Поиск аномалий

HBOS и ECOD

Обо мне

- Старший специалист по машинному обучению
- Deep learning engineer
- NLP, CV, anomaly detection
- Open source contributor
- Выпускник и амбассадор Яндекс Практикума
- Выпускник DLS ФПМИ МФТИ

Аномалии

Применение

Свойства

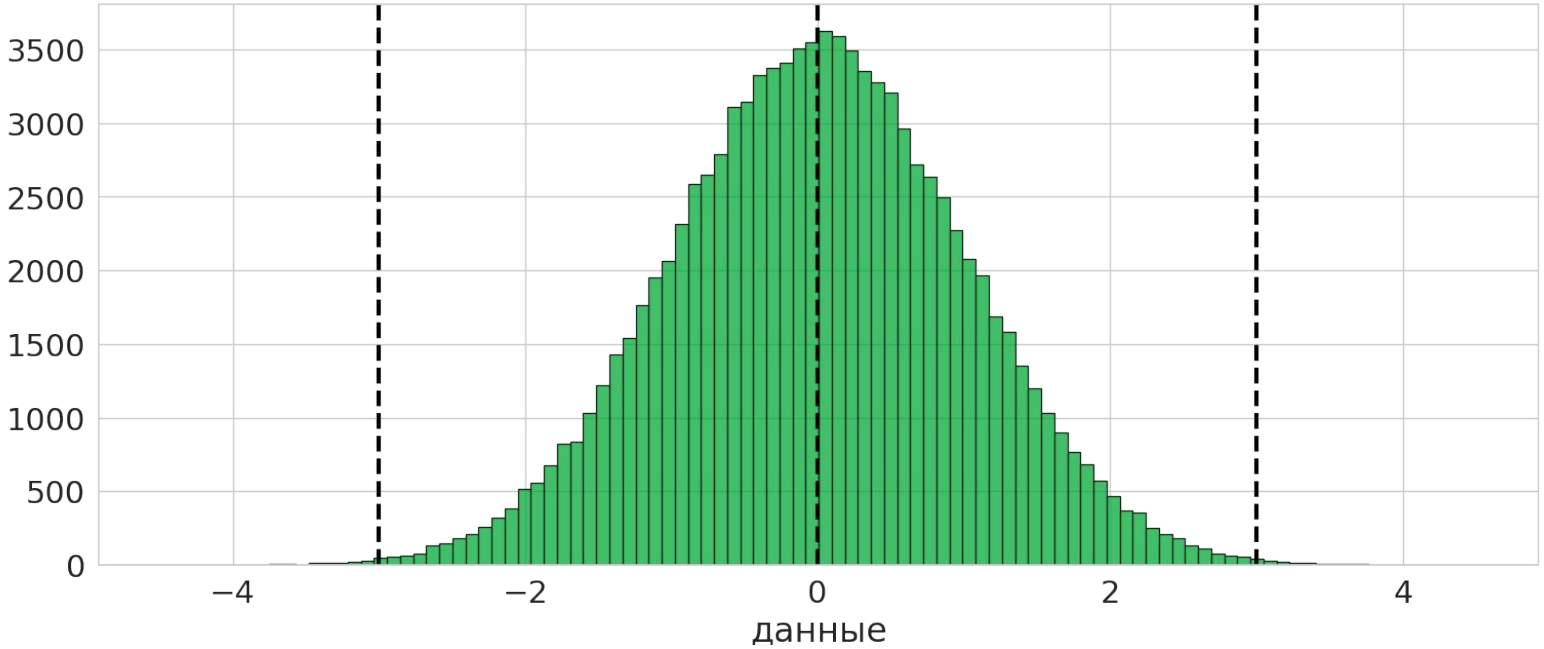
Методы

[1] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): a fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012.

