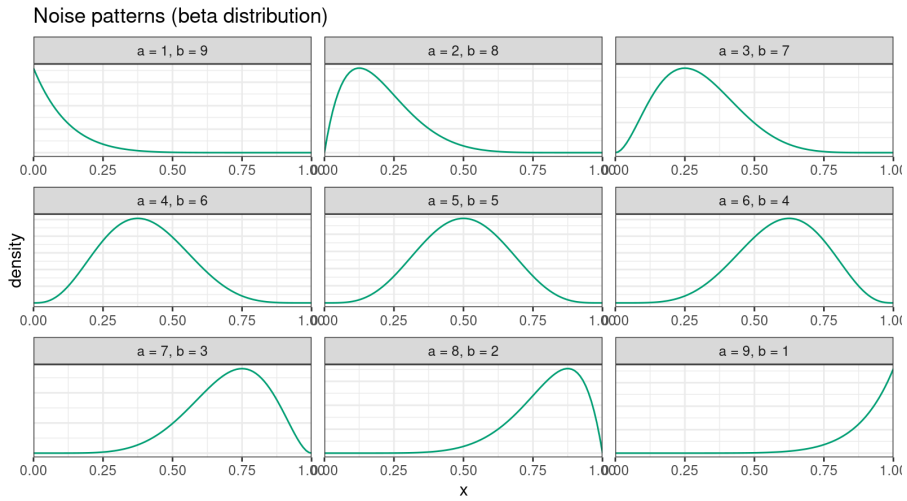




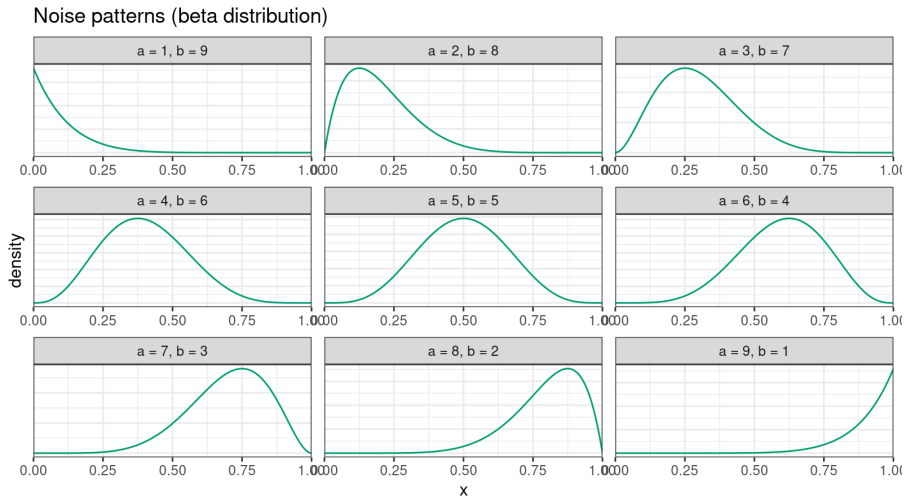
https://aakinshin.net/posts/qrde-hd/

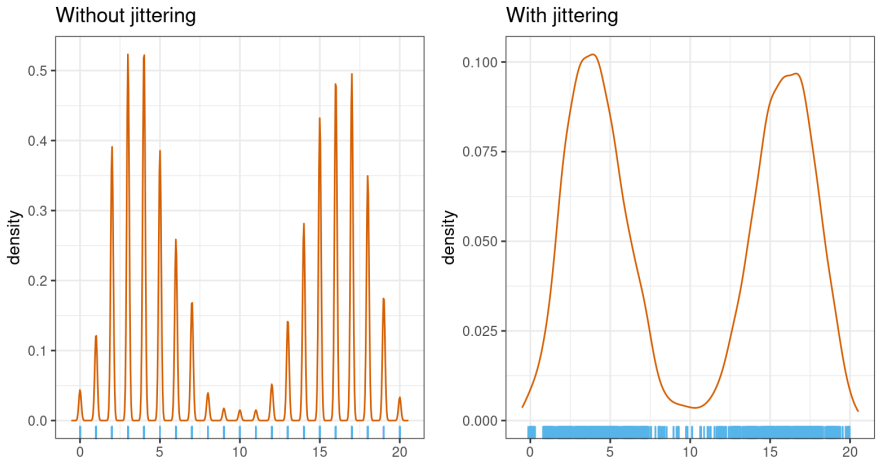
$$x = \{1, 4, 3, 2, 4, 2, 3, 4, 3, 2, 4, \ldots\}$$

Улучшаем KDE/QRDE с помощью джиттеринга



Улучшаем KDE/QRDE с помощью джиттеринга

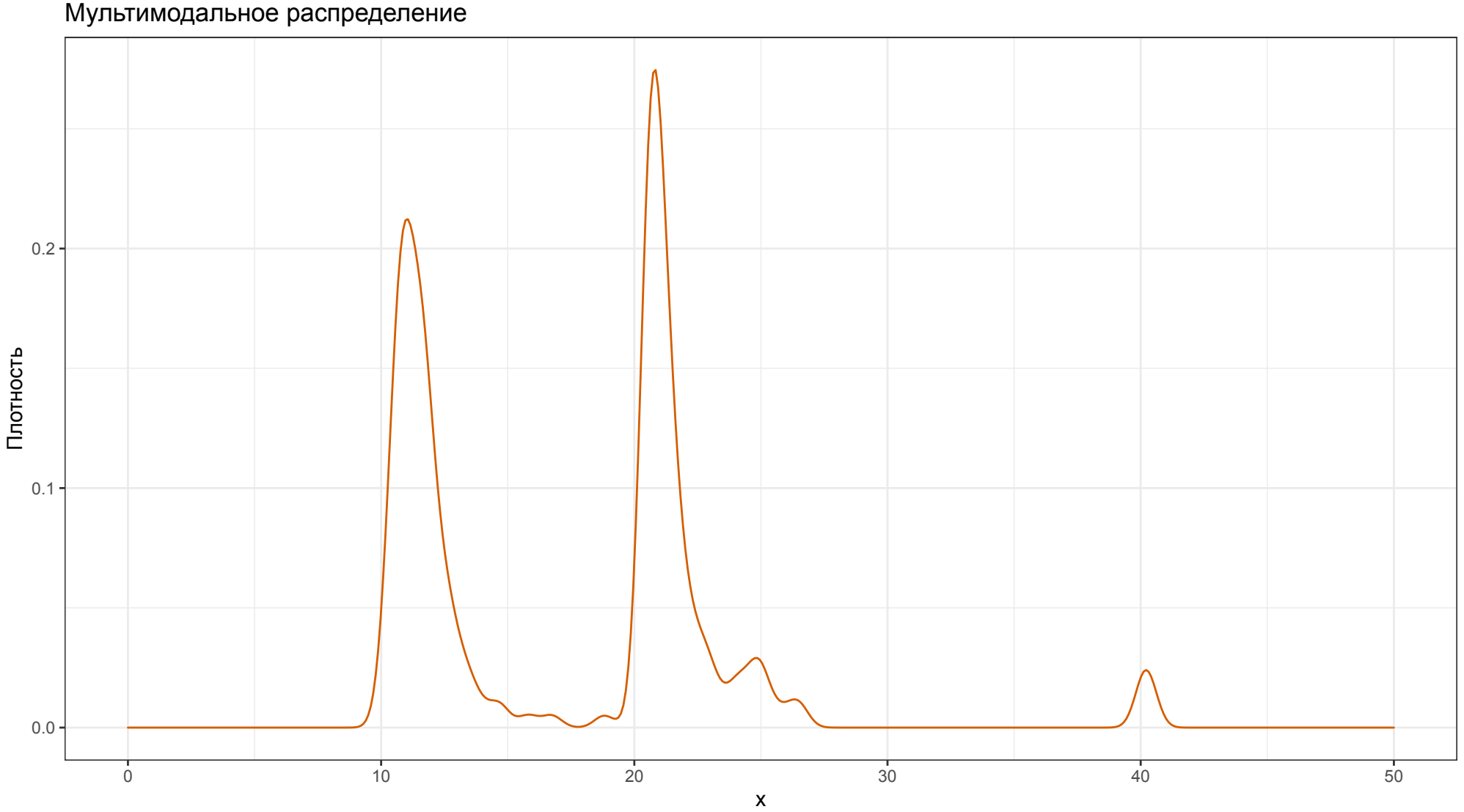


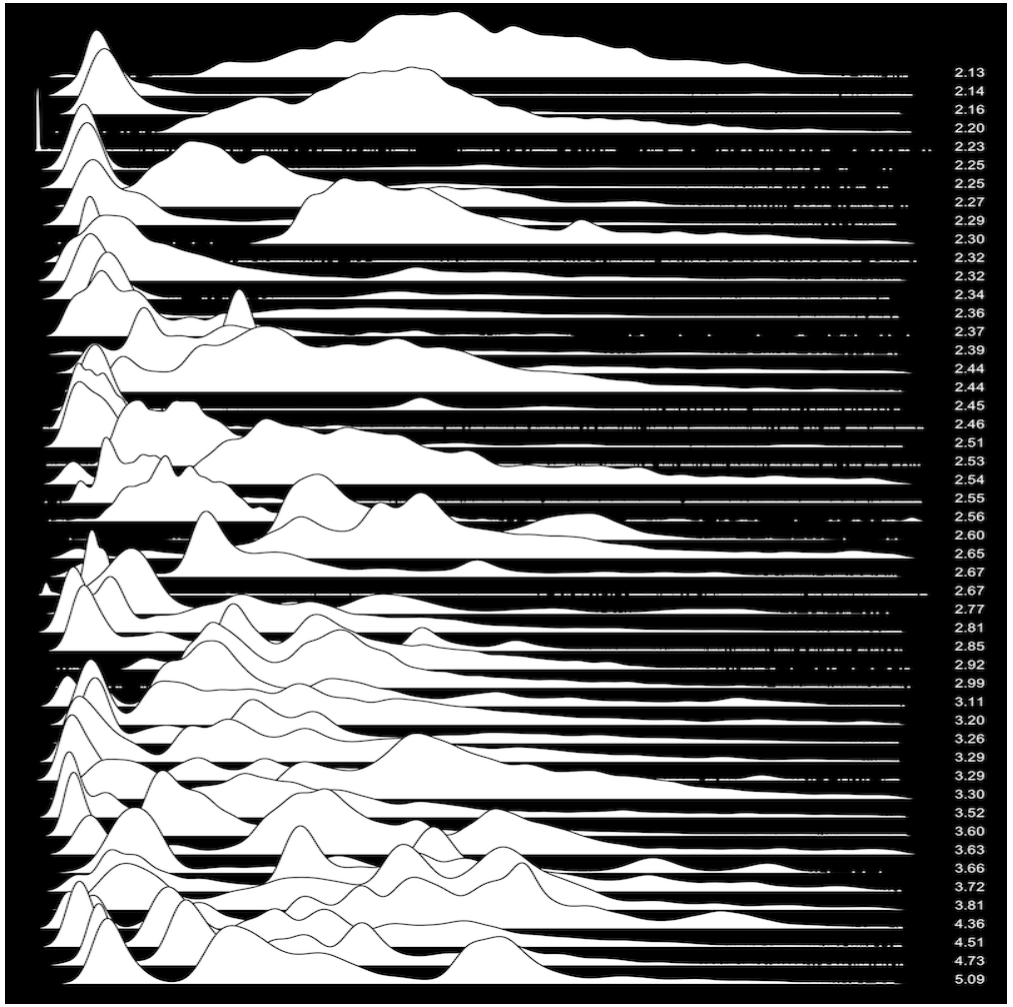


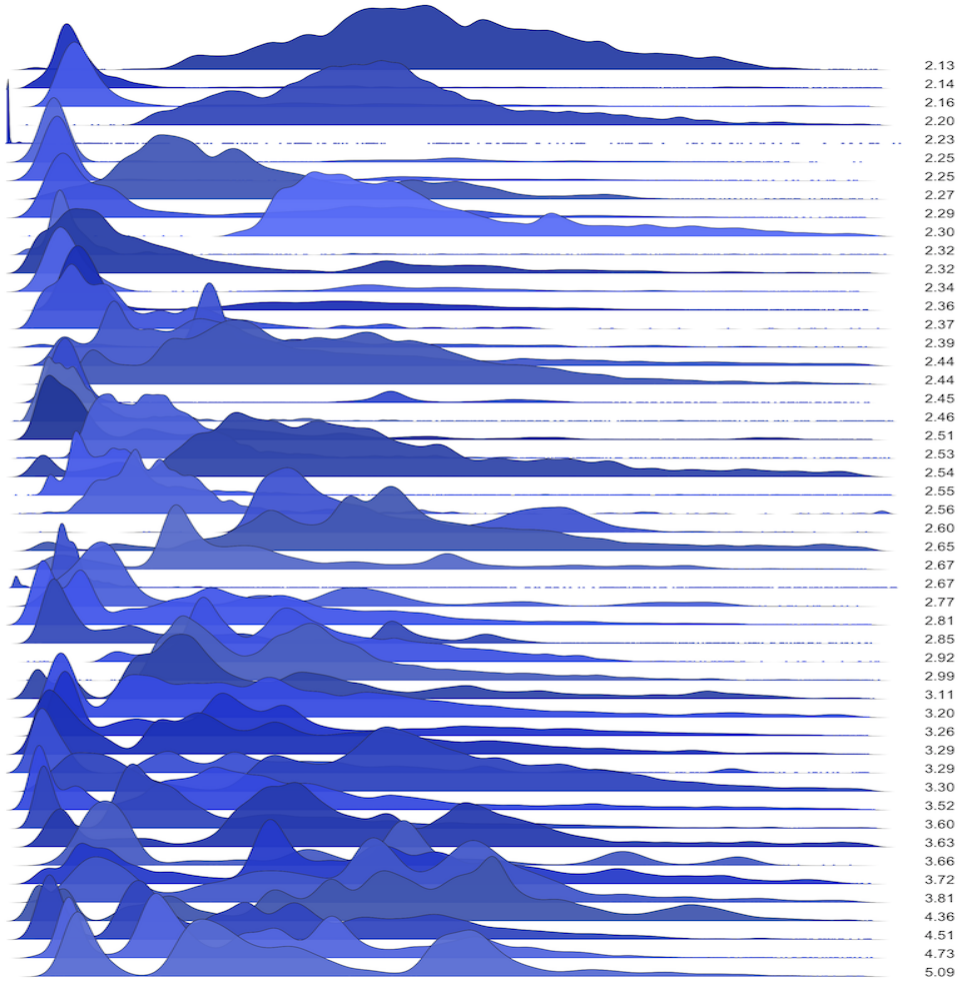
https://aakinshin.net/posts/discrete-sample-jittering/
https://aakinshin.net/posts/qrde-discrete/

