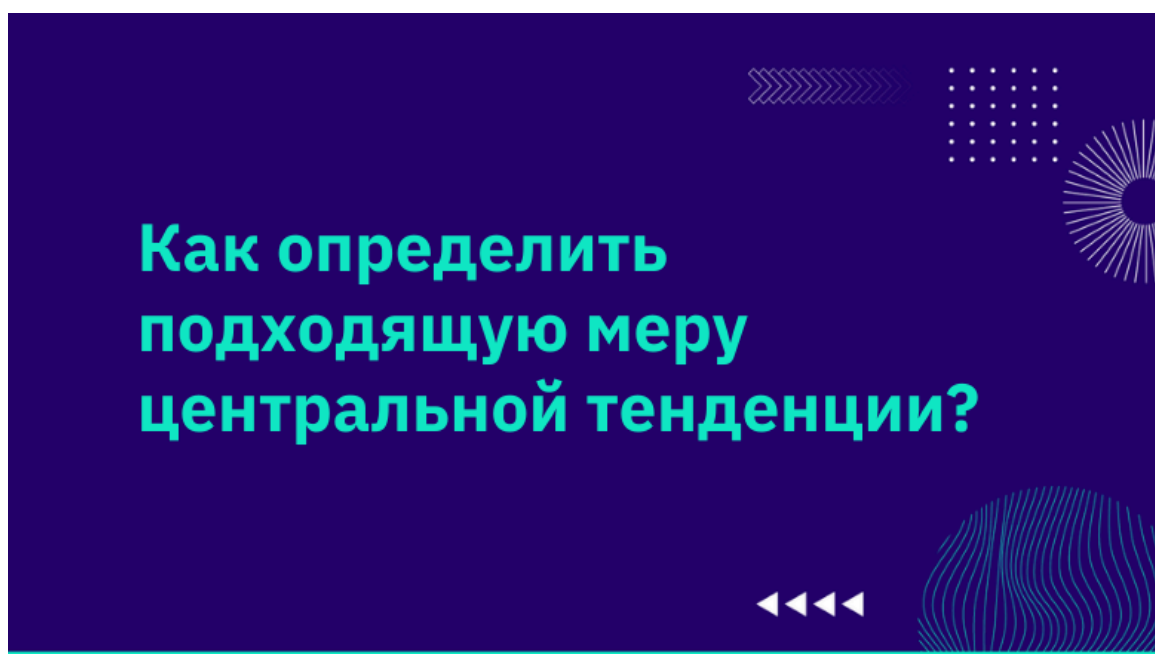


# Как определить подходящую меру центральной тенденции?

rikki\_tikki

6 мин

17K



*Мера центральной тенденции (measure of central tendency) представляет из себя статистическую величину, которая характеризует целый набор данных одним единственным числом. Ее также называют мерой центрального расположения (measure of central location). Она описывает, как выглядит приблизительный центр набора данных.*

Но сам по себе термин “центр” может подразумевать немного разные значения в зависимости от конкретной ситуации. Вы можете считать “центром” среднее арифметическое. Вы также можете назвать “центром” данные, которые просто находятся в середине вашей

выборки. А еще вы можете рассматривать в качестве “центра” данные, которые повторяются чаще всего. Все эти центры по-своему характеризуют ваши данные.

Поскольку человеческое понимание “центра” может разниться, статистика позаботилась определить каждый вариант. Таким образом мы имеем следующие общепринятые меры центральной тенденции:

1. Среднее арифметическое.
2. Медиана.
3. Мода.

В этой статье я расскажу, каким образом распределение вашего набора данных играет роль в выборе подходящей меры центральной тенденции. А объяснять я буду это на примере реальных наборов данных.

## **1. Среднее арифметическое**

Среднее арифметическое — это среднее значение всех элементов в наборе данных. Оно рассчитывается как сумма всех значений, деленная на общее количество значений.

Среднее арифметическое = сумма всех значений / общее количество значений

### **Когда следует использовать среднее арифметическое?**

Среднее арифметическое лучше всего использовать для описания данных, которые имеют нормальное распределение. Нормальное распределение — это когда построив график по “значениям” и их “частоте” (количеству появлений каждого значения в наборе данных), вы получаете кривую, по форме напоминающую колокол. Центр этой кривой совпадает со средним арифметическим.