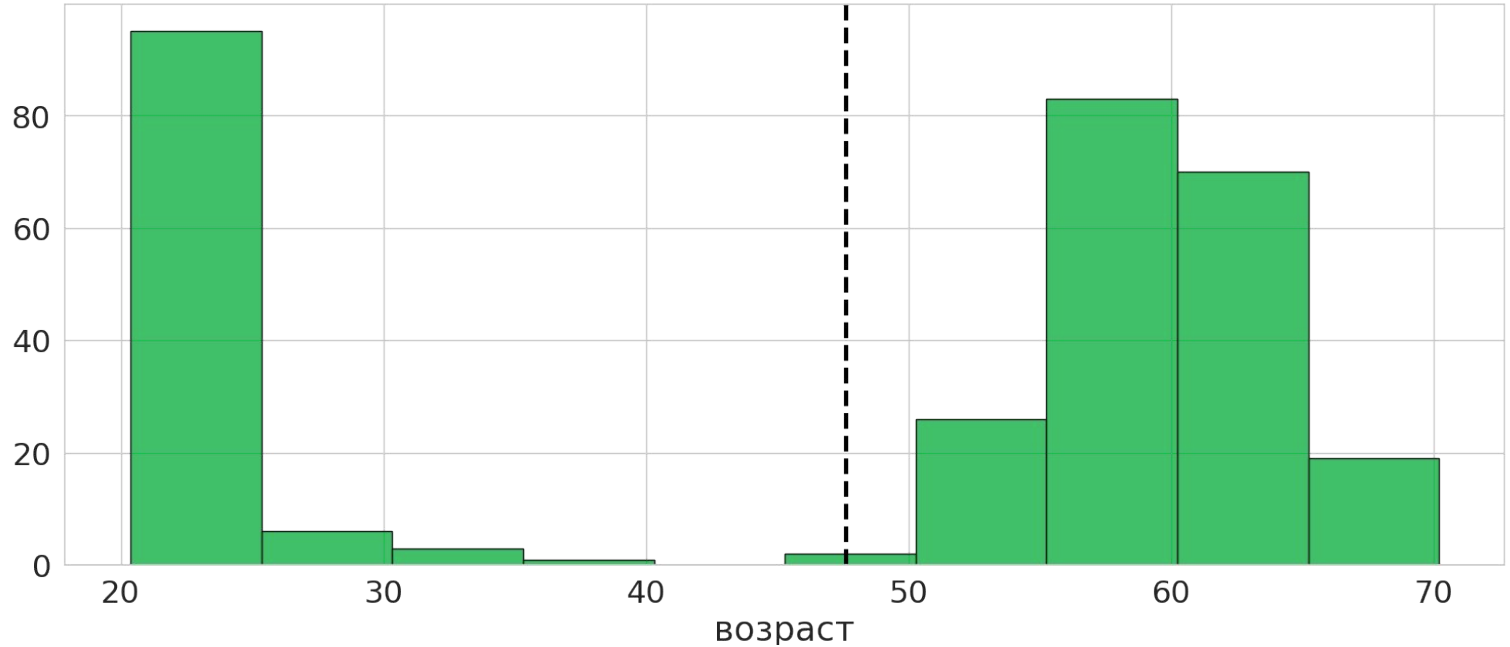
[2] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and H. George Chen. Ecod: unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.


```
$ pip install pyod
```

Три сигмы



Пионеры и пенсионеры



Гистограмма в таблице

(20.4, 25.4]	95			
(25.4, 30.3]	6			
(30.3, 35.3]	3			
(35.3, 40.3]	1			
(40.3, 45.3]	0			
(45.3, 50.3]	2			
(50.3, 55.2]	26			
(55.2, 60.2]	83			
(60.2, 65.2]	70			
(65.2, 70.2]	19			

Плотность вероятности

(20.4, 25.4]	95	**95 / 305 / 5.0**		
(25.4, 30.3]	6	**6 / 305 / 5.0**		
(30.3, 35.3]	3	**3 / 305 / 5.0**		
(35.3, 40.3]	1	**1 / 305 / 5.0**		
(40.3, 45.3]	0	**0 / 305 / 5.0**		
(45.3, 50.3]	2	**2 / 305 / 5.0**		
(50.3, 55.2]	26	**26 / 305 / 5.0**		
(55.2, 60.2]	83	**83 / 305 / 5.0**		
(60.2, 65.2]	70	**70 / 305 / 5.0**		
(65.2, 70.2]	19	**19 / 305 / 5.0**		

Плотность вероятности

(20.4, 25.4]	95	**0.063**		
(25.4, 30.3]	6	**0.004**		
(30.3, 35.3]	3	**0.002**		
(35.3, 40.3]	1	**0.001**		
(40.3, 45.3]	0	**0.000**		
(45.3, 50.3]	2	**0.001**		
(50.3, 55.2]	26	**0.017**		
(55.2, 60.2]	83	**0.055**		
(60.2, 65.2]	70	**0.046**		
(65.2, 70.2]	19	**0.013**		