Для выбора оценки плотности
нужно подумать

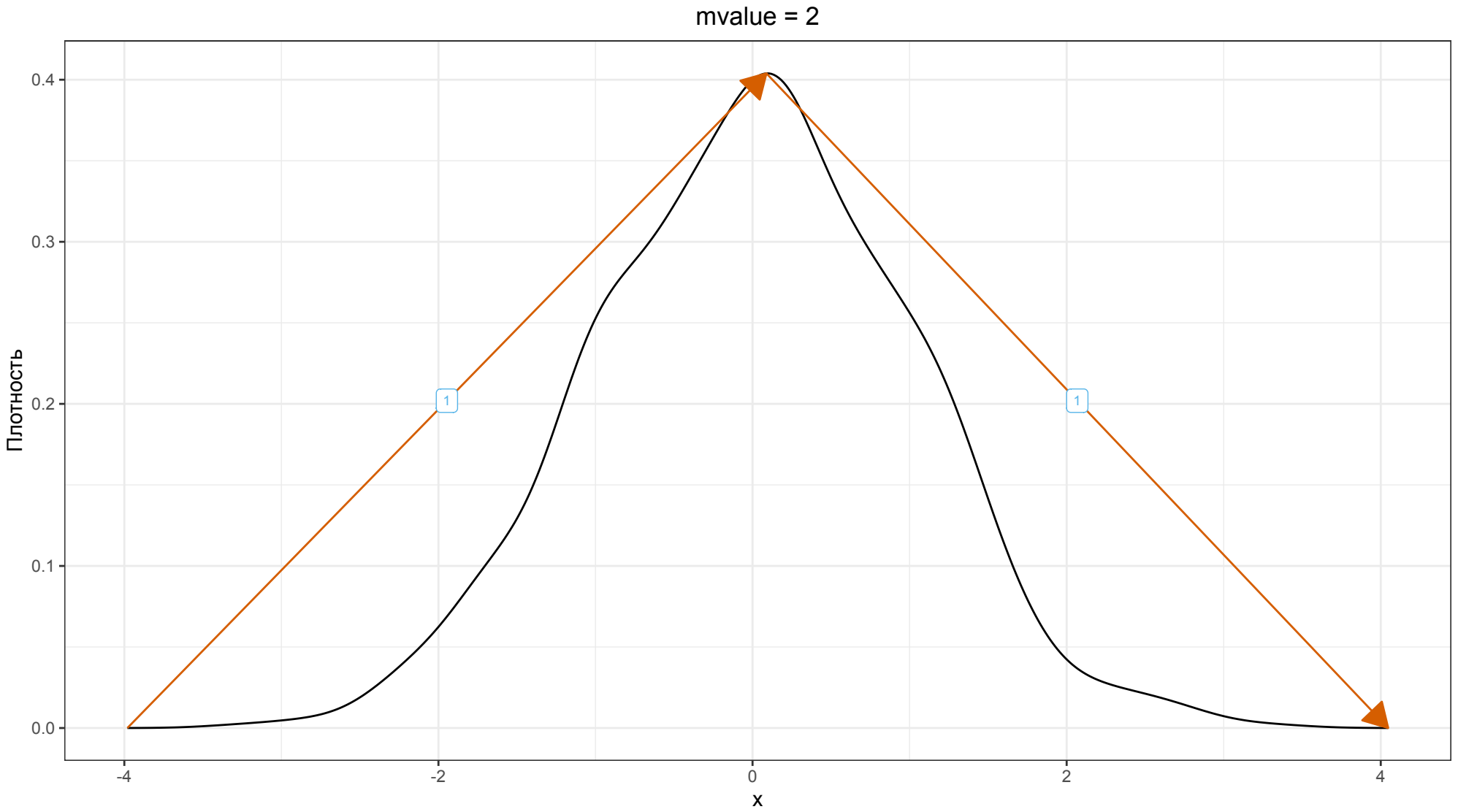
Мультимодальность


Мультимодальное распределение
Плотность
0.2
0.1
0.0
0
10
20
30
40
50
x

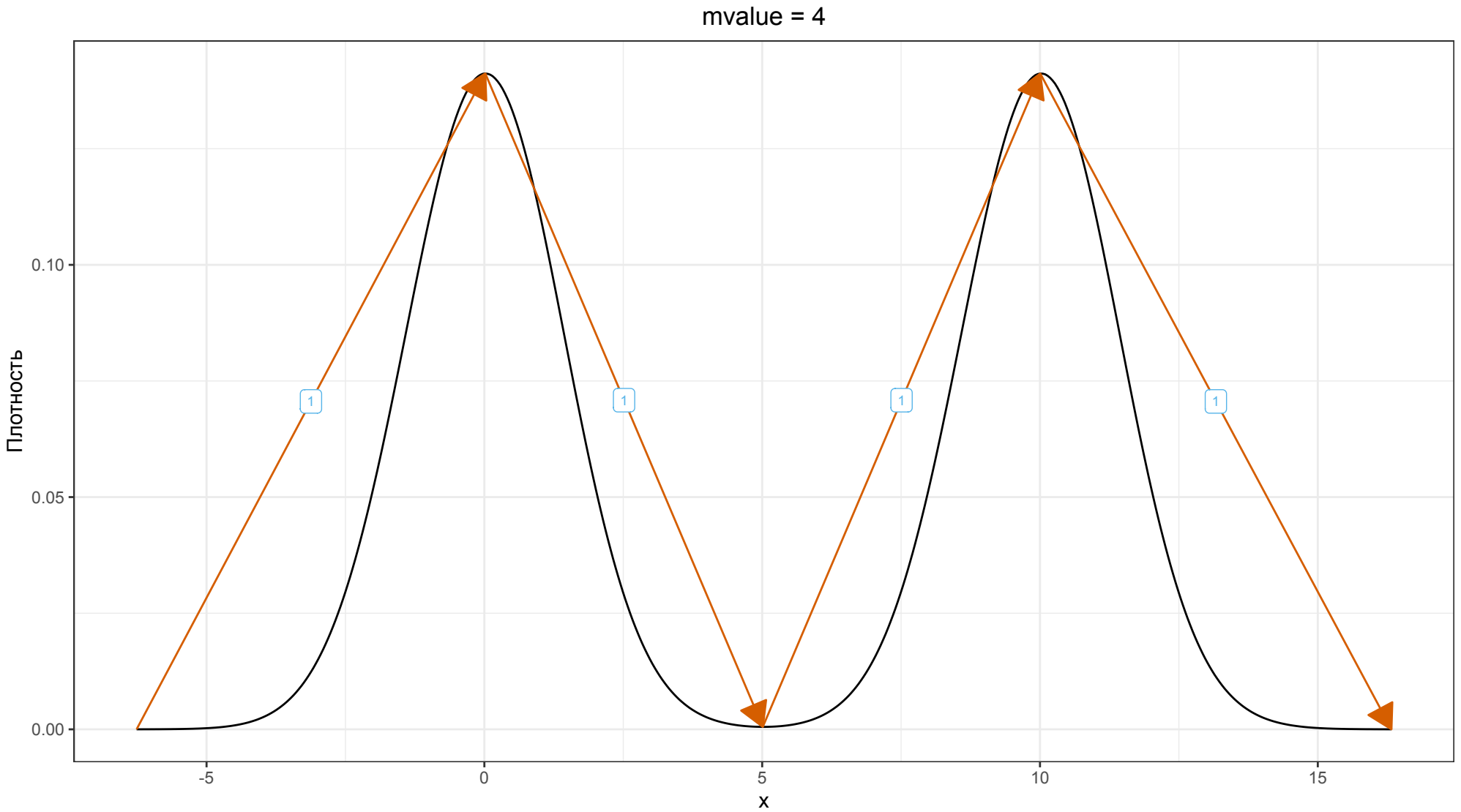




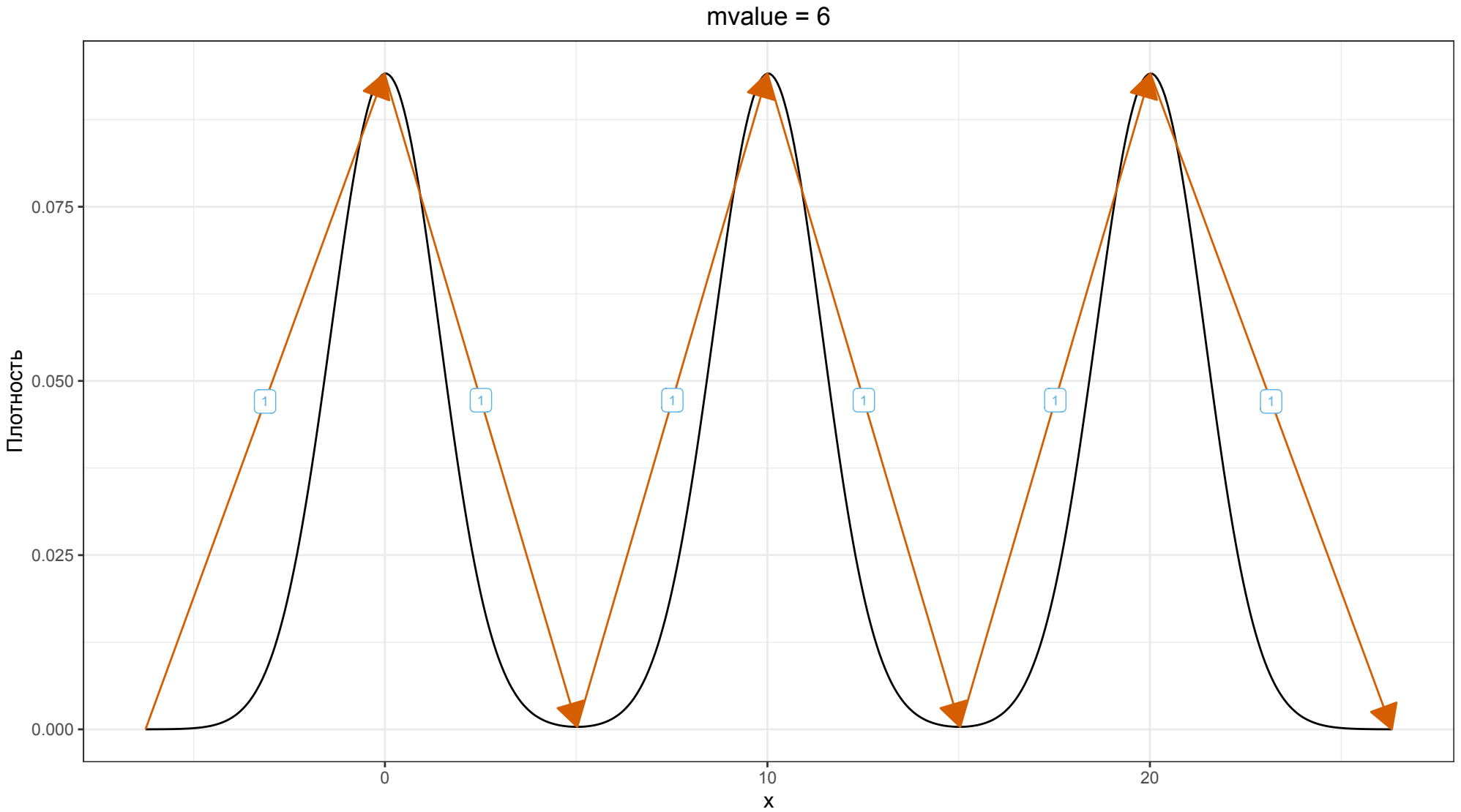
http://www.brendangregg.com/FrequencyTrails/modes.html