## Пример — набор данных с длинами крыльев комнатной мухи

В качестве примера я буду использовать реальный набор данных — это набор данных с [длинами крыльев комнатной мухи](#), который естественным образом имеет нормальное распределение.

Источник набора данных: [Sokal, R.R. and F.J. Rohlf, 1968. *Biometry*, Freeman Publishing Co., p 109. Original data from Sokal, R.R. and P.E. Hunter. 1955. A morphometric analysis of DDT-resistant and non-resistant housefly strains *Ann. Entomol. Soc. Amer.* 48: 499-507.]

Набор данных содержит длины крыльев комнатной мухи в миллиметрах. В нем 100 элементов.

housefly\_wing\_length - Notepad

File Edit Format View Help

36

37

38

38

39

39

40

40

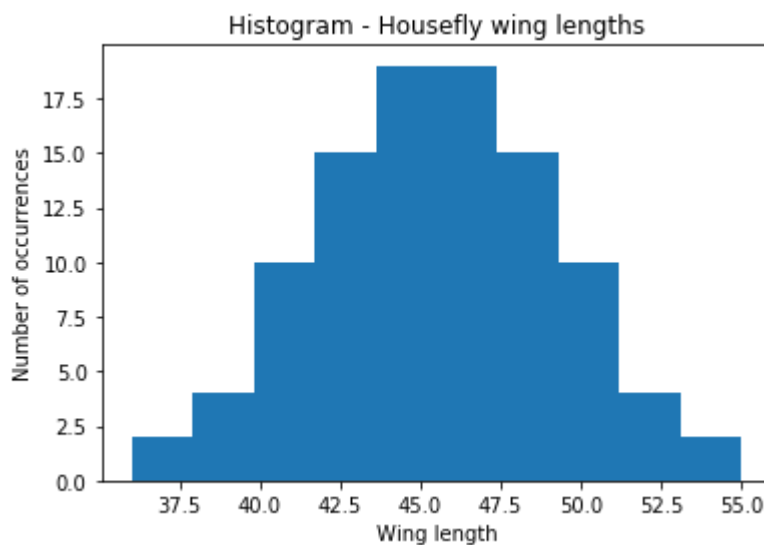
40

Часть набора данных с длинами крыльев комнатной мухи

Я построил гистограмму (по “значениям” и “количествам повторений этих значений”) этих данных, которую вы можете наблюдать ниже. Если мы проведем по внешним краям столбцов плавную линию, то она образует колоколообразную кривую. Вычислив среднее арифметическое значение этих данных, мы получим 45,5. А теперь давайте поищем на приведенном ниже графике полученное значение 45,5. Он находится прямо по середине.

Колоколообразная кривая со средним значением в центре дает нам четкое понимание, что этот набор данных имеет нормальное распределение.

```
import numpy as np
import matplotlib.pyplot as plt
data_housefly = np.loadtxt("housefly_wing_length.txt")
plt.hist(data_housefly)
plt.xlabel("Wing length")
plt.ylabel("Number of occurrences")
plt.title("Histogram - Housefly wing lengths")
plt.show()
```



### Длина крыла комнатной мухи – гистограмма

Это хороший пример, наглядно демонстрирующий, что для нормально распределенных данных имеет смысл использовать “среднее арифметическое” как меру центральной тенденции.