~~Перемножаем~~

(20.4, 25.4]	95	0.063	**~~0.063 * …~~**	
(25.4, 30.3]	6	0.004		
(30.3, 35.3]	3	0.002		
(35.3, 40.3]	1	0.001		
(40.3, 45.3]	0	0.000		
(45.3, 50.3]	2	0.001		
(50.3, 55.2]	26	0.017		
(55.2, 60.2]	83	0.055		
(60.2, 65.2]	70	0.046		
(65.2, 70.2]	19	0.013		

~~Перемножаем~~ $\log_2(x\ y) = \log_2(x) + \log_2(y)$

(20.4, 25.4]	95	0.063	~~**0.063 * …**~~	
(25.4, 30.3]	6	0.004		
(30.3, 35.3]	3	0.002		
(35.3, 40.3]	1	0.001		
(40.3, 45.3]	0	0.000		
(45.3, 50.3]	2	0.001		
(50.3, 55.2]	26	0.017		
(55.2, 60.2]	83	0.055		
(60.2, 65.2]	70	0.046		
(65.2, 70.2]	19	0.013		

Складываем логарифмы

(20.4, 25.4]	95	0.063	**$\log_2(0.063)$ + …**	
(25.4, 30.3]	6	0.004	**$\log_2(0.004)$ + …**	
(30.3, 35.3]	3	0.002	**$\log_2(0.002)$ + …**	
(35.3, 40.3]	1	0.001	**$\log_2(0.001)$ + …**	
(40.3, 45.3]	0	0.000	**$\log_2(0.000)$ + …**	
(45.3, 50.3]	2	0.001	**$\log_2(0.001)$ + …**	
(50.3, 55.2]	26	0.017	**$\log_2(0.017)$ + …**	
(55.2, 60.2]	83	0.055	**$\log_2(0.055)$ + …**	
(60.2, 65.2]	70	0.046	**$\log_2(0.046)$ + …**	
(65.2, 70.2]	19	0.013	**$\log_2(0.013)$ + …**	