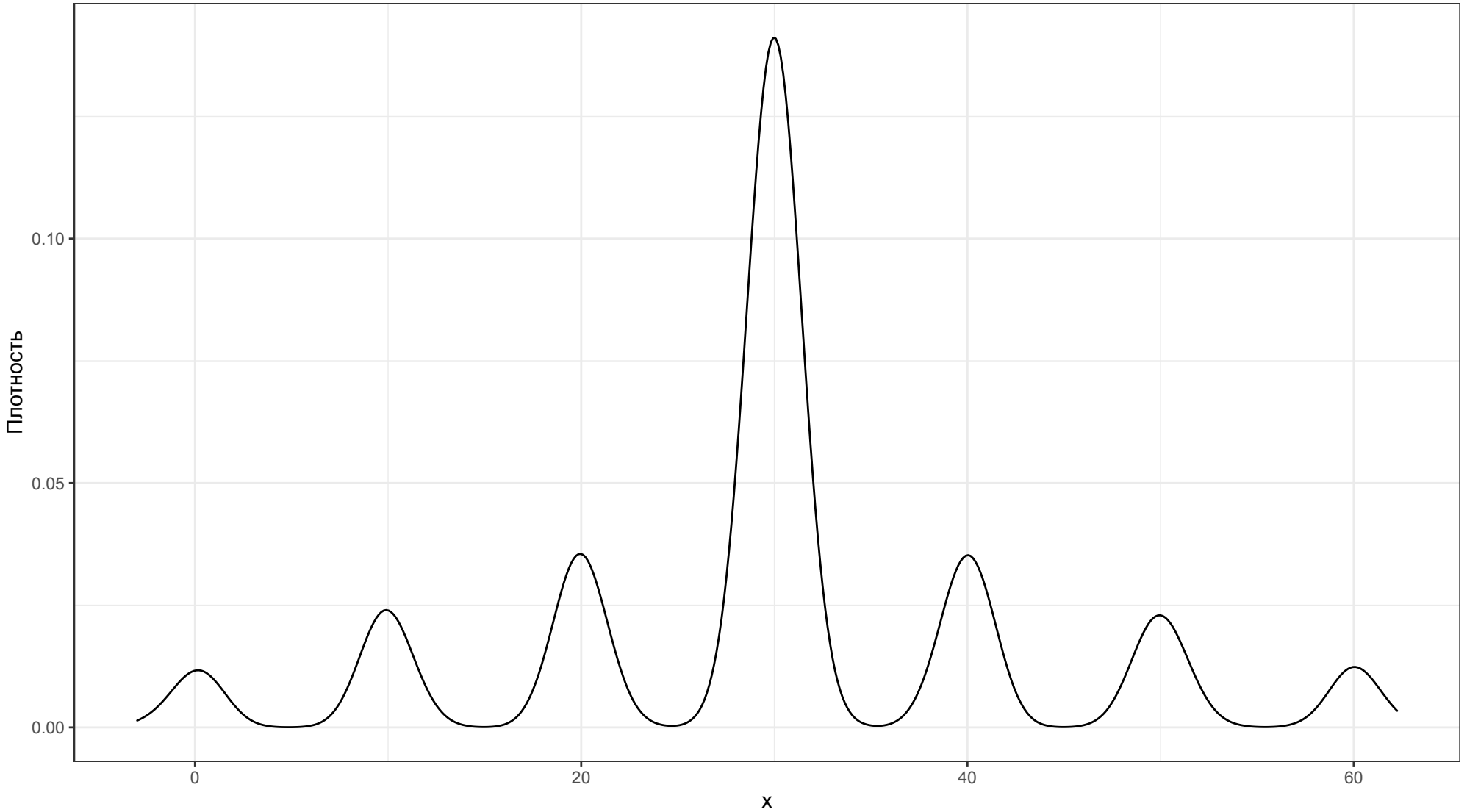
http://www.brendangregg.com/FrequencyTrails/modes.html



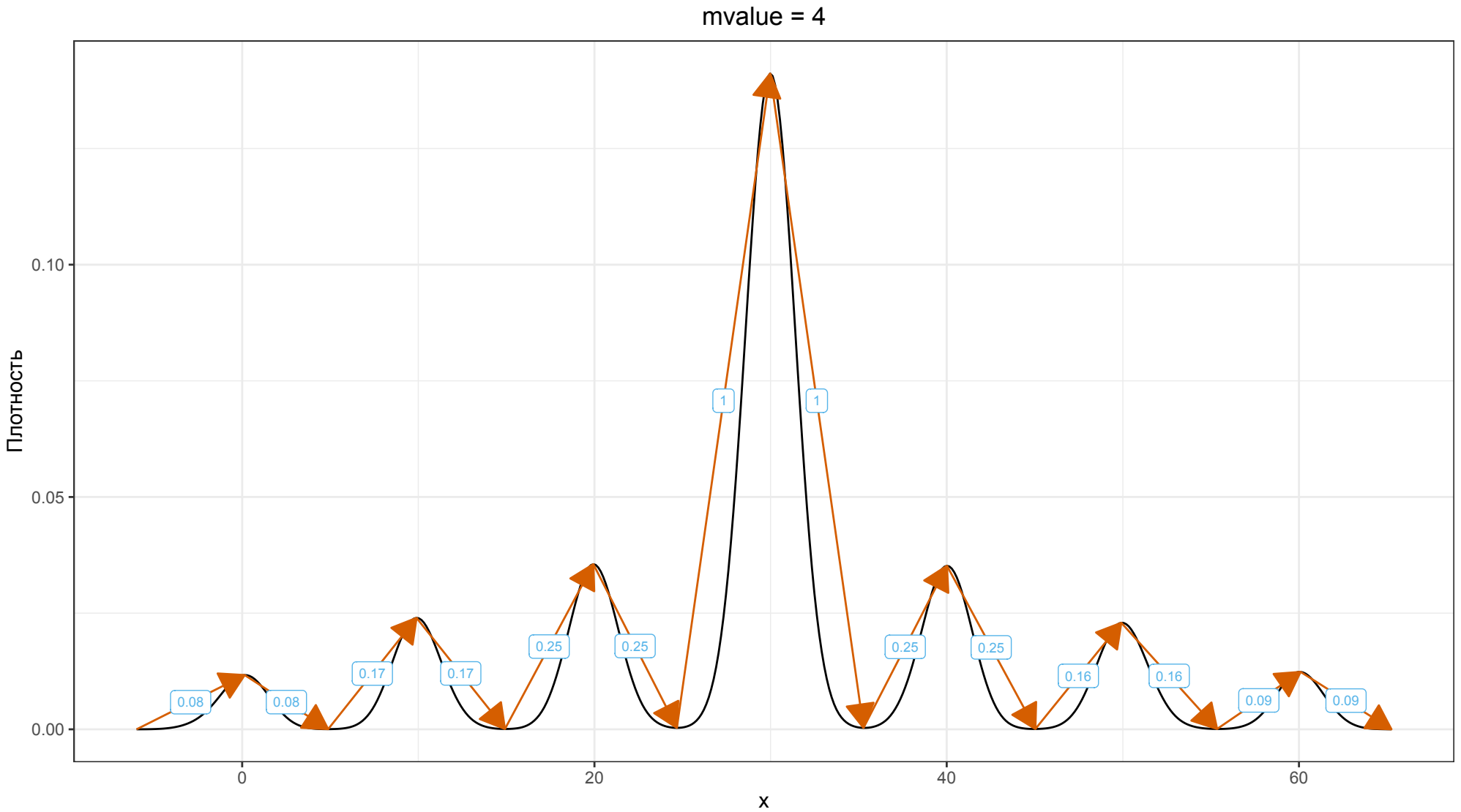
http://www.brendangregg.com/FrequencyTrails/modes.html



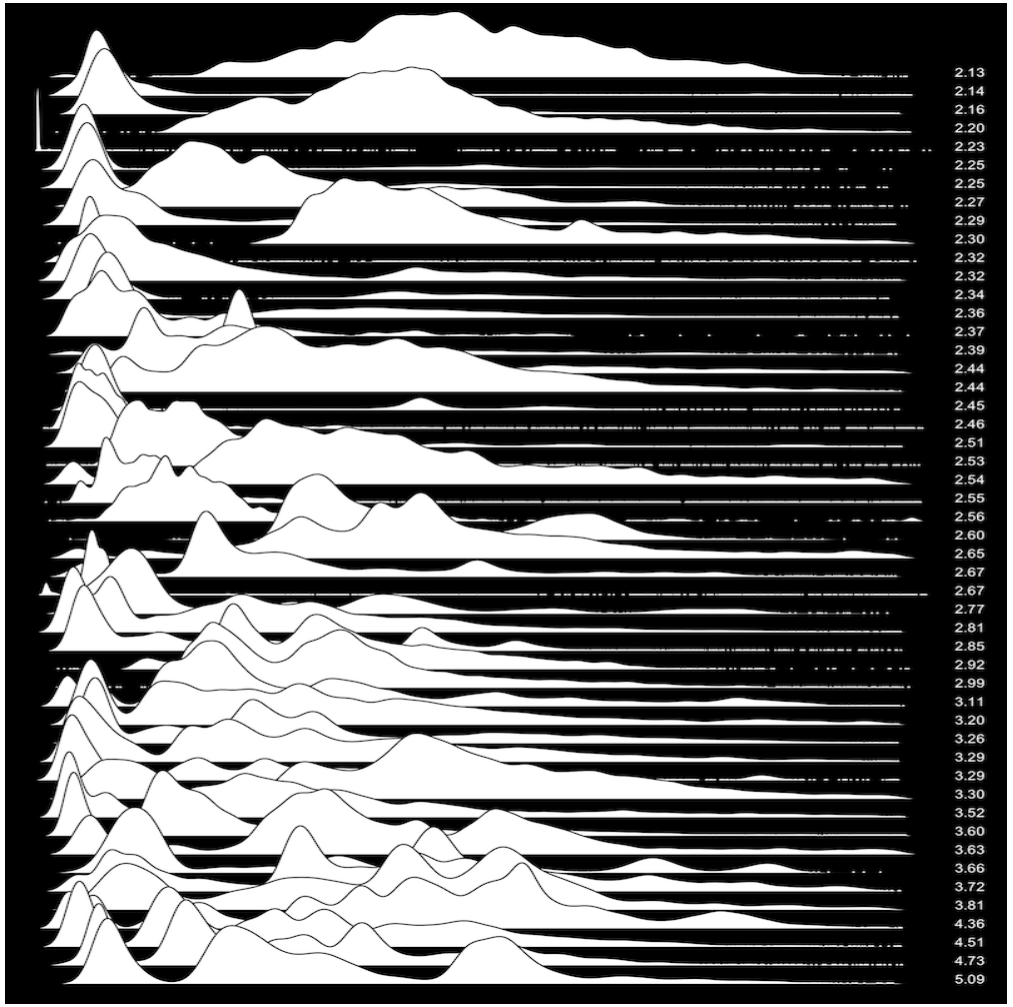
http://www.brendangregg.com/FrequencyTrails/modes.html