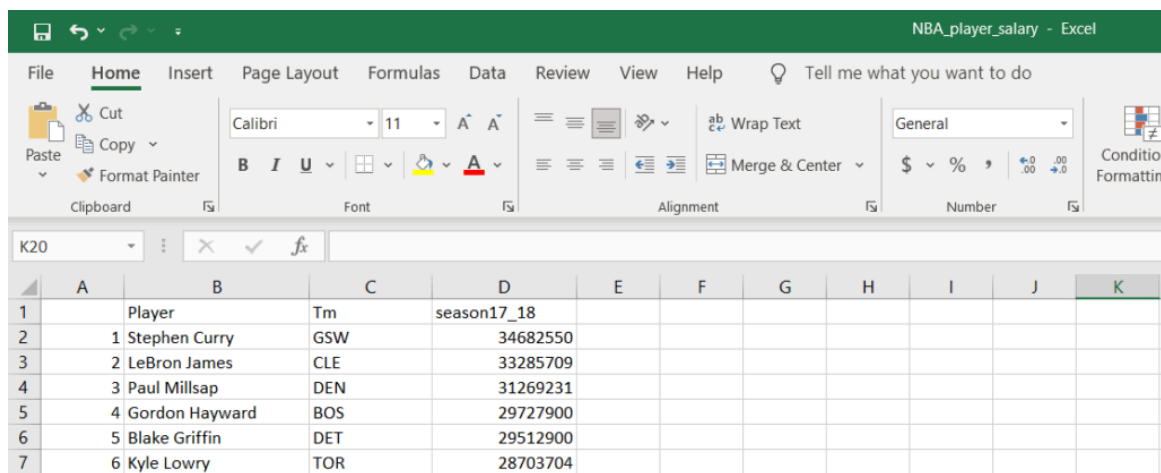
### Когда НЕ стоит использовать среднее арифметическое?

Хотя среднее арифметическое является одной из основных мер центральной тенденции, иногда (на самом деле очень часто) оно наоборот может ввести вас в заблуждение. Данные из реального мира не всегда имеют нормальное распределение. В подавляющем большинстве случаев есть вероятность, что ваши данные ассиметричны.

Ассиметричные данные — это данные, в которых несколько элементов у верхнего или нижнего пределов имеют заметно отличающийся паттерн по сравнению с остальной частью набора данных.

## Пример — набор данных с зарплатами игроков NBA

Давайте посмотрим на [набор данных с зарплатами игроков NBA](#). Этот набор данных содержит зарплаты в долларах США за период с 2017 по 2018 годы.

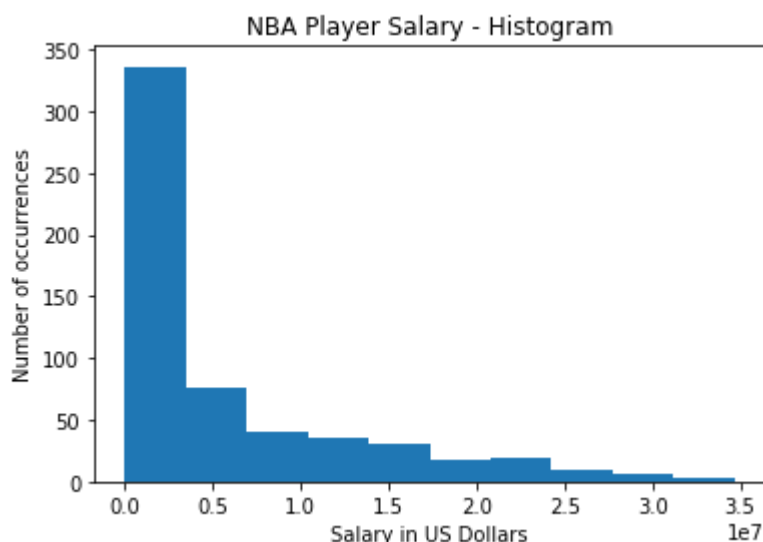


	A	B	C	D	E	F	G	H	I	J	K
1		Player	Tm	season17_18							
2	1	Stephen Curry	GSW	34682550							
3	2	LeBron James	CLE	33285709							
4	3	Paul Millsap	DEN	31269231							
5	4	Gordon Hayward	BOS	29727900							
6	5	Blake Griffin	DET	29512900							
7	6	Kyle Lowry	TOR	28703704							

### Часть набора данных с зарплатами игроков NBA

Я построил гистограмму столбца с зарплатой (название столбца “season17\_18”).

```
import numpy as np
import matplotlib.pyplot as plt
data_nba = pd.read_csv("NBA_player_salary.csv")
plt.hist(data_nba.season17_18)
plt.xlabel("Salary in US Dollars")
plt.ylabel("Number of occurrences")
plt.title("NBA Player Salary - Histogram")
plt.show()
```



### Зарплата игрока NBA – гистограмма

Глядя на приведенное выше распределение, становится очевидным, что данные распределены не нормально. Из 573 игроков более 300 получают зарплату ниже 2,5 миллионов долларов (из графика выше). Но когда мы вычисляем среднее арифметическое заработной платы, оно составляет 5,85 миллиона долларов.