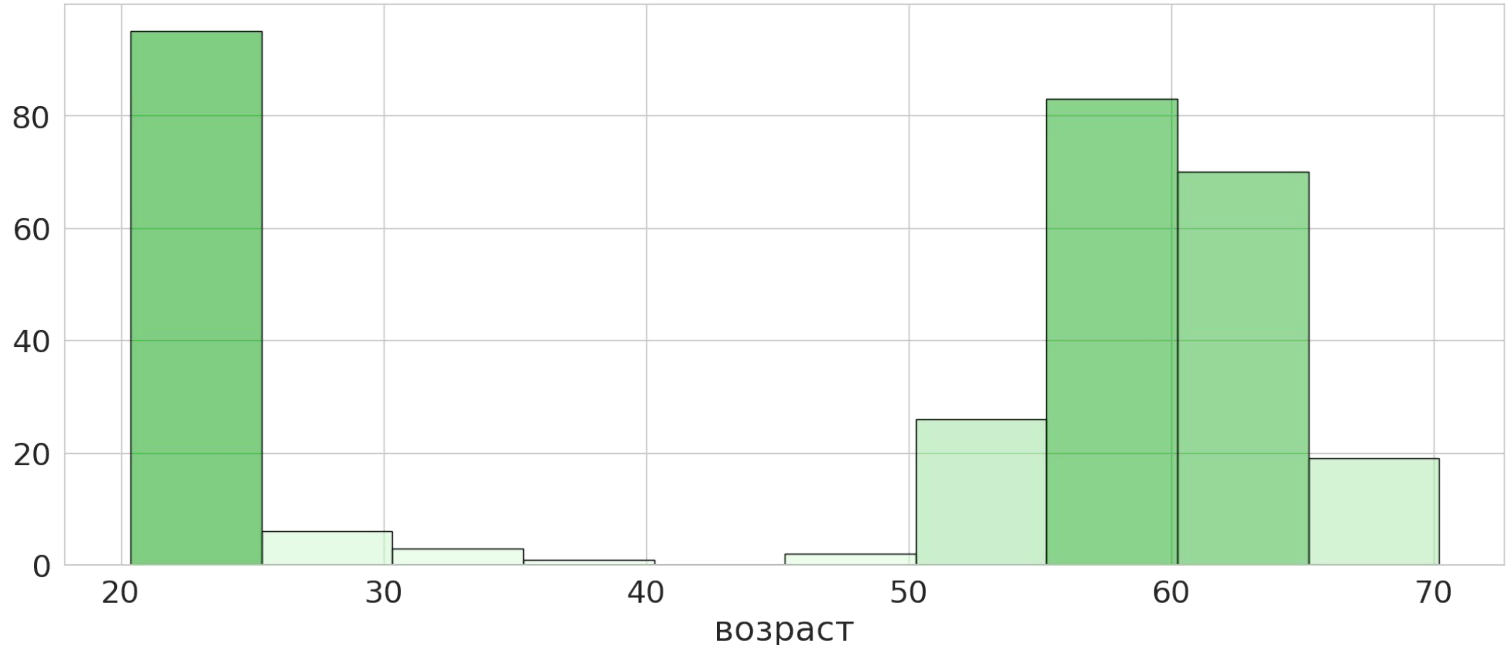
Складываем логарифмы

(20.4, 25.4]	95	0.063	**-2.621**	
(25.4, 30.3]	6	0.004	**-3.266**	
(30.3, 35.3]	3	0.002	**-3.294**	
(35.3, 40.3]	1	0.001	**-3.312**	
(40.3, 45.3]	0	0.000	**-3.322**	
(45.3, 50.3]	2	0.001	**-3.303**	
(50.3, 55.2]	26	0.017	**-3.094**	
(55.2, 60.2]	83	0.055	**-2.693**	
(60.2, 65.2]	70	0.046	**-2.775**	
(65.2, 70.2]	19	0.013	**-3.152**	

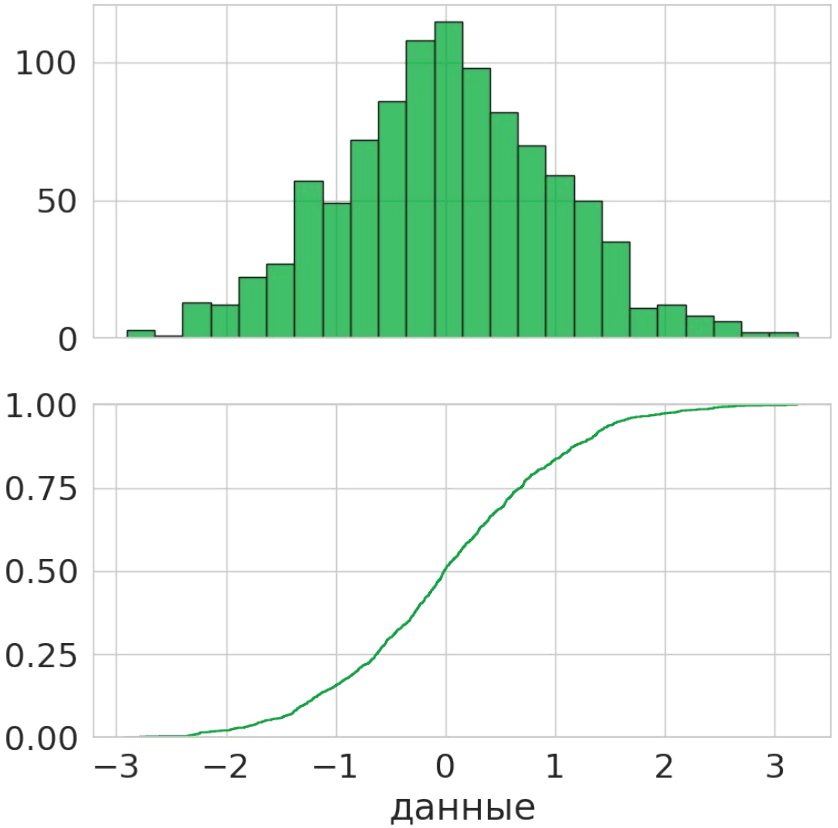
Меняем знак

(20.4, 25.4]	**95**	0.063	-2.621	**2.621**
(25.4, 30.3]	**6**	0.004	-3.266	**3.266**
(30.3, 35.3]	**3**	0.002	-3.294	**3.294**
(35.3, 40.3]	**1**	0.001	-3.312	**3.312**
(40.3, 45.3]	**0**	0.000	-3.322	**3.322**
(45.3, 50.3]	**2**	0.001	-3.303	**3.303**
(50.3, 55.2]	**26**	0.017	-3.094	**3.094**
(55.2, 60.2]	**83**	0.055	-2.693	**2.693**
(60.2, 65.2]	**70**	0.046	-2.775	**2.775**
(65.2, 70.2]	**19**	0.013	-3.152	**3.152**