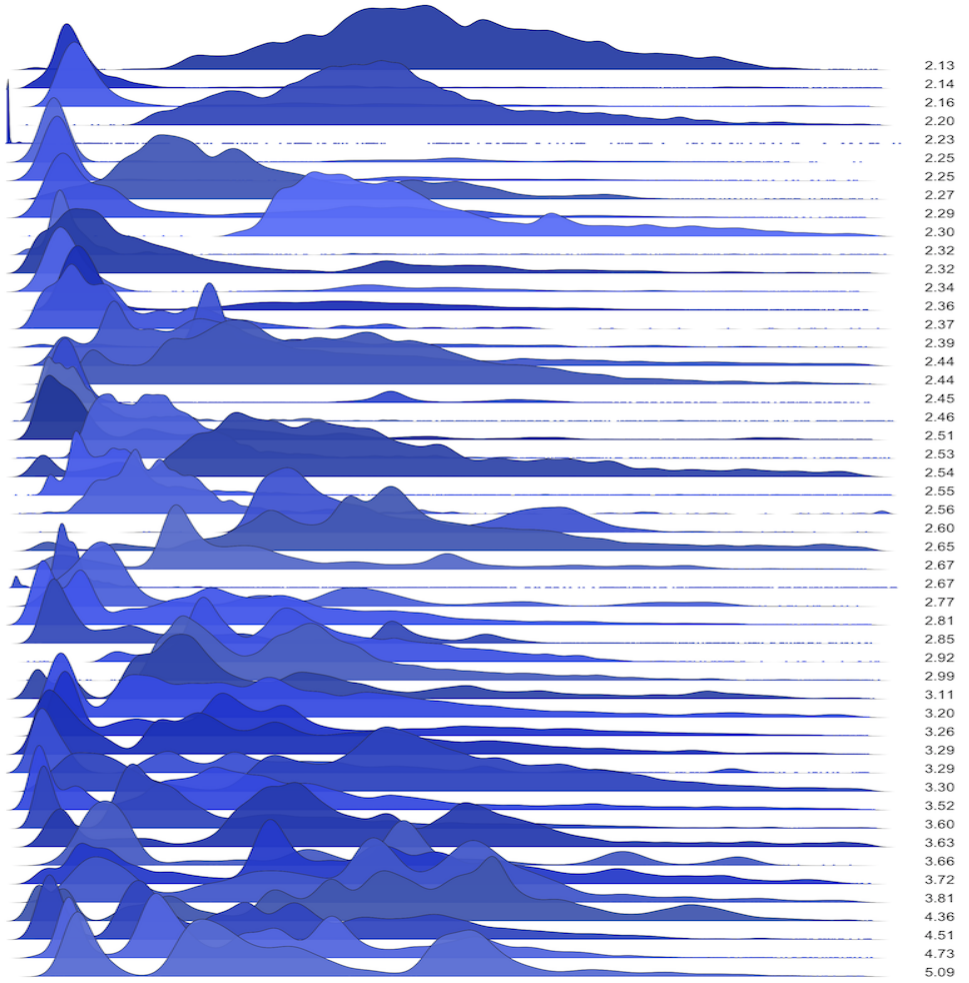


http://www.brendangregg.com/FrequencyTrails/modes.html



http://www.brendangregg.com/FrequencyTrails/modes.html

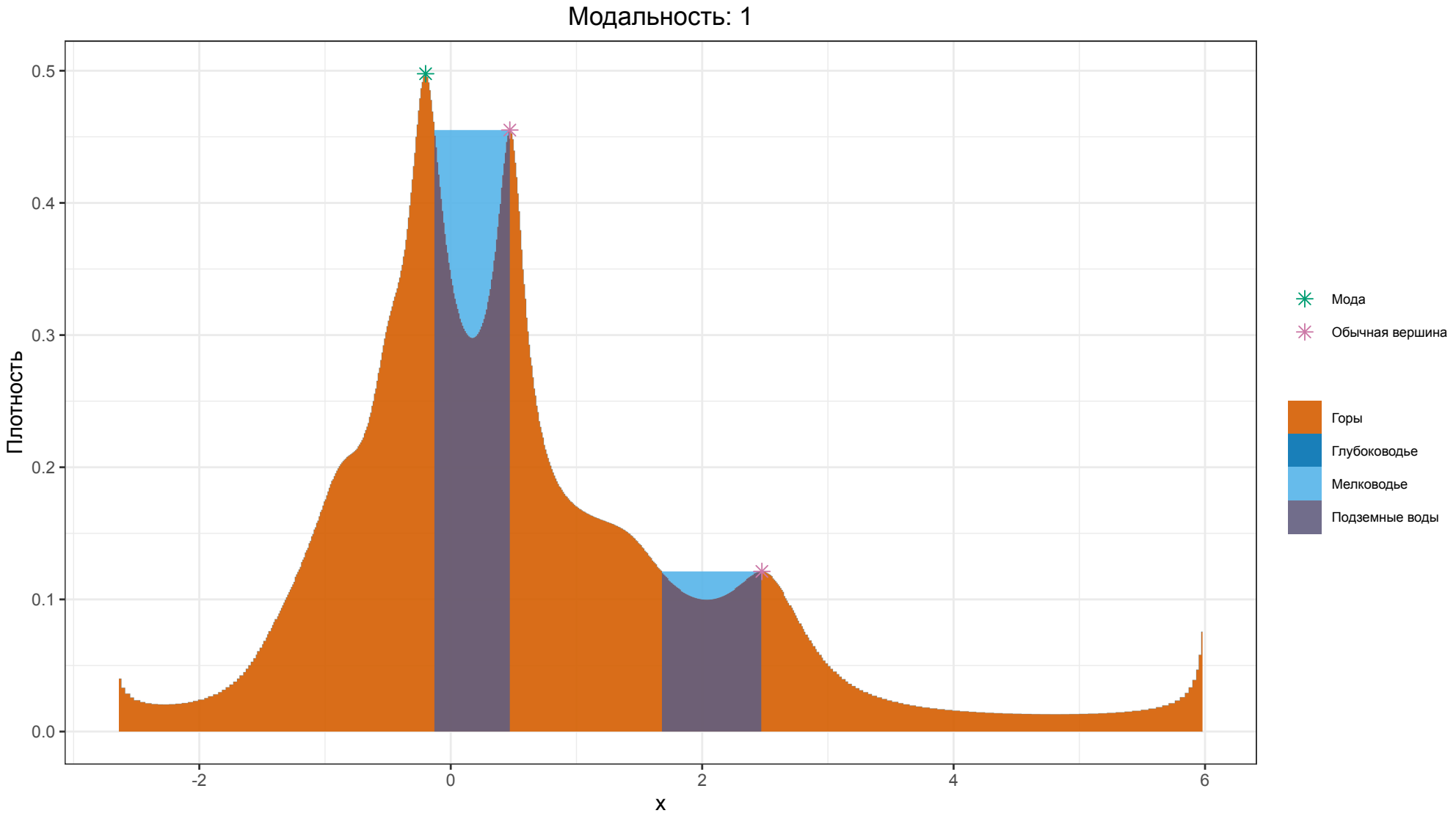




http://www.brendangregg.com/FrequencyTrails/modes.html

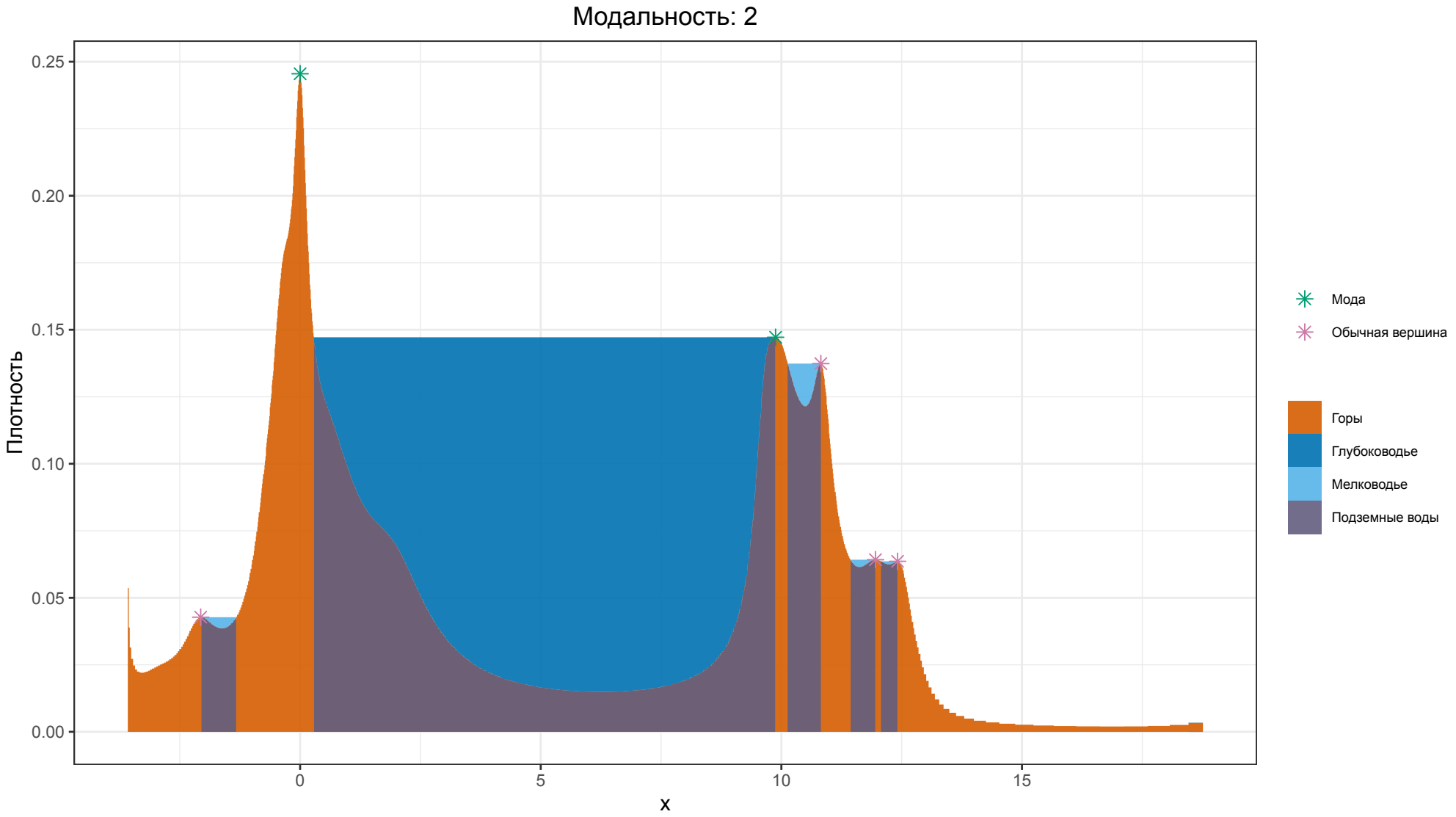
Lowland multimodality detection

Lowland multimodality detection



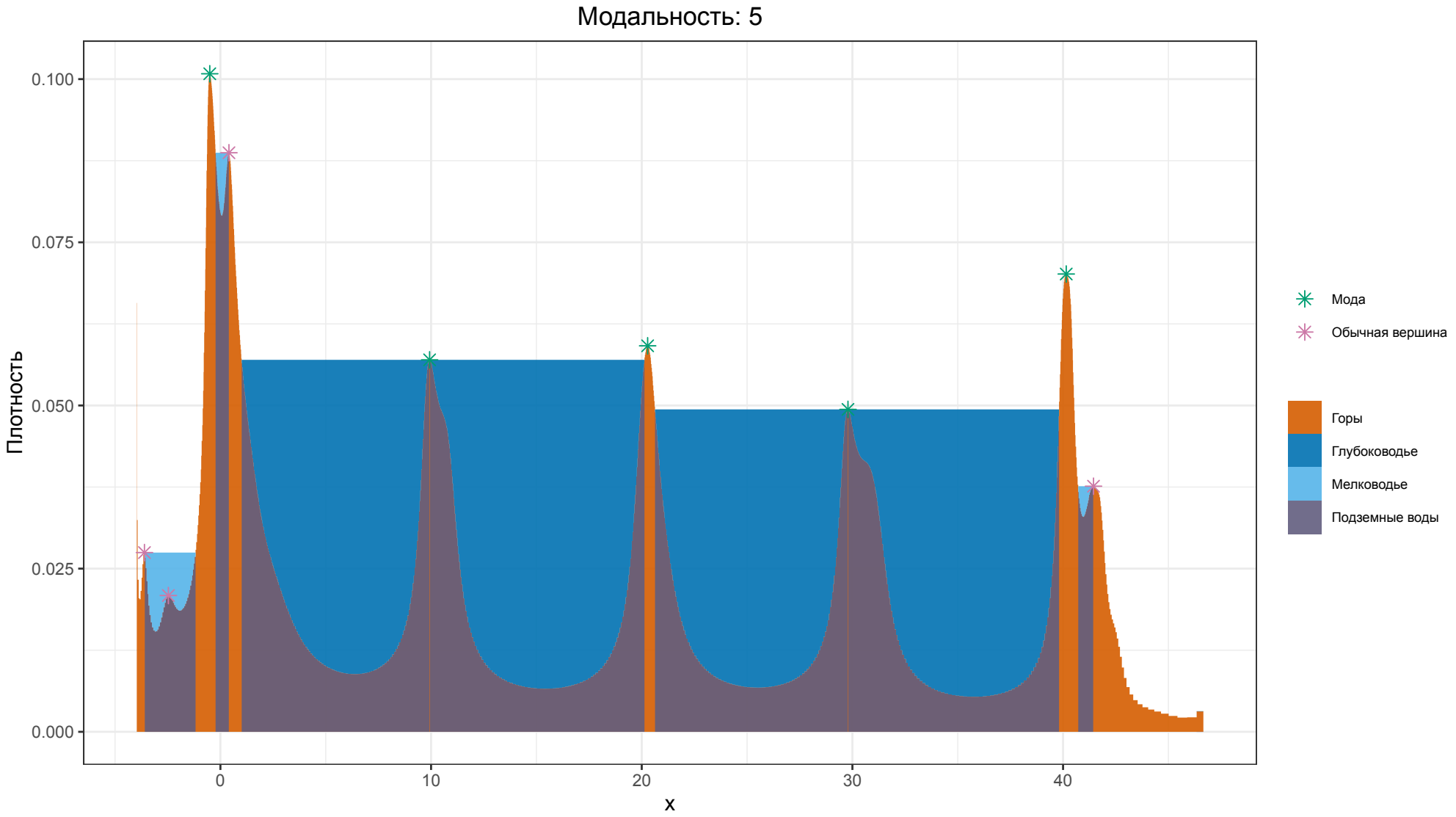
https://aakinshin.net/posts/lowland-multimodality-detection/

Lowland multimodality detection



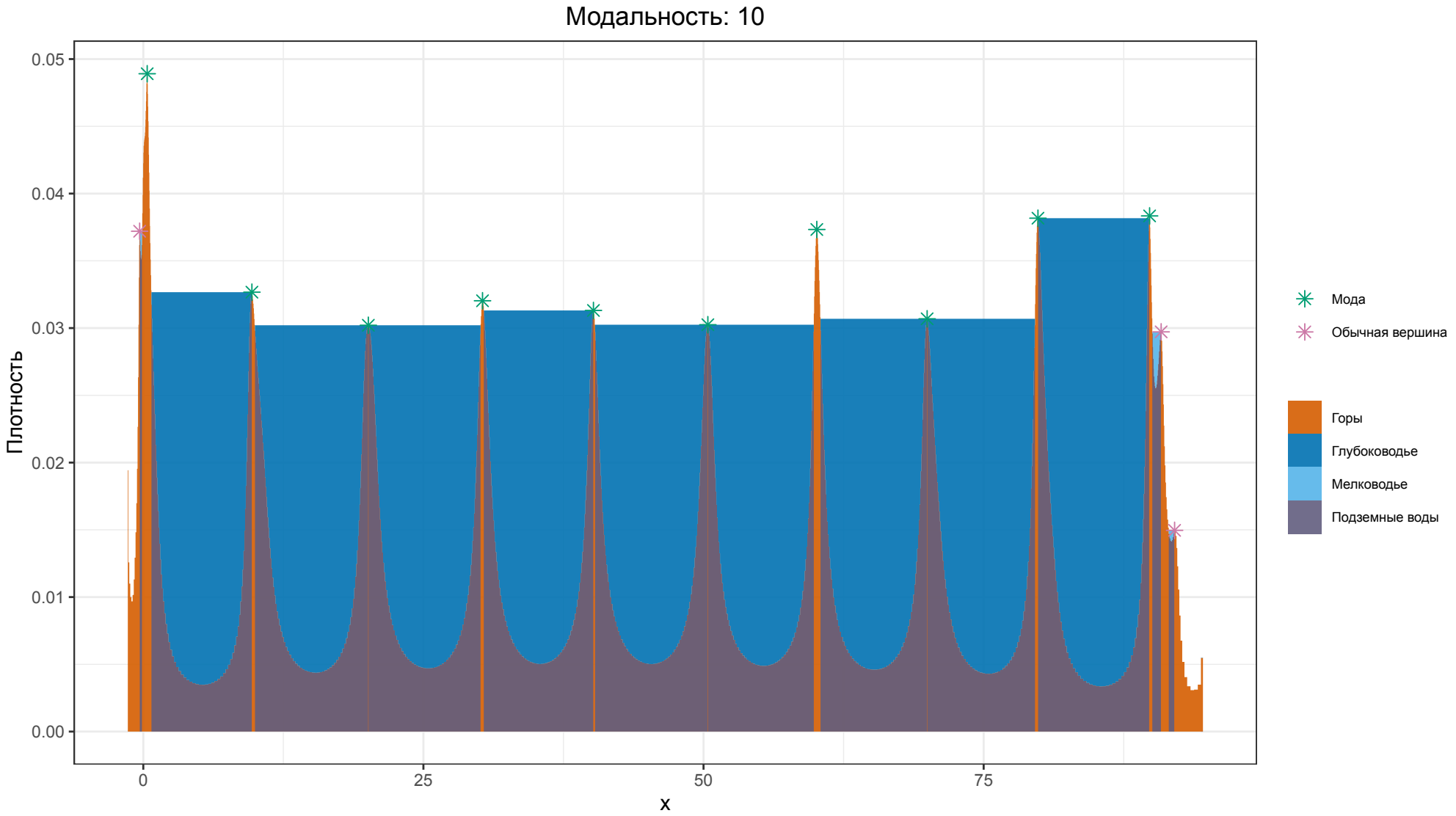
https://aakinshin.net/posts/lowland-multimodality-detection/