Как вы считаете, годится ли среднее арифметическое в качестве лучшего представления этих данных в целом?

Уж точно нет. Те немногие игроки, которые получали огромные зарплаты, утащили среднее арифметическое далеко от центра. Это называется асимметрией данных.

Не имеет смысла и говорить о том, что среднее арифметическое, которое составляет 5,85 миллиона, является центром, потому что абсолютное большинство из игроков получили зарплату менее 2,5 миллиона долларов.

Таким образом, в случае подобных асимметрий наборов данных среднее арифметическое хорошим выбором для представления данных не является. Здесь нам может помочь медиана.

## 2. Медиана

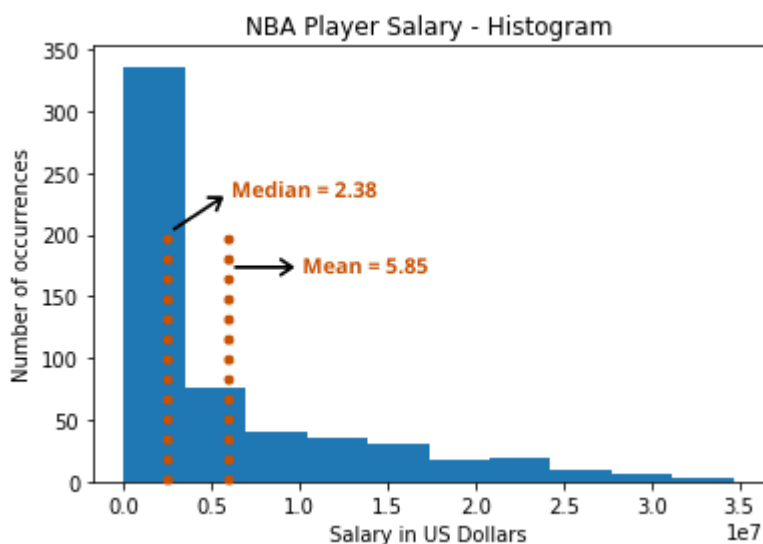
Медиана — это значение, которое находится в центре (прямо посередине), если данные расположены в порядке возрастания или убывания.

Если общее количество значений в наборе данных нечетное, то в центральной позиции будет только одно число. Это и будет наша медиана. Если общее количество значений в наборе данных четное, в центральной позиции будет два значения. В этом случае медиана представляет собой среднее значение этих двух значений.

### Когда следует использовать медиану?

Если набор данных асимметричен или содержит выбросы, среднее арифметическое — не лучший способ представления данных. В таком случае как меру центральной тенденции можно использовать медиану. Выбросы не портят медиану. Потому что само название “выбросы” означает, что они располагаются снаружи, либо в нижнем, либо в верхнем диапазоне. В таком случае медиана — это среднее значение, не нарушенное выбросами.

Еще раз давайте рассмотрим ассиметричный набор данных с зарплатами игроков NBA. (Который мы рассматривали в предыдущем разделе “Когда НЕ стоит использовать среднее арифметическое?”). Медиана по зарплате составляет 2,38 миллиона долларов.



## Диаграмма зарплаты игроков NBA, демонстрирующая среднее арифметическое и медиану

Это значение находится в первой столбце. Обратите внимание, что ось X это  $10^7$ . Итак, первый столбик представляет зарплату до 2,5 миллионов. Таким образом, медианное значение 2,38 миллиона лучше всего представляет эти данные, потому что большинство игроков получают зарплату, близкую к этому показателю.