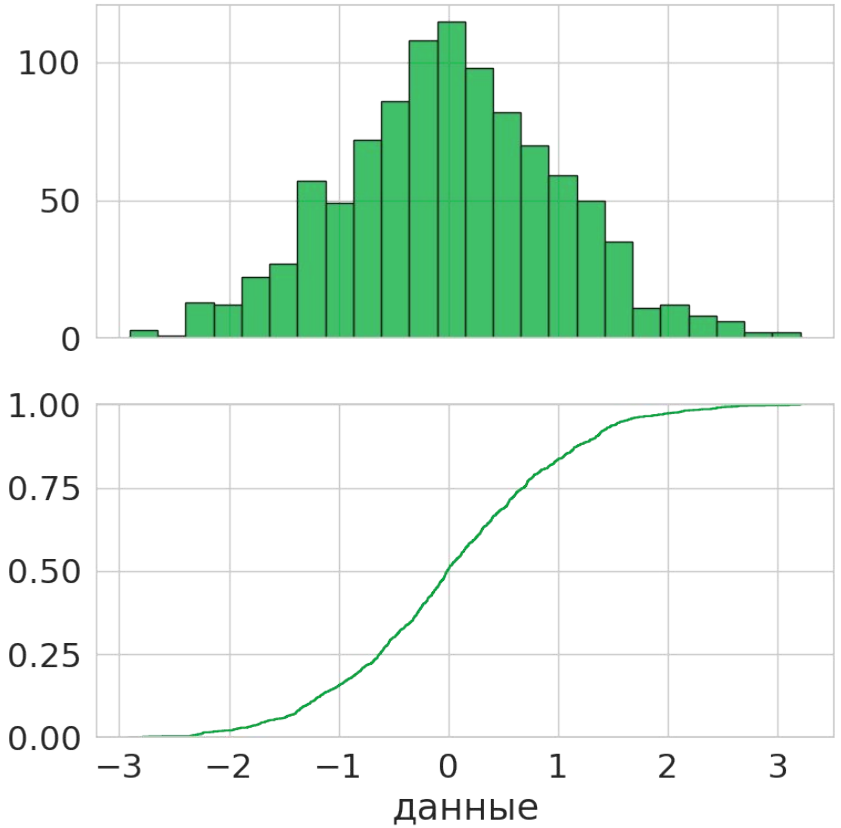
Histogram-Based Outlier Score [1]



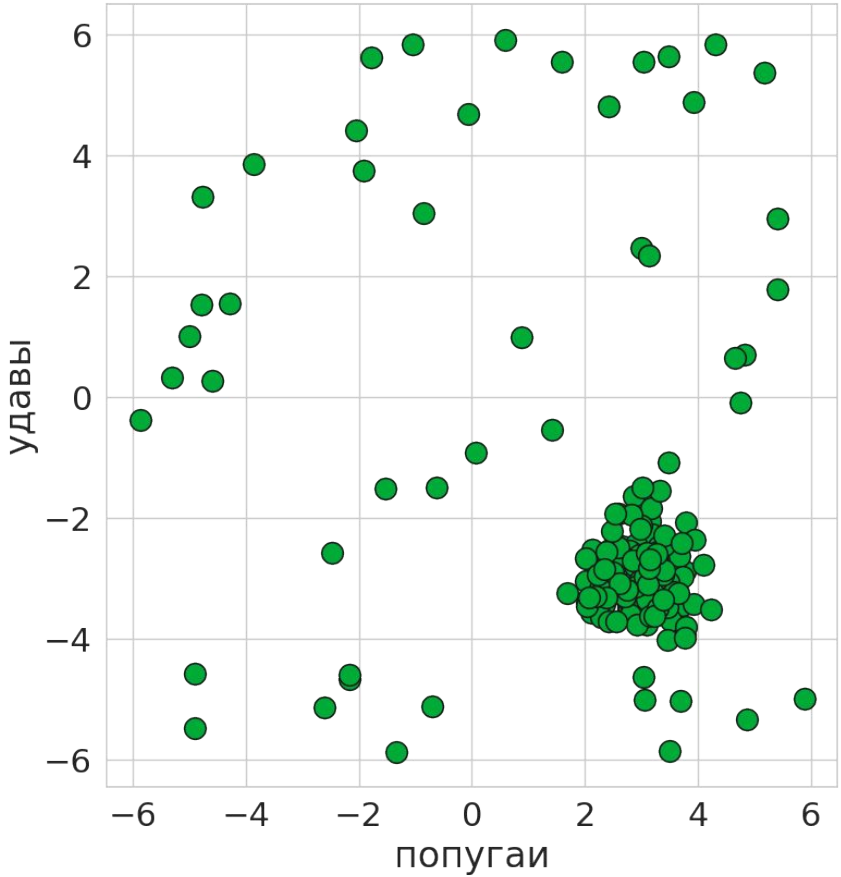
Выборочная эмпирическая функция распределения