Lowland multimodality detection



https://aakinshin.net/posts/lowland-multimodality-detection/

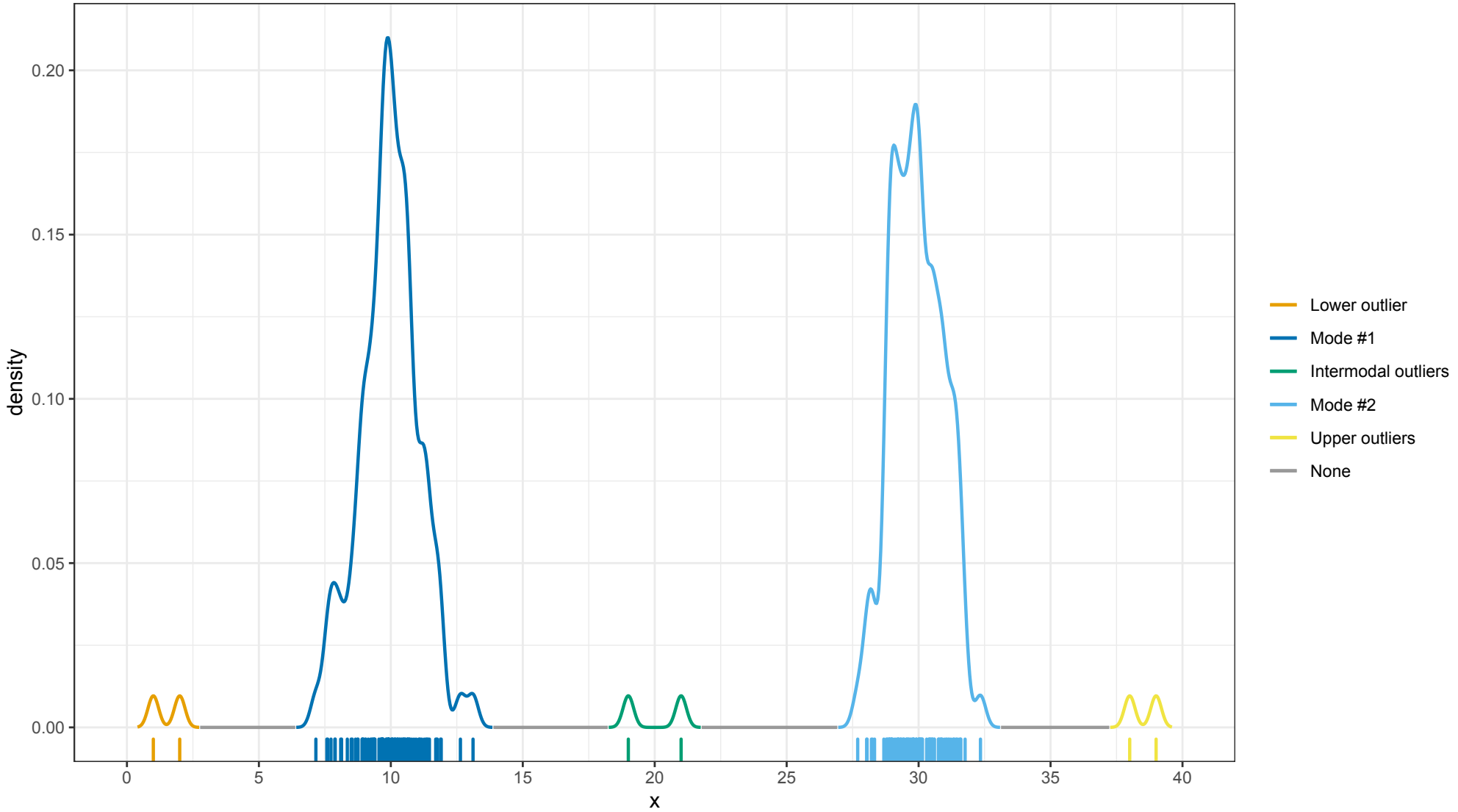
Lowland multimodality detection



https://aakinshin.net/posts/lowland-multimodality-detection/

Интермодальные выбросы

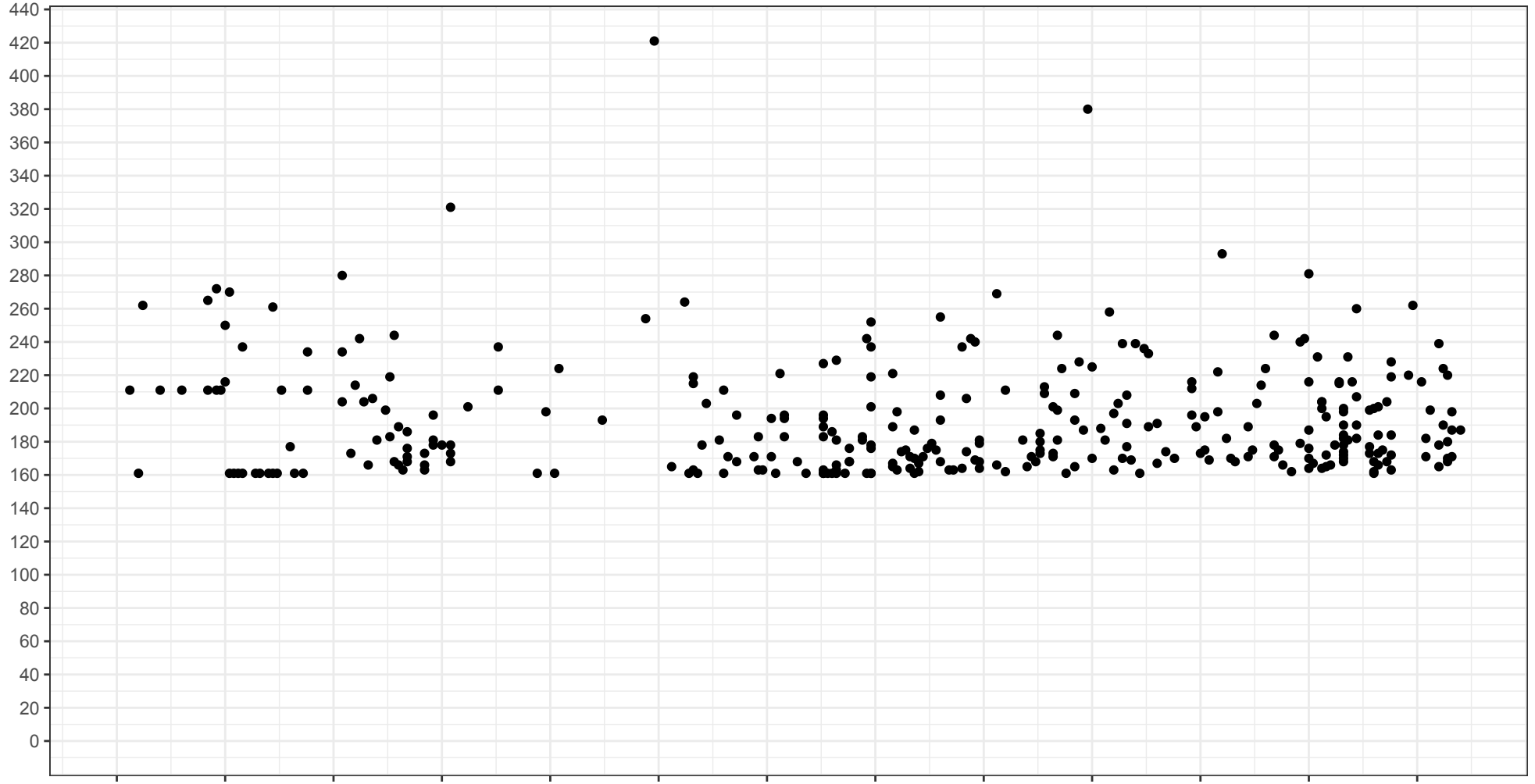
Интермодальные выбросы

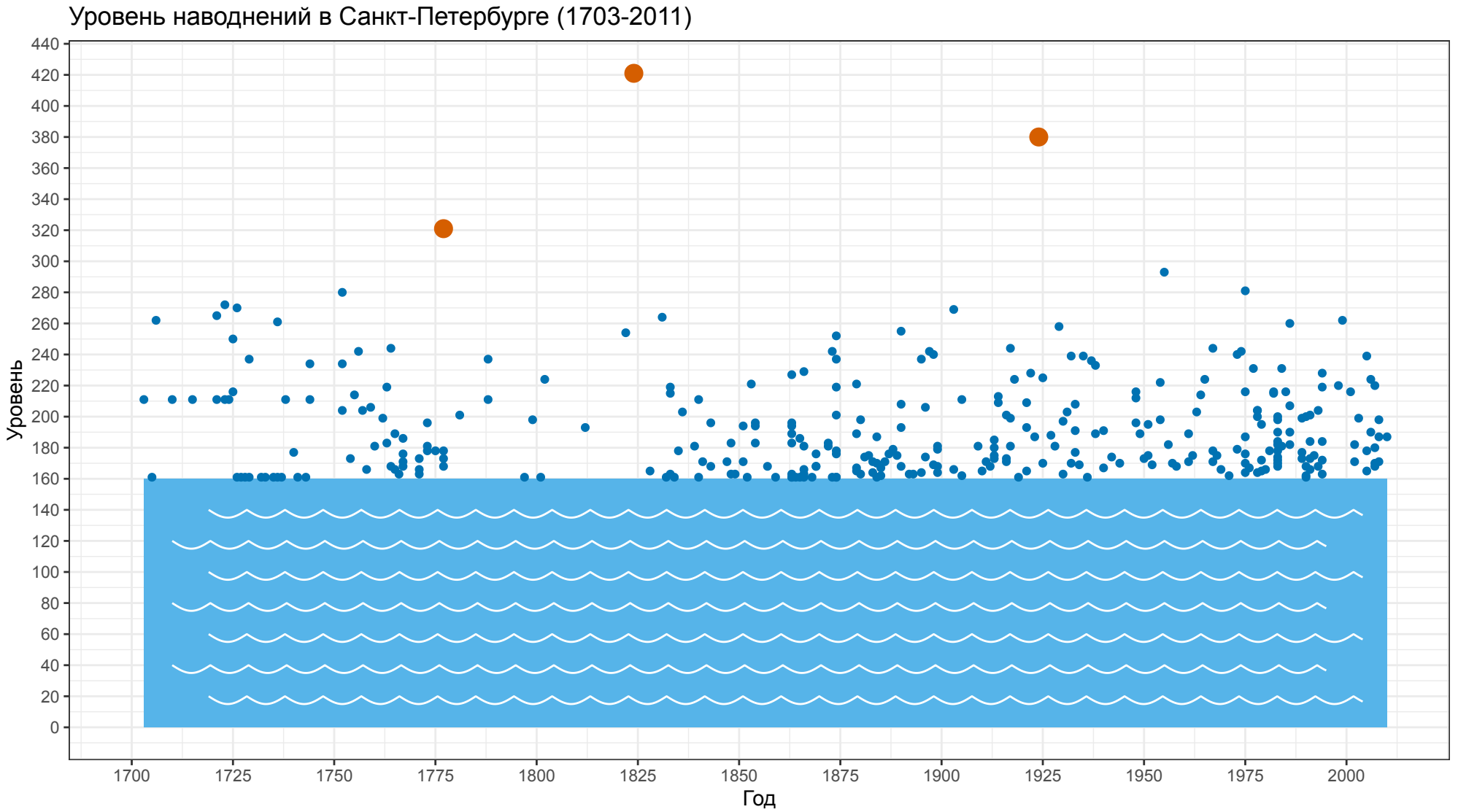


https://aakinshin.net/posts/intermodal-outliers/

Для детектирования мультимодальность
нужно подумать

Теория экстремальных значений


440
420
400
380
360
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0

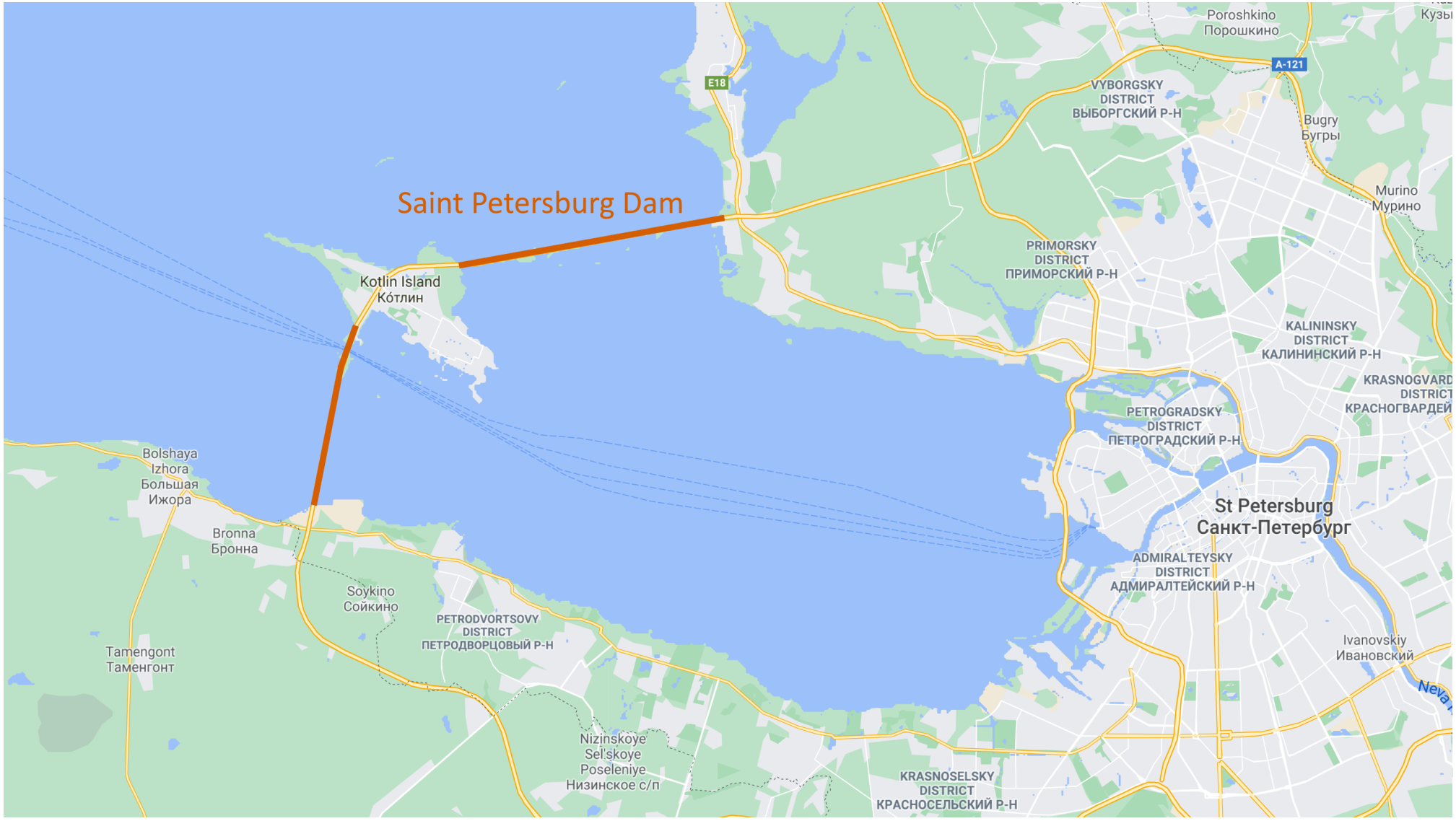


Источник данных: http://www.nevariver.ru/flood_list.php

Наводнение в Санкт-Петербурге (1824г)

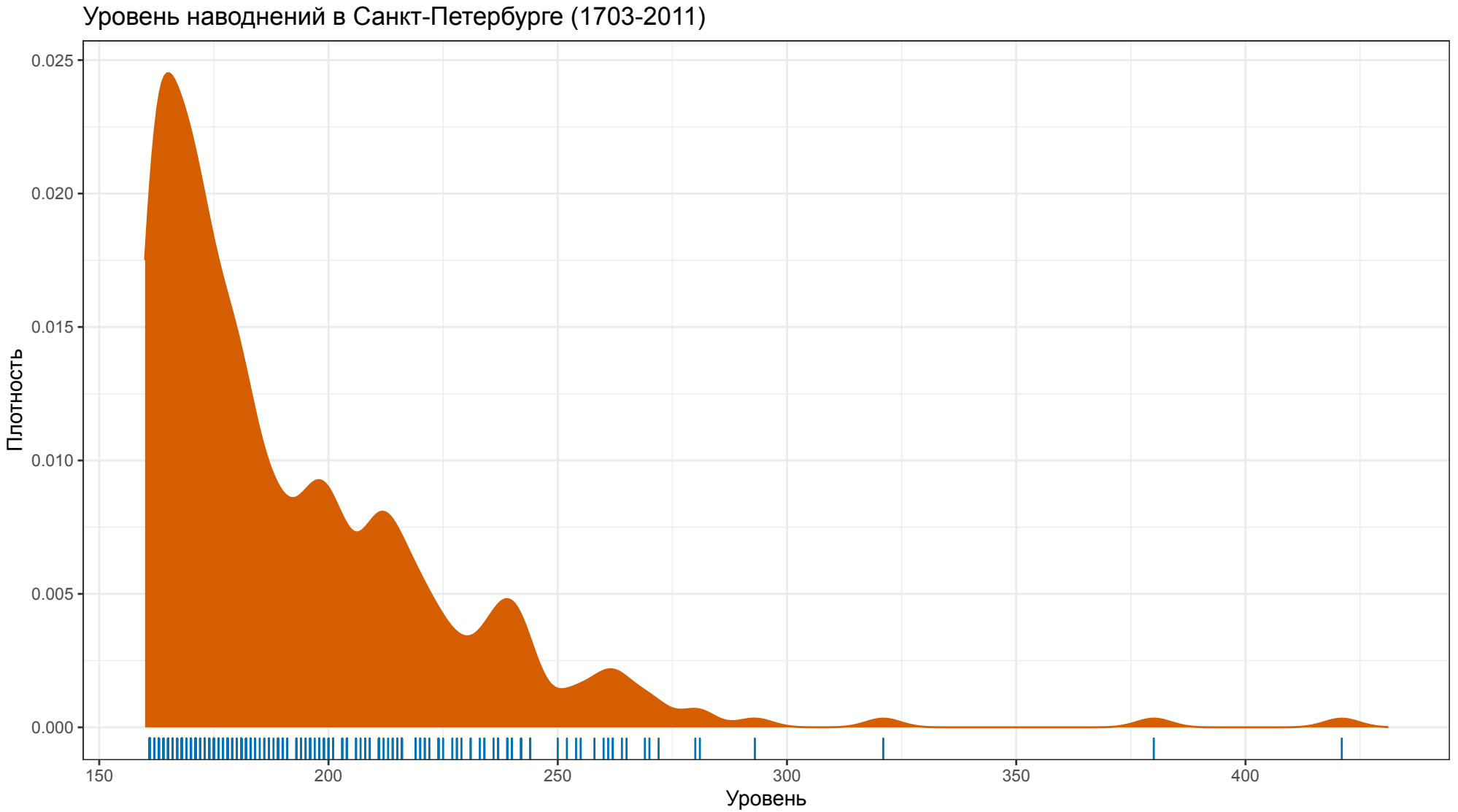
https://ru.m.wikipedia.org/wiki/Файл:7_ноября_1824_года_на_площади_у_Большого_театра.jpg

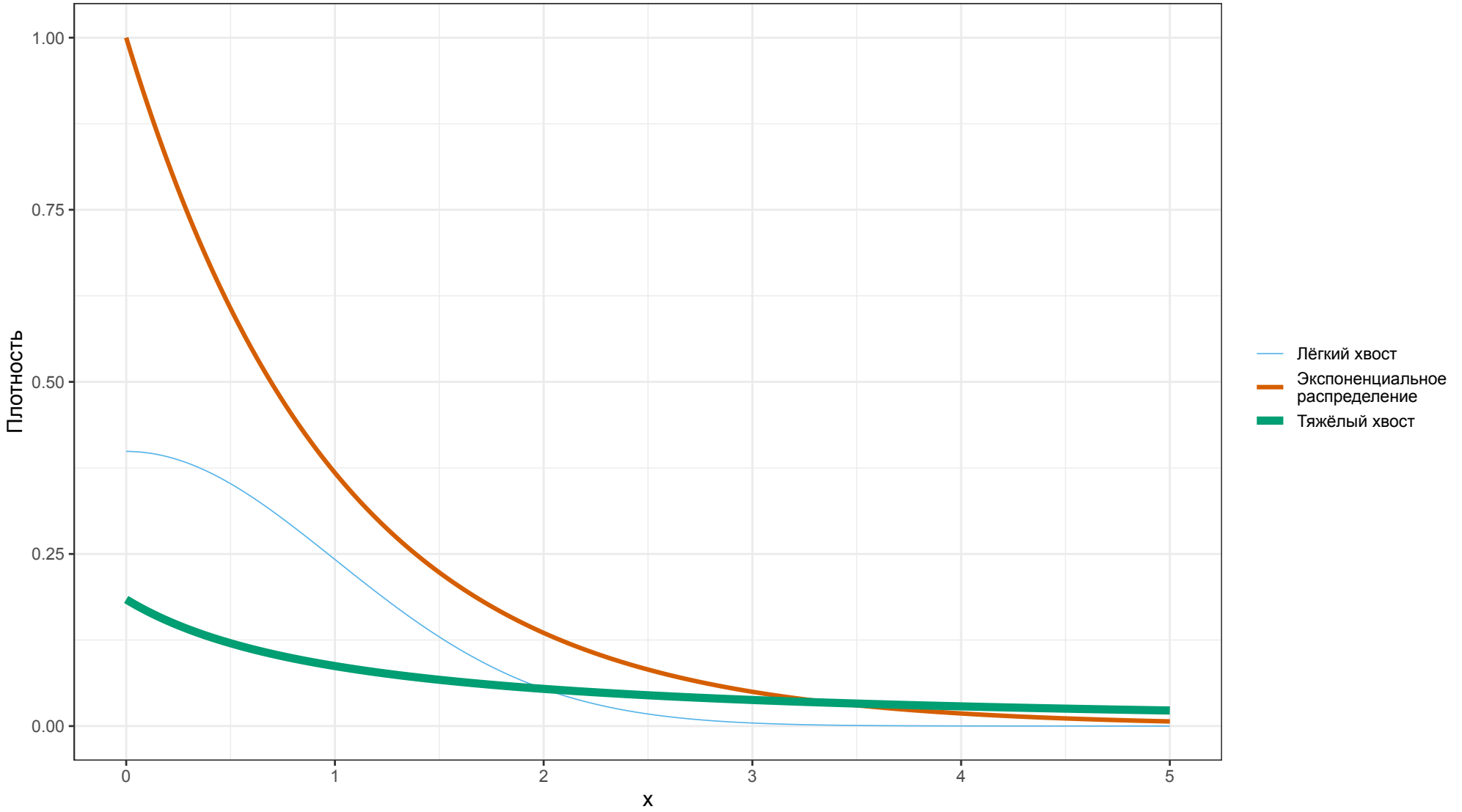
Санкт-Петербургская дамба



