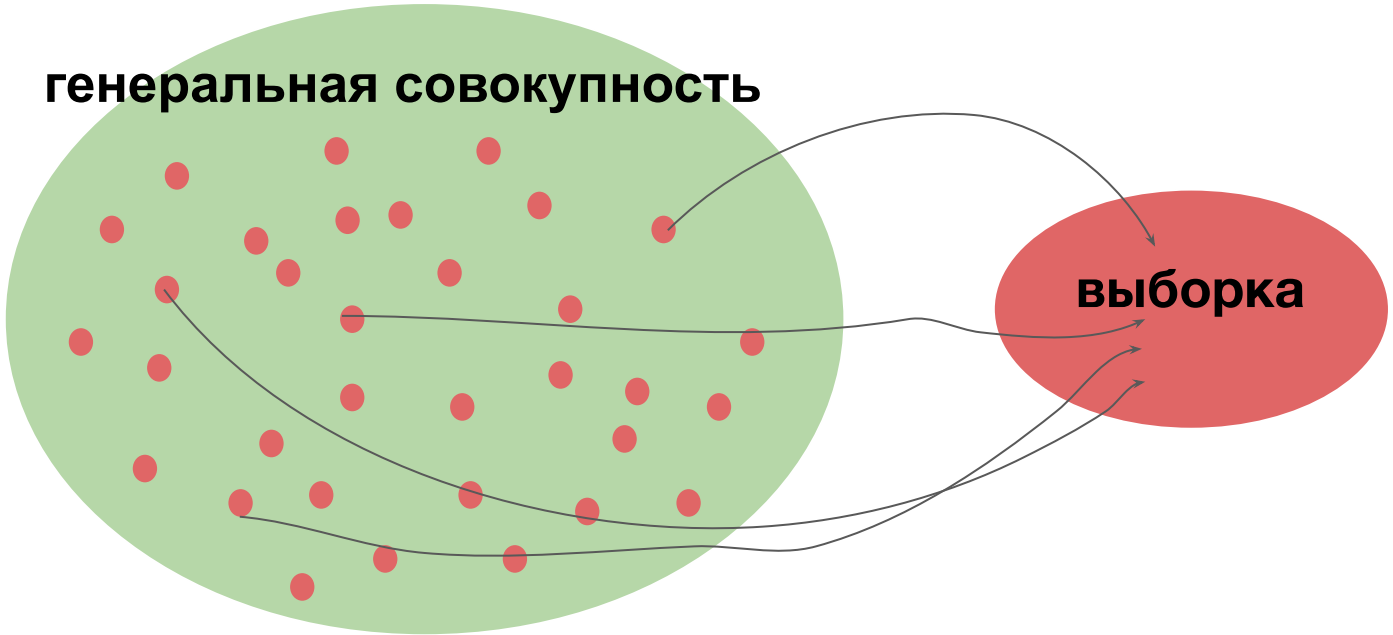
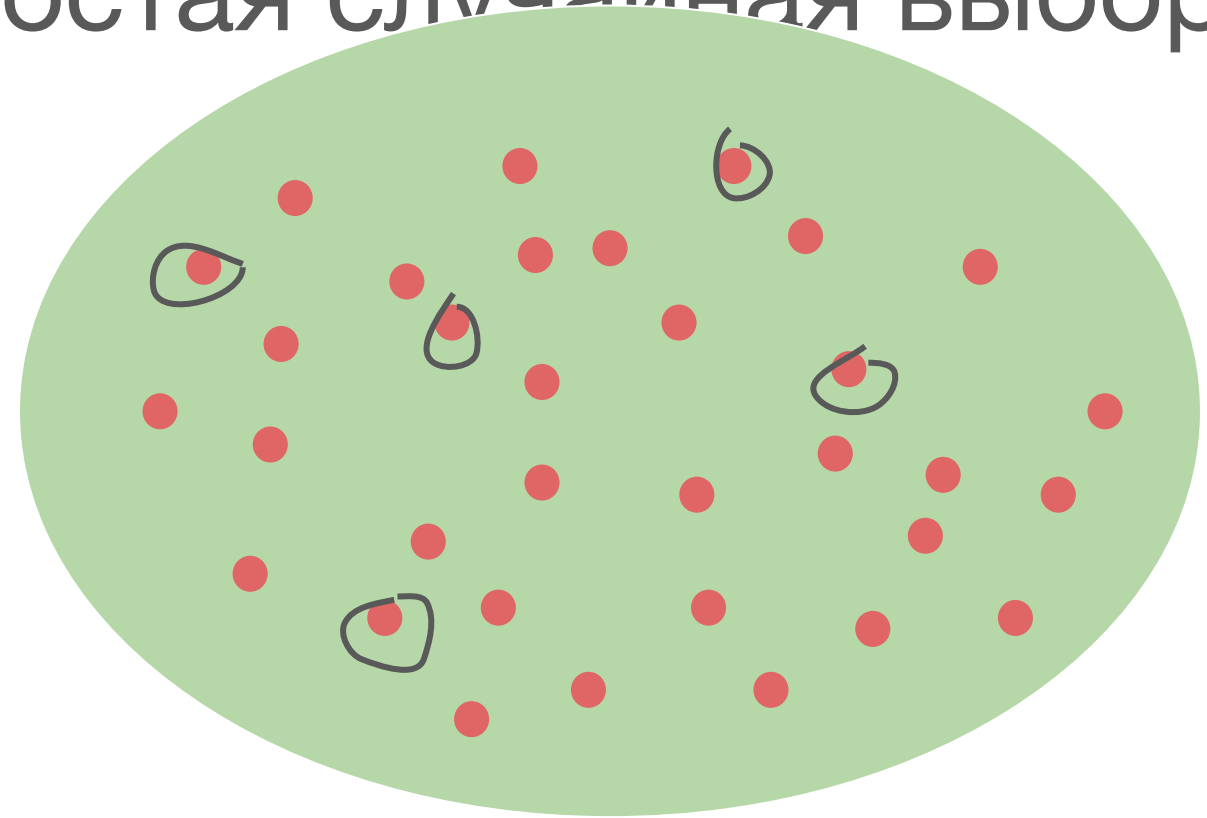
лекция 4, спецкурс по биоинформатике 2022