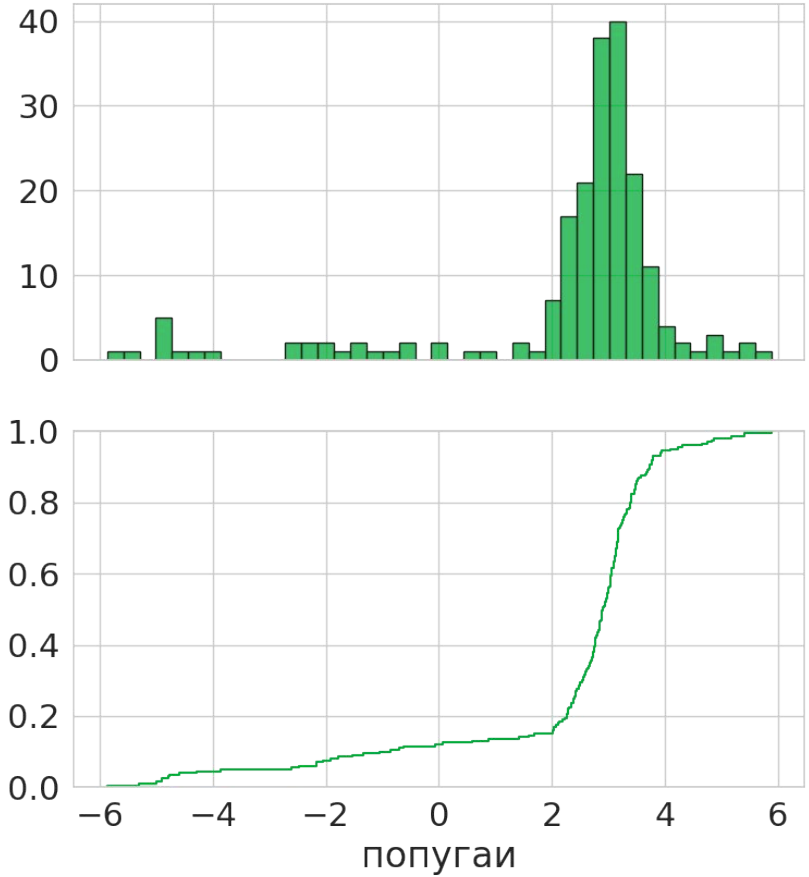


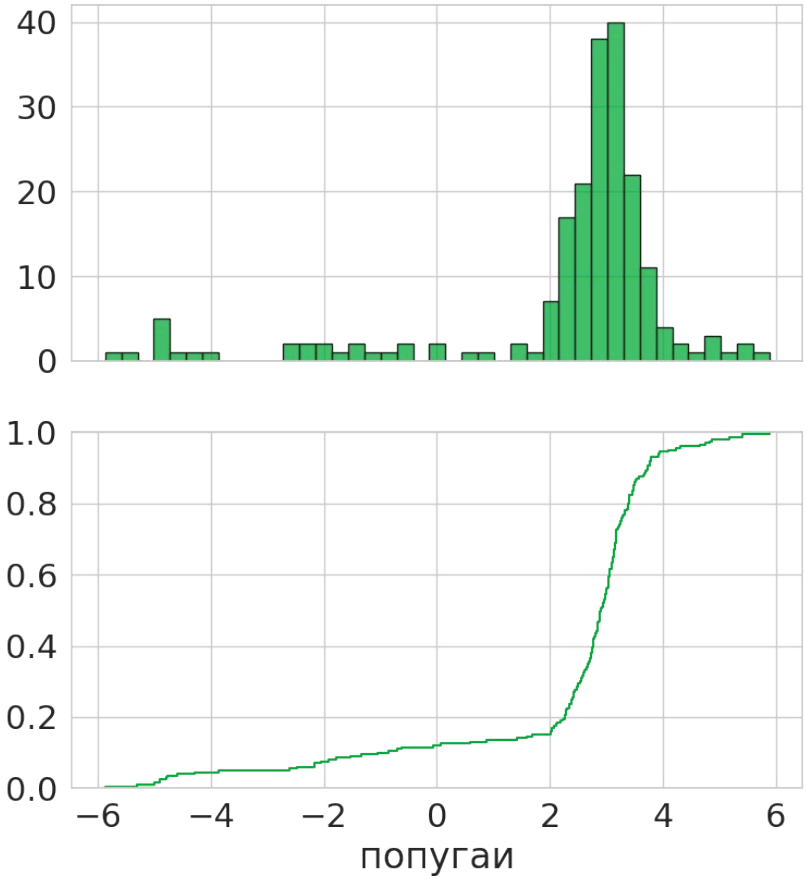
```
>>> x = -1
>>> (df['данные'] <= x).mean()
np.float64(0.155)
```