https://www.google.com/maps/@59.9552059,29.915138,10z

Распределение уровней наводнений




1.00
0.75
0.50
0.25
0.00
Плотность
0
1
2
3
4
5
x
Лёгкий хвост
Экспоненциальное распределение
Тяжёлый хвост

Лёгкий хвост	**Тяжёлый хвост**
Равномерное	Парето
Нормальное	Логнормальное
Бета	Коши
Экспоненциальное	Фреше
Вейбул (k>1)	Вейбул (k<1)

Приложения теории экстремальных значений

- Гидрология
- Эпидемиология
- ...

Приложения теории экстремальных значений

- Гидрология
- Эпидемиология
- ...
- Финансы
- Страхование
- ...

Приложения теории экстремальных значений

- Гидрология
- Эпидемиология
- ...
- Финансы
- Страхование
- ...
- **Перформанс-анализ**

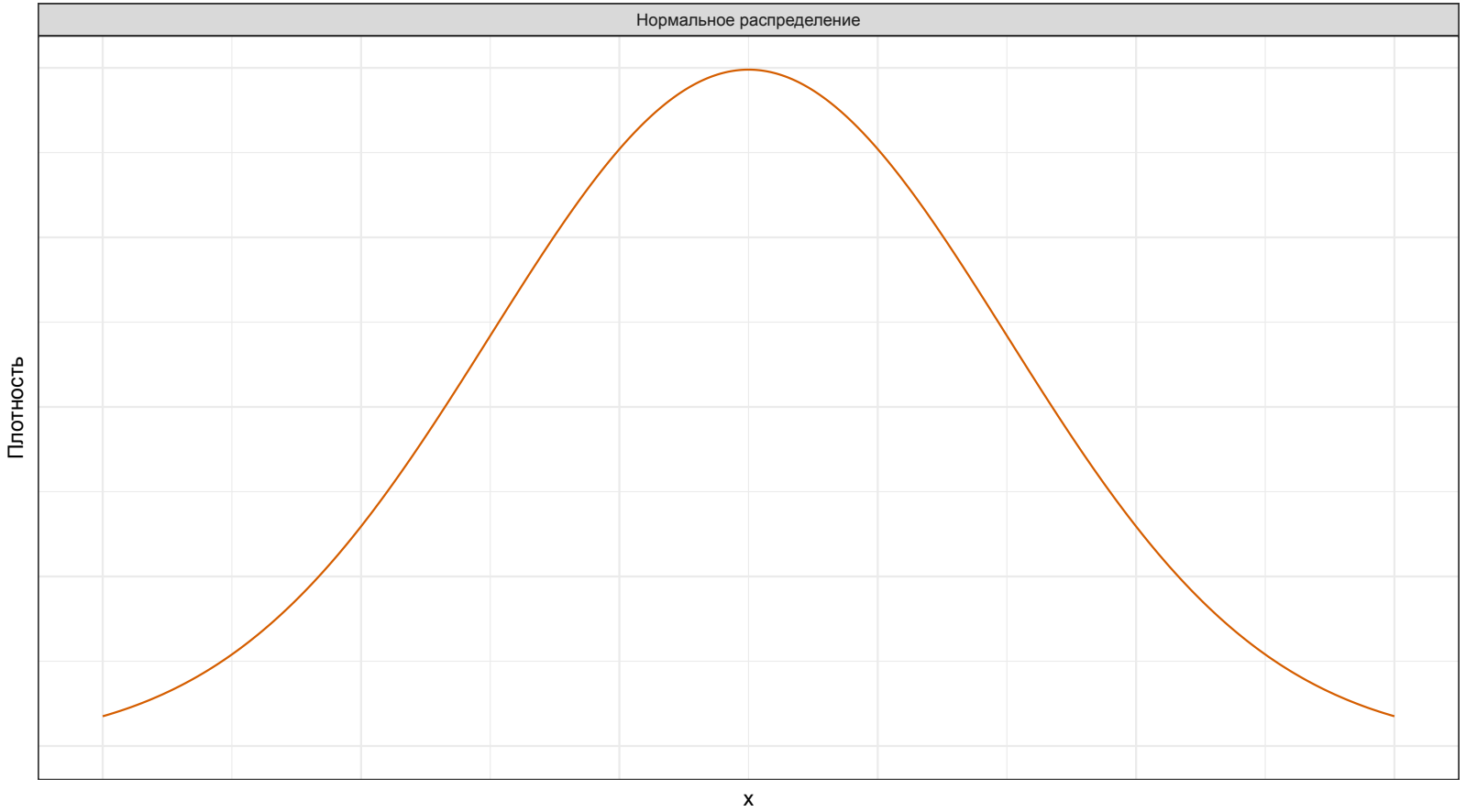
$$x = \{x_1, x_2, \ldots, x_n\}$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$y = (x_1 + x_2 + \ldots + x_n)$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$y = (x_1 + x_2 + \ldots + x_n)$$