### Когда НЕ стоит использовать медиану?

Если и среднее арифметическое, и медиана одного и того же набора данных не сильно отклоняются, то можно использовать обе эти меры. В любом случае расчет среднего арифметического предполагает учет всех элементов данных и их усреднение. Таким образом, логичнее, что среднее арифметическое является более точной мерой (когда среднее арифметическое и медиана не сильно отклоняются).

### Как определить, является ли ваш набор данных асимметричным или содержит выбросы?

Самый банальный способ определить, является ли ваш набор данных асимметричным или содержит выбросы, — это вычислить среднее арифметическое и медиану. Если обе меры не сильно отклоняются, то с вашим набором данных все в порядке. И вы сэкономили время, которое в противном случае было бы потрачено на очистку и преобразование данных.

Если среднее арифметическое и медиана очень сильно отклоняются, ваш набор данных асимметричен или содержит выбросы.

Следующий шаг — провести исследование с целью выявить и удалить выбросы, если таковые имеются. Или применить какое-либо преобразование, чтобы уменьшить асимметрию в ваших данных, если таковая имеется.



### 3. Мода

Мода — это значение, которое чаще всего встречается в наборе данных. В гистограмме мода — это значение с самым высоким столбцом.

Если набор данных имеет более одного значения с одинаковой максимальной частотой появления, набор данных имеет мультимодальное распределение, поскольку он имеет несколько мод. Если в наборе данных нет повторяющихся значений, то и моды у него тоже нет.

#### Когда стоит использовать моду?

Моду можно использовать для анализа часто встречающихся значений как числовых, так и категориальных данных.

Мода — единственная мера центральной тенденции, которую можно использовать с категориальными данными. Для категориальных данных вы не можете вычислить среднее арифметическое или медиану. Мода - единственный выбор в таких случаях.

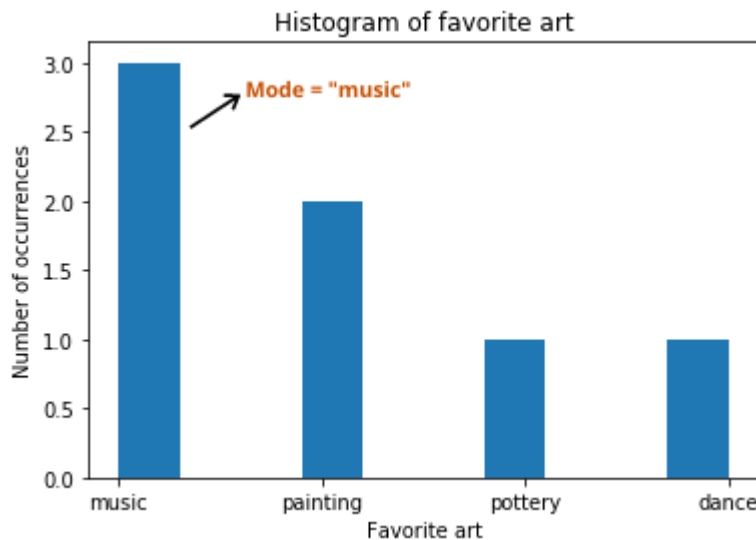
#### Пример — Простое перечисление

Ниже приведен учебный набор данных, отражающий любимый вид искусства семерых человек. Построим частотный график (гистограмму).

```
data_art = ['music', 'painting', 'pottery', 'painting', 'dance', 'music', 'music']
```

```
import matplotlib.pyplot as plt
data_art = ['music', 'painting', 'pottery', 'painting',
'dance', 'music', 'music']
plt.hist(data_art)
plt.xlabel("Favorite art")
plt.ylabel("Number of occurrences")
```

```
plt.title("Histogram of favorite art")  
plt.show()
```



Гистограмма любимого вида искусства — пример моды

---

Во многих областях машинного обучения возникают функции многих переменных и их производные. Такие производные ещё называют "матричными". На открытом уроке мы поговорим про отличие таких производных от обычных, изучаемых в школе, разберём необходимую теорию, научимся такие производные считать, а также посмотрим, где и как матричные производные используются. Регистрация открыта [по ссылке](#) для всех желающих.