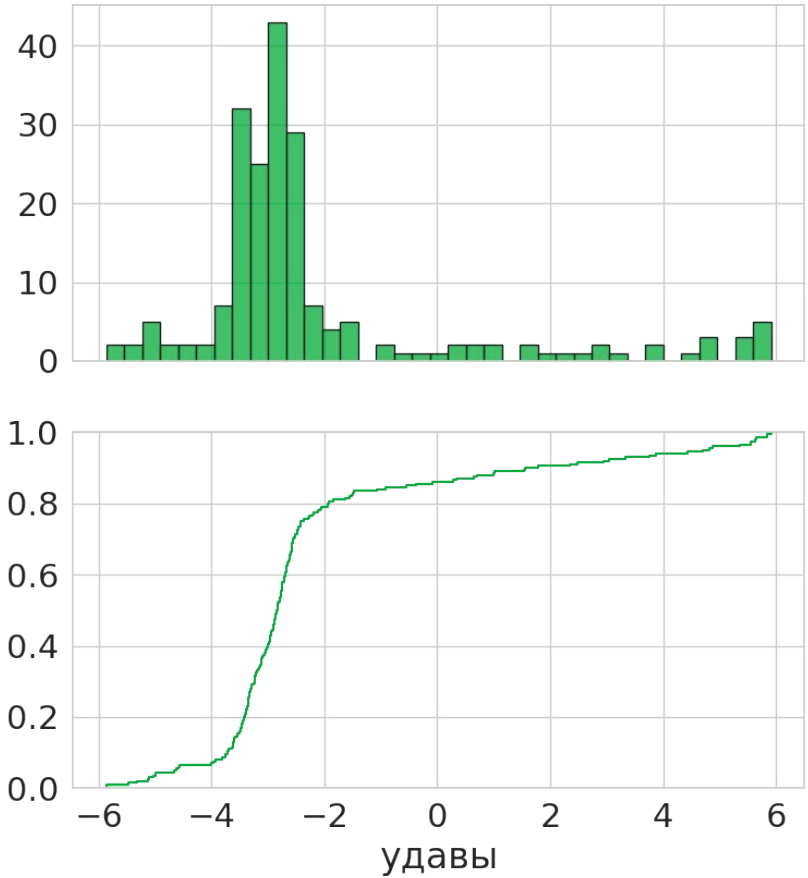


Данные




40
30
20
10
0
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
попугаи


40
30
20
10
0
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
попугаи


40
30
20
10
0
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
удавы

ЭФР и 1 - ЭФР

Удавы	ЭФР	1 - ЭФР			
-2.87	**0.48**	**0.52**			
-2.83	**0.51**	**0.50**			
-3.34	**0.26**	**0.75**			
-2.88	**0.48**	**0.53**			
…	**…**	**…**			
-0.55	**0.85**	**0.16**			
1.53	**0.90**	**0.11**			
1.01	**0.89**	**0.12**			
4.81	**0.96**	**0.05**			