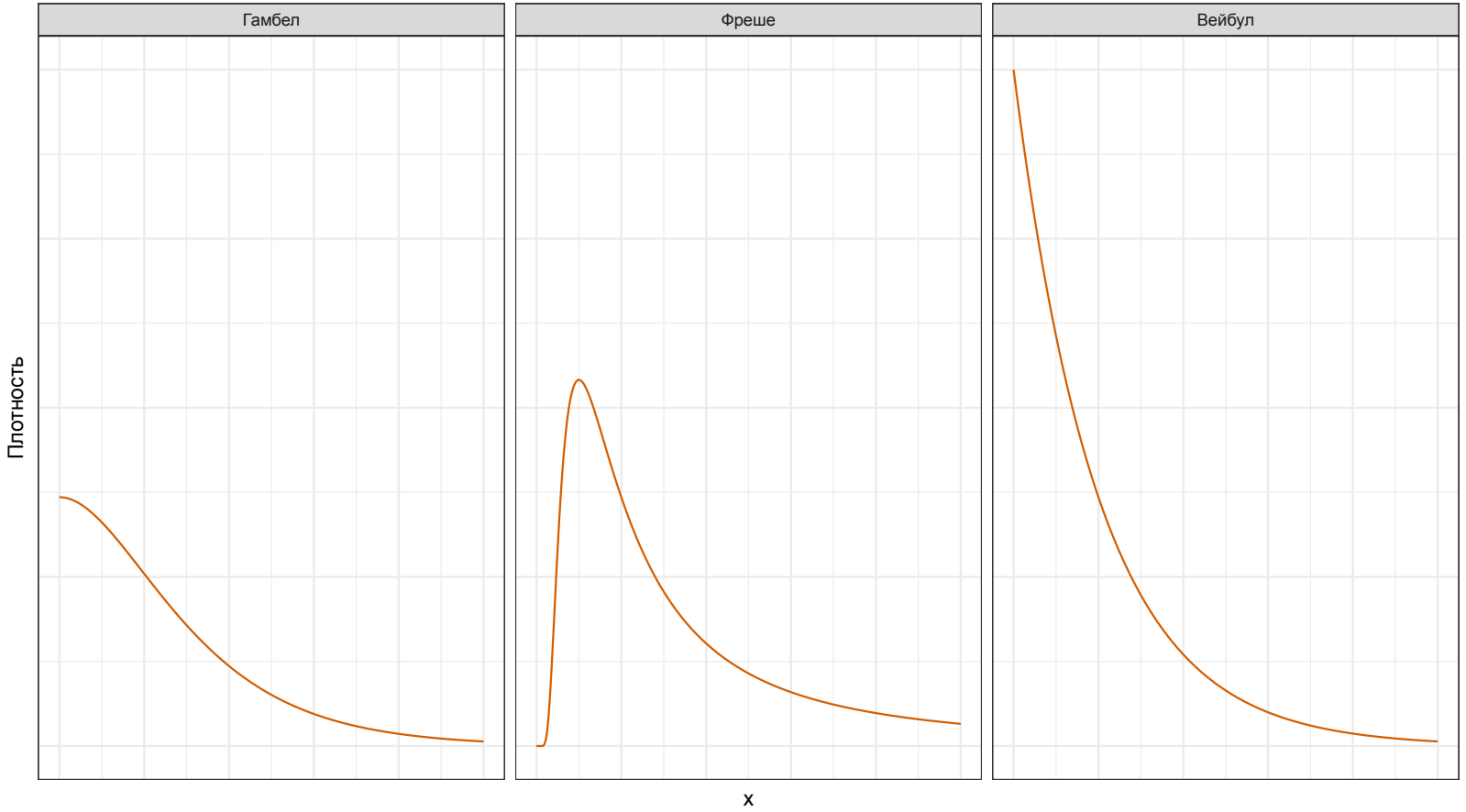
$$x = \{x_1, x_2, \ldots, x_n\}$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$y = \max(x_1, x_2, \ldots, x_n)$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$y = \max(x_1, x_2, \ldots, x_n)$$



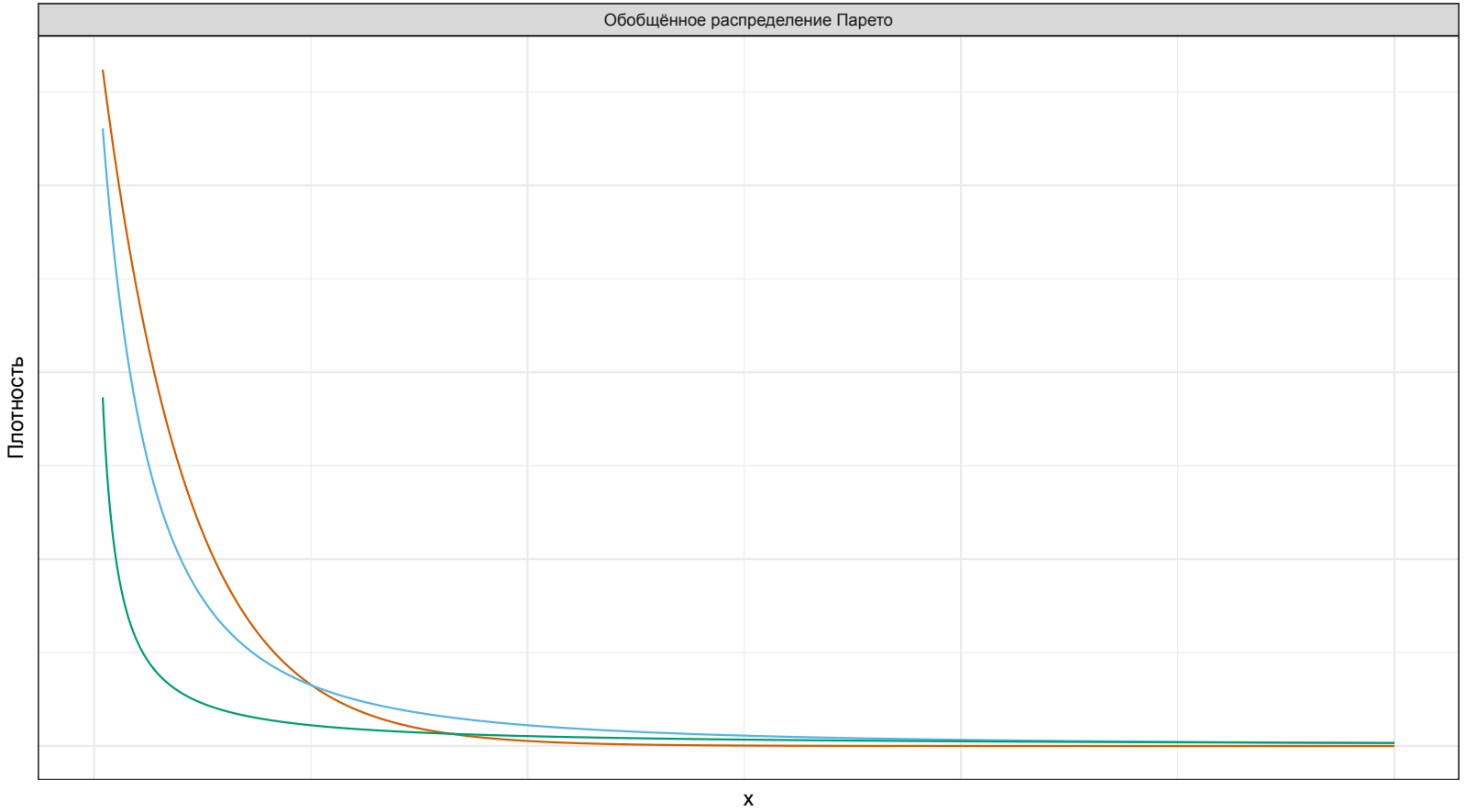
$$x = \{x_1, x_2, \ldots, x_n\}$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$y_u = \{x_i | x_i > u\}$$

$$x = \{x_1, x_2, \ldots, x_n\}$$

$$y_u = \{x_i | x_i > u\}$$



- **Block Maxima approach:**
  - The package climextRemes provides functions for fitting GEV via point process fitting for extremes in climate data, providing return values, return probabilities, and return periods for stationary and nonstationary models.
  - The package evd provides functions for a wide range of univariate distributions. Modelling function allow estimation of parameters for standard univariate extreme value methods.
  - The package evir performs modelling of univariate GEV distributions by maximum likelihood fitting.
  - The package extRemes provides EVDs univariate estimation for block maxima model approache by MLE. It also incorporates a non-stationarity through the parameters of the EVDs and L-moments estimation for the stationary case for the GEV distributions. Finally, it has also Bayes estimation capabilities. A separate package in2extRemes provides some GUI interfaces to extRemes.
  - The package extremeStat includes functions to fit multiple GEV distributions types available in the package lmomco using linear moments to estimate the parameters.
  - The package fExtremes provides univariate data processing and modelling. It includes clustering, block maxima identification and exploratory analysis. The estimation of stationary models for the GEV is provided by maximum likelihood and probability weighted moments.
  - The package ismev provides a collection of three functions to fit the GEV (diagnostic plot, MLE, likelihood profile) and follows the book of Coles (2001).
  - The package lmom has functions to fit probability distributions from GEV distributions to data using the low-order L-moments.
  - The package lmomRFA extends package lmom and implements all the major components for regional frequency analysis using L-moments.
  - The package QRM provides a function to fit GEV
    in Quantitative Risk Management perspective.
  - The package Renext provides various functions to fit the GEV distribution using an aggregated marked POT process.

https://cran.r-project.org/web/views/ExtremeValue.html

- **Peak-Over-Threshold by GPD approach:**
  - The package ercv provides a methodology to fit a generalized Pareto distribution, together with an automatic threshold selection algorithm.
  - The package eva provides Goodness-of-fit tests for selection of r in the r-largest order statistics and threshold selection.
  - The package evd includes univariate estimation for GPD approach by MLE.
  - The package evir performs modelling of univariate GPD by maximum likelihood fitting.
  - The package extRemes provides EVDs univariate estimation for GPD approach by MLE. A non-stationarity through the parameters of the EVDs and L-moments estimation for the stationnary case for the GPD distributions is also included.
  - The package extremeStat includes functions to fit multiple GPD distributions types available in the package lmomco using linear moments to estimate the parameters.
  - The package fExtremes includes the estimation of stationary models for the GPD by maximum likelihood and probability weighted moments.
  - The package ismev provides a collection of three functions to fit the GPD (diagnostic plot, MLE over a range of thresholds, likelihood profile) and follows the book of Coles (2OO1).
  - The package lmom includes functions to fit probability distributions from GPD to data using the low-order L-moments.
  - The package lmomRFA extends package lmom and implements all the major components for regional frequency analysis using L-moments.
  - The package mev provides functions to simulate data from GPD and multiple method to estimate the parameters (optimization, MLE, Bayesian methods and the method used in the ismev package).
  - The package POT provides multiple estimators of the GPD parameters (MLE, L-Moments, method of median, minimum density power divergence). L-moments diagrams and from the properties of a non-homogeneous Poisson process techniques are provided for the selection of the threshold.
  - The package QRM provides functions to fit and graphically assess the fit of the GPD.
  - The package ReIns provides a function to fit the GPD distribution as well as the extended Pareto distribution.
  - The package Renext provides various functions to fit and assess the GPD distribution using an aggregated marked POT process.
  - The package SpatialExtremes provides a function to fit the GPD distribution.
  - The package SpatialExtremes provides different approaches for fitting/selecting the threshold in generalized Pareto distributions. Most of them are based on minimizing the AMSE-criterion or at least by reducing the bias of the assumed GPD-model.
  - The package texmex fit GPD models by using maximum (optionally penalised-)likelihood, or Bayesian estimation, and both classes of models may be fitted with covariates in any/all model parameters.

https://cran.r-project.org/web/views/ExtremeValue.html

POT (Peak Over Threshold)

POT (Peak Over Threshold)

```
tcplot(data, u.range, cmax = FALSE, r = 1,
  ulow = -Inf, rlow = 1, nt = 25, which = 1:npar, conf = 0.95,
  lty = 1, lwd = 1, type = "b", cilty = 1, ask = nb.fig <
  length(which) && dev.interactive(), ...)

mrlplot(data, u.range, main, xlab, ylab,
  nt = max(100, length(data)), lty = rep(1,3),
  col = c('grey', 'black', 'grey'), conf = 0.95, lwd = c(1, 1.5, 1), ...)

lmomplot(data, u.range, nt = max(50, length(data)), identify = TRUE, ...)

exiplot(data, u.range, tim.cond = 1, n.u = 50, xlab, ylab, ...)

diplot(data, u.range, main, xlab, ylab, nt = max(200, nrow(data)), conf=0.95, ...)
```

https://cran.r-project.org/web/packages/POT/index.html

POT (Peak Over Threshold)

`Fit the GP Distribution`

Fitting a GPD to Peaks Over a Threshold

Description

Maximum (Penalized) Likelihood, Unbiased Probability Weighted Moments,Biased Probability Weighted Moments, Moments, Pickands', Minimum Density Power Divergence, Medians, Likelihood Moment and Maximum Goodness-of-Fit Estimators to fit Peaks Over a Threshold to a GP distribution.