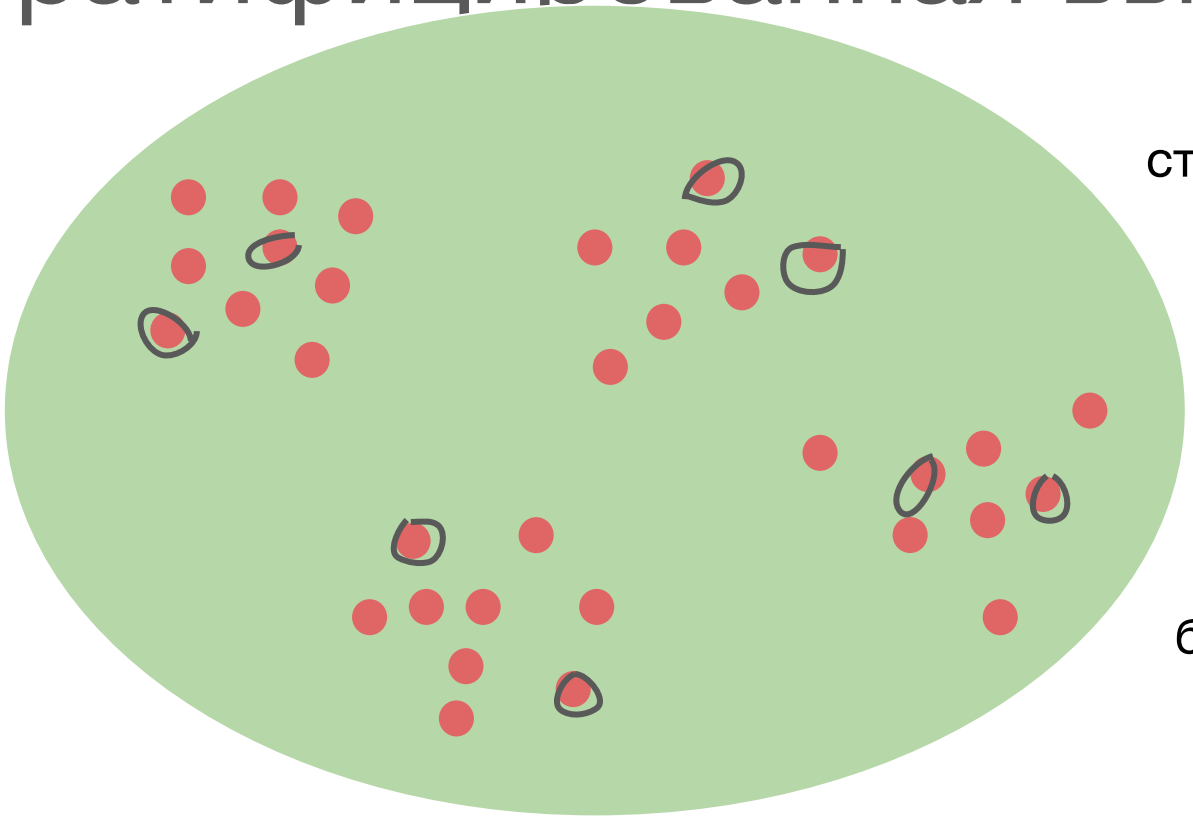


<все объекты исследования>

простая случайная выборка



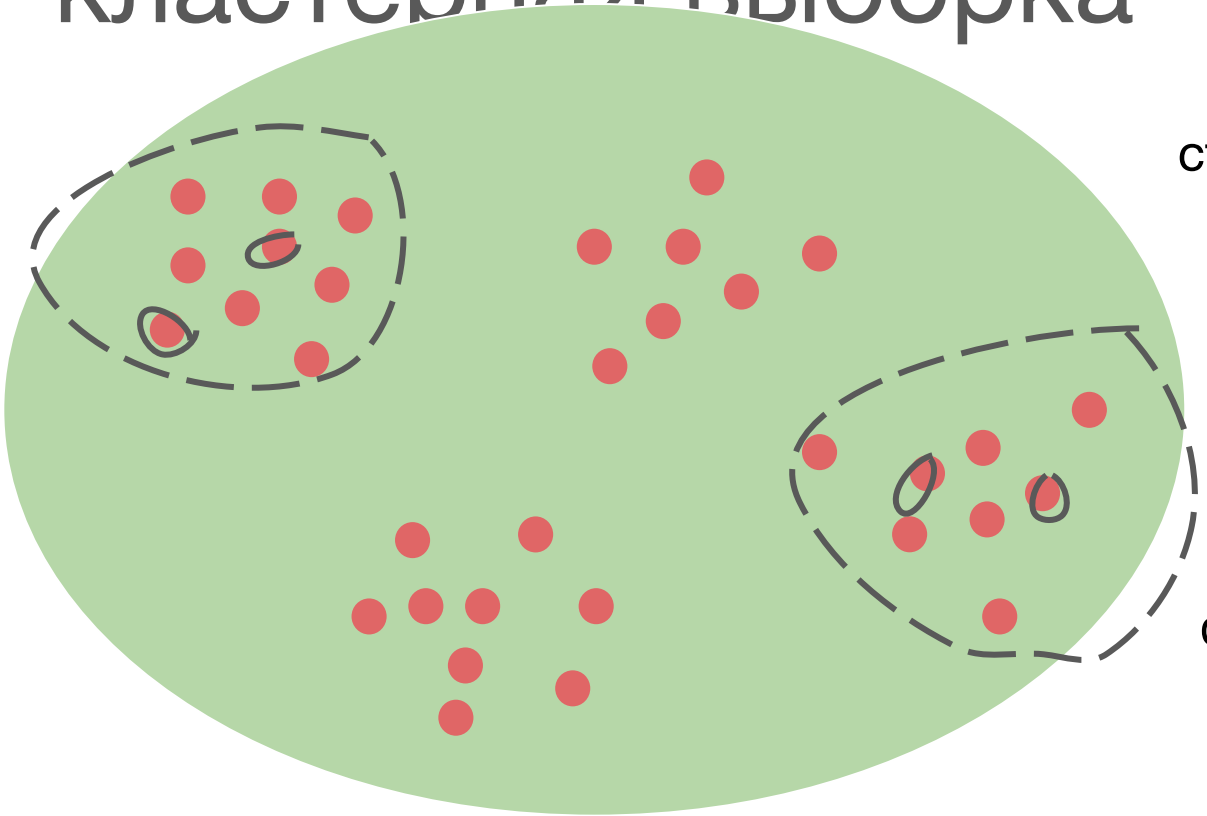
стратифицированная выборка



страты между
собой разнородные,
внутри однородные

оказывается
более точной

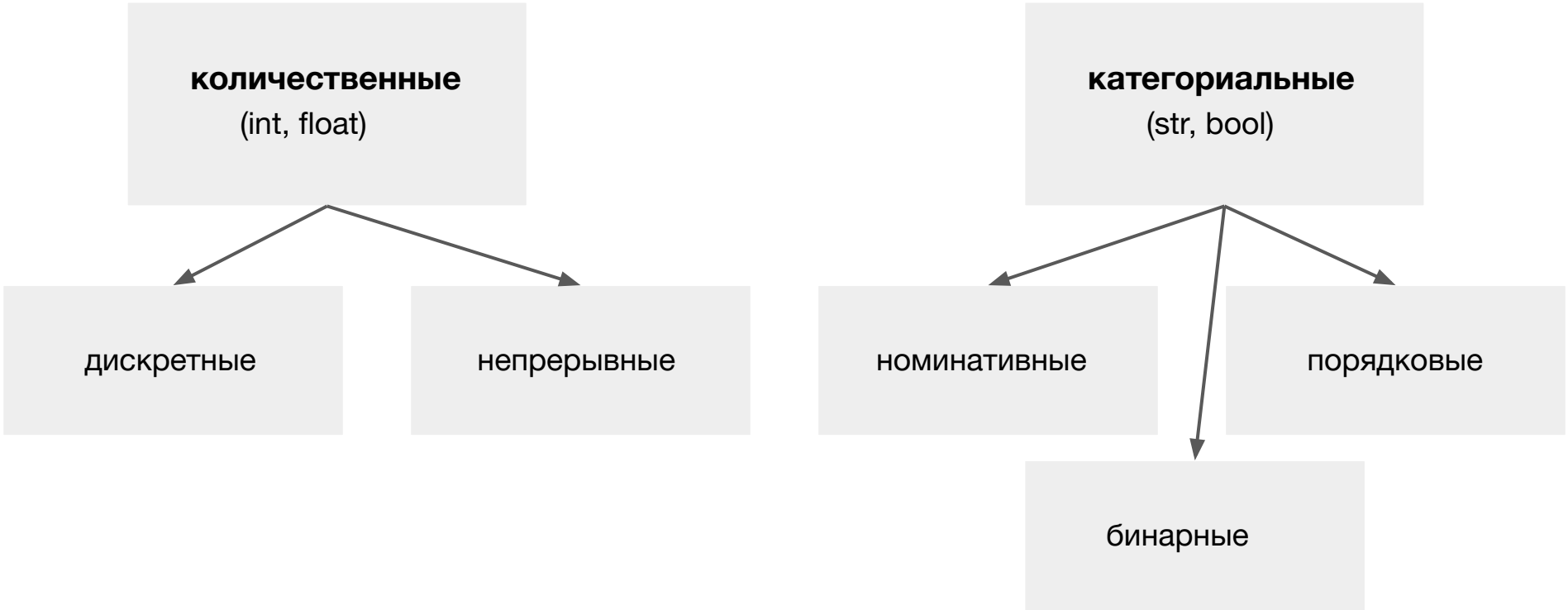
кластерная выборка



страты между
собой однородные,
внутри разнородные

проще и дешевле
формировать

типы переменных



дискретные

непрерывные

раса

ответы акинатора

преподы по биоинфе в 1514

високосность года

высота этажа

гендер

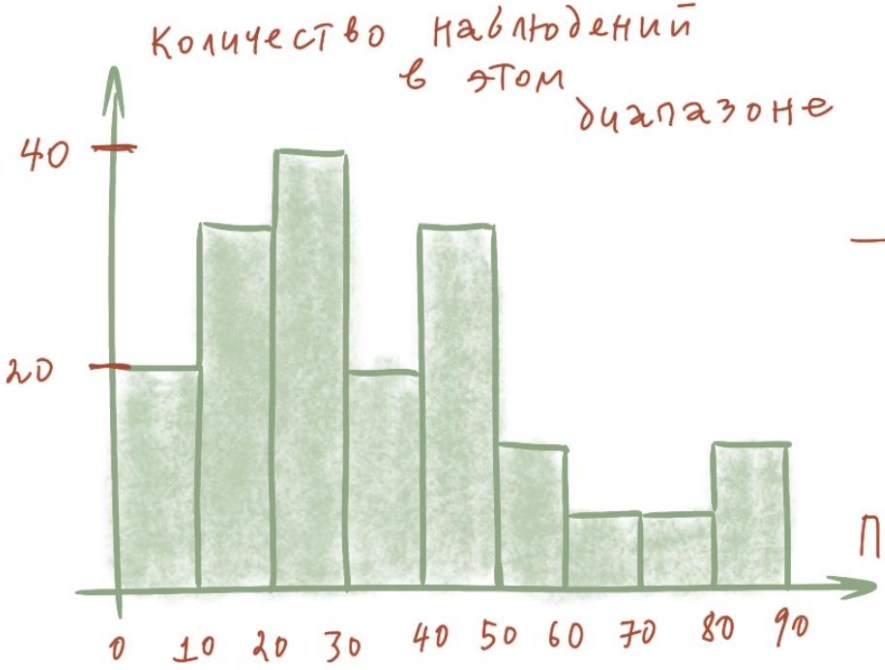
количество родственников

номинативные

порядковые

бинарные