Negative log probs

Удавы	**ЭФР**	**1 - ЭФР**	**- log(ЭФР)**	**-log(1 - ЭФР)**	
-2.87	0.48	0.52	**0.73** + …	**0.64** + …	
-2.83	0.51	0.50	**0.67** + …	**0.70** + …	
-3.34	0.26	0.75	**1.37** + …	**0.29** + …	
-2.88	0.48	0.53	**0.74** + …	**0.63** + …	
…	…	…	**…**	**…**	
-0.55	0.85	0.16	**0.16** + …	**1.86** + …	
1.53	0.90	0.11	**0.11** + …	**2.21** + …	
1.01	0.89	0.12	**0.12** + …	**2.16** + …	
4.81	0.96	0.05	**0.05** + …	**3.00** + …	

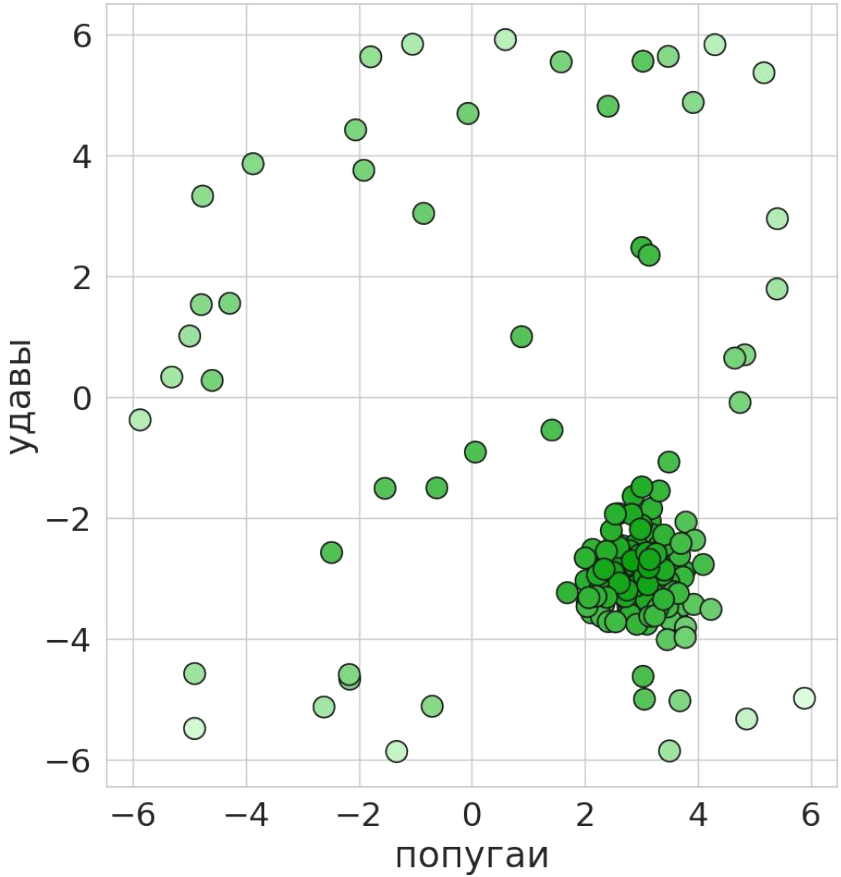
Negative log probs

Удавы	ЭФР	1 - ЭФР	- log(ЭФР)	-log(1 - ЭФР)	
-2.87	0.48	0.52	**1.03**	**1.99**	
-2.83	0.51	0.50	**1.34**	**1.42**	
-3.34	0.26	0.75	**1.61**	**1.78**	
-2.88	0.48	0.53	**0.84**	**2.99**	
…	…	…	**…**	**…**	
-0.55	0.85	0.16	**2.13**	**2.01**	
1.53	0.90	0.11	**3.62**	**2.23**	
1.01	0.89	0.12	**4.32**	**2.17**	
4.81	0.96	0.05	**1.39**	**3.29**	

ECOD

Удавы	ЭФР	1 - ЭФР	- log(ЭФР)	-log(1 - ЭФР)	max
-2.87	0.48	0.52	1.03	1.99	**1.99**
-2.83	0.51	0.50	1.34	1.42	**1.42**
-3.34	0.26	0.75	1.61	1.78	**1.78**
-2.88	0.48	0.53	0.84	2.99	**2.99**
…	…	…	…	…	**…**
-0.55	0.85	0.16	2.13	2.01	**2.13**
1.53	0.90	0.11	3.62	2.23	**3.62**
1.01	0.89	0.12	4.32	2.17	**4.32**
4.81	0.96	0.05	1.39	3.29	**3.29**

Empirical Cumulative Outlier Detection [2]



Резюме

Вопросы?