Usage

```
fitgpd(data, threshold, est = "mle", ...)
```

Arguments

`data`	A numeric vector.
`threshold`	A numeric value giving the threshold for the GPD. The `'mle'` estimator allows varying threshold; so that threshold could be for this case a numeric vector. Be careful, varying thresholds are used cyclically if length doesn't match with `data`.
`est`	A string giving the names of the estimator. It can be `'mle'` (the default), `'mple'`, `'moments'`, `'pwmu'`, `'pwmb'`, `'mdpd'`, `'med'`, `'pickands'`, `'lme'` and `'mgf'` for the maximum likelihood, maximum penalized likelihood, moments, unbiased probability weighted moments, biased probability weigthed moments, minimum density power divergence, medians, Pickands', likelihood moment and maximum goodness-of-fit estimators respectively.
`...`	Other optional arguments to be passed to the `optim` function, allow hand fixed parameters (only for the `mle`, `mple` and `mgf` estimators) or passed several options to specific estimators - see the Note section.

https://cran.r-project.org/web/packages/POT/index.html

Для использования
теории экстремальных значений
нужно подумать

Заключение

Особенности перформанс-распределений

- Большой разброс значений
- Тяжелые хвосты и экстремальные выбросы
- Мультимодальность
- Дискретизация
- Асимметрия

Решения

- Непараметрическая статистика

- Непараметрическая статистика
- Робастная статистика

- Непараметрическая статистика
- Робастная статистика
- Поддержка мультимодальности

- Непараметрическая статистика
- Робастная статистика
- Поддержка мультимодальности
- Поддержка дискретизации

- Непараметрическая статистика
- Робастная статистика
- Поддержка мультимодальности
- Поддержка дискретизации
- Фокус на конечных выборках

- Непараметрическая статистика
- Робастная статистика
- Поддержка мультимодальности
- Поддержка дискретизации
- Фокус на конечных выборках

Perfolizer

Performance analysis toolkit

Нелинейность умножения процентов

$$0.95^{20}$$

$$0.95^{20} \approx 35.8\%$$

$$0.95^{20} \approx 35.8\%$$

$$0.96^{20} \approx 44.2\%$$

$$0.95^{20} \approx 35.8\%$$

$$0.96^{20} \approx 44.2\%$$

$$0.97^{20} \approx 54.4\%$$

$$0.95^{20} \approx 35.8\%$$

$$0.96^{20} \approx 44.2\%$$

$$0.97^{20} \approx 54.4\%$$

$$0.98^{20} \approx 66.8\%$$

$$0.95^{20} \approx 35.8\%$$

$$0.96^{20} \approx 44.2\%$$

$$0.97^{20} \approx 54.4\%$$

$$0.98^{20} \approx 66.8\%$$

$$0.99^{20} \approx 81.8\%$$

$$0.95^{20} \approx 35.8\%$$

$$0.96^{20} \approx 44.2\%$$

$$0.97^{20} \approx 54.4\%$$

$$0.98^{20} \approx 66.8\%$$

$$0.99^{20} \approx 81.8\%$$

$$0.995^{20} \approx 90.5\%$$

$$0.95^{20} \approx 35.8\%$$

$$0.96^{20} \approx 44.2\%$$

$$0.97^{20} \approx 54.4\%$$

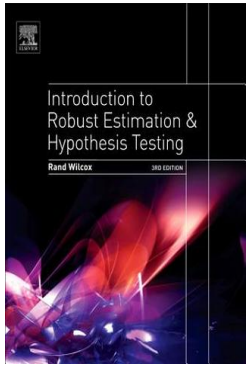
$$0.98^{20} \approx 66.8\%$$

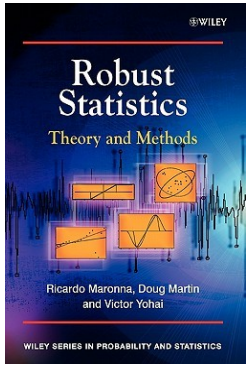
$$0.99^{20} \approx 81.8\%$$

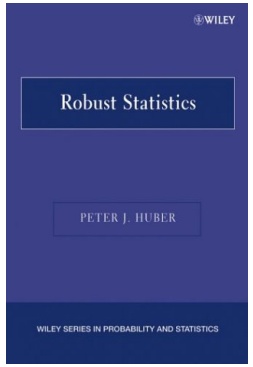
$$0.995^{20} \approx 90.5\%$$

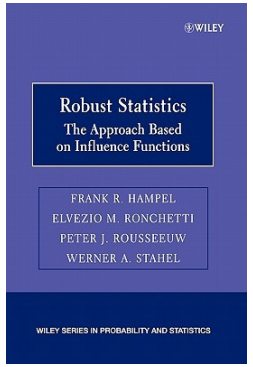
$$0.999^{20} \approx 98.0\%$$

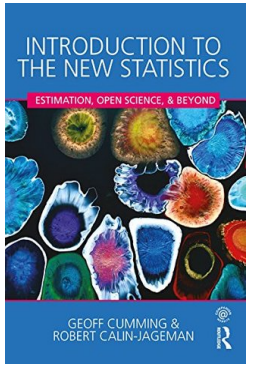
Хорошие книжки

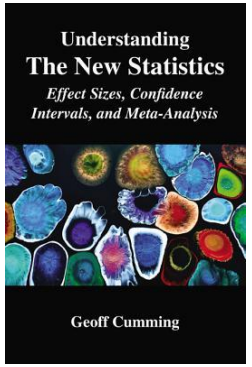


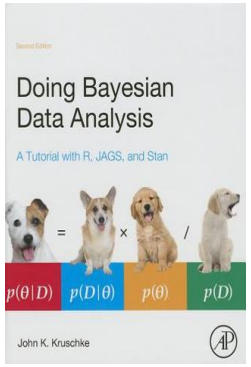


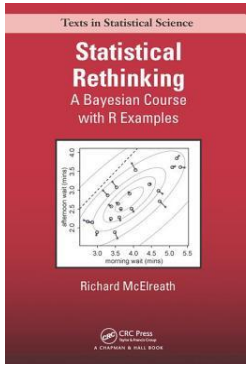


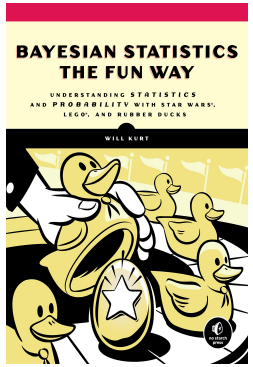


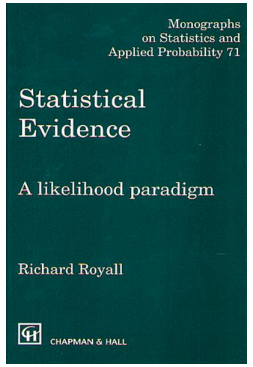


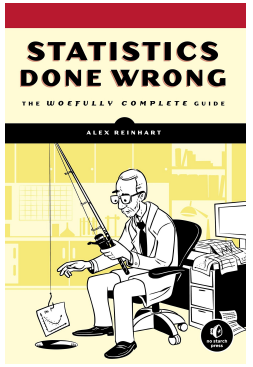


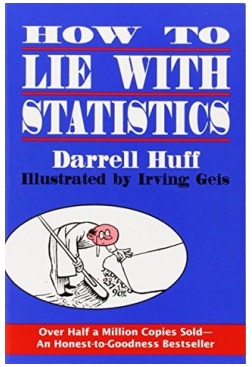


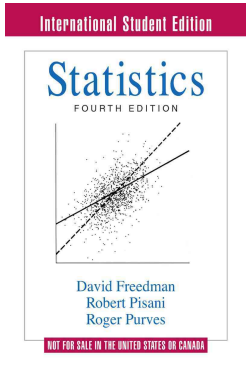


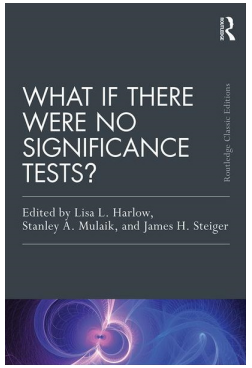


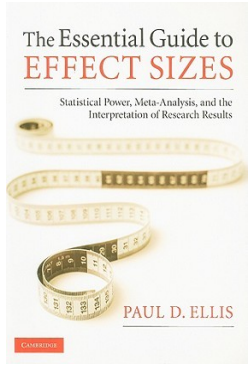


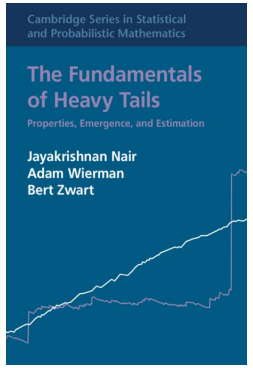


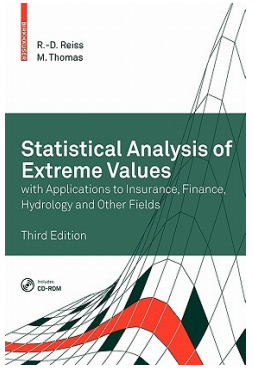


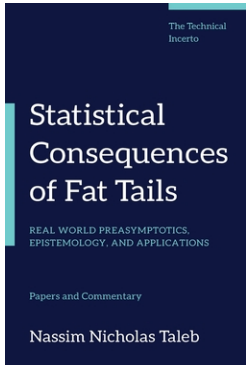


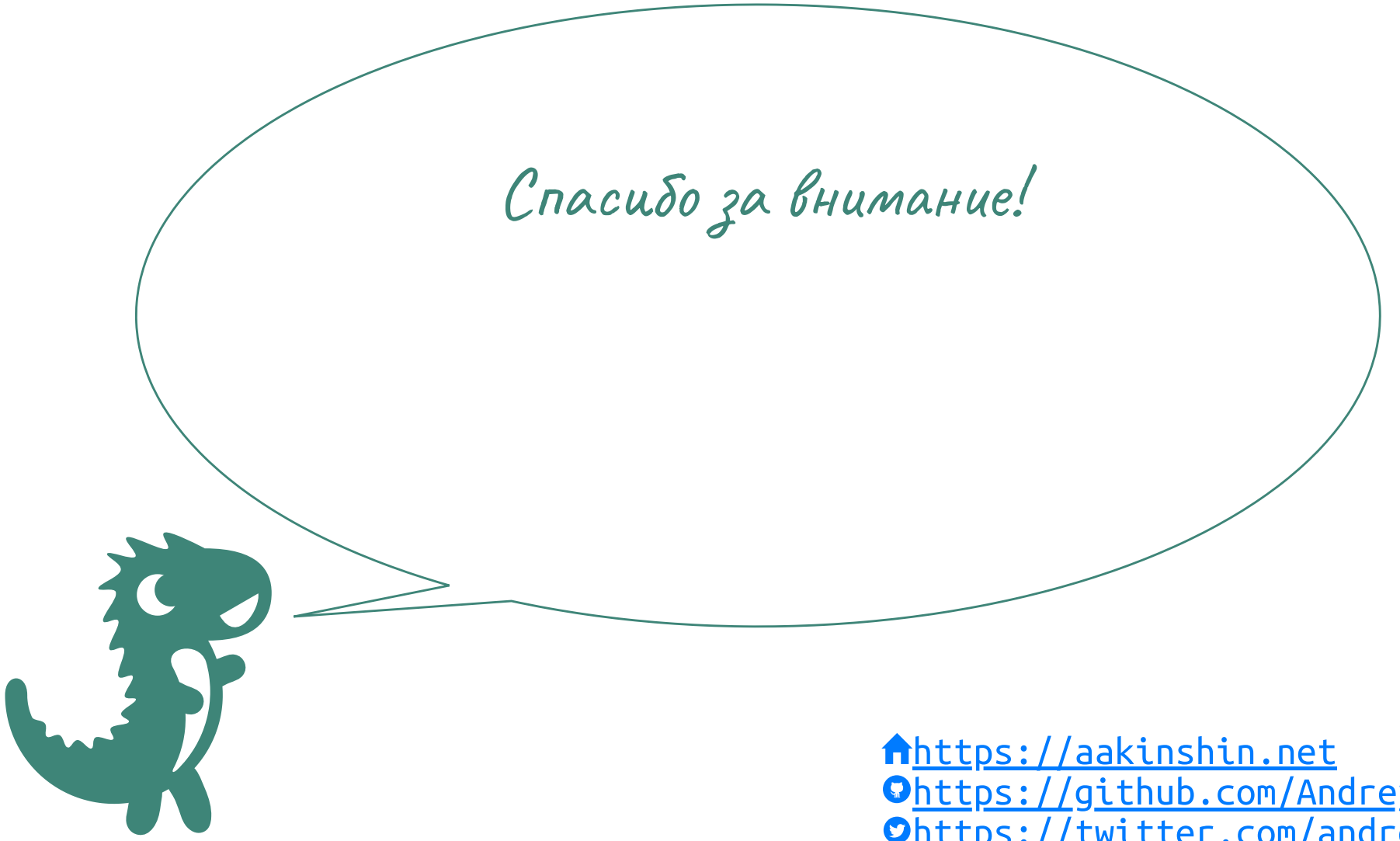












https://aakinshin.net
https://github.com/AndreyAkinshin
https://twitter.com/andrey_akinshin
mailto:andrey.akinshin@gmail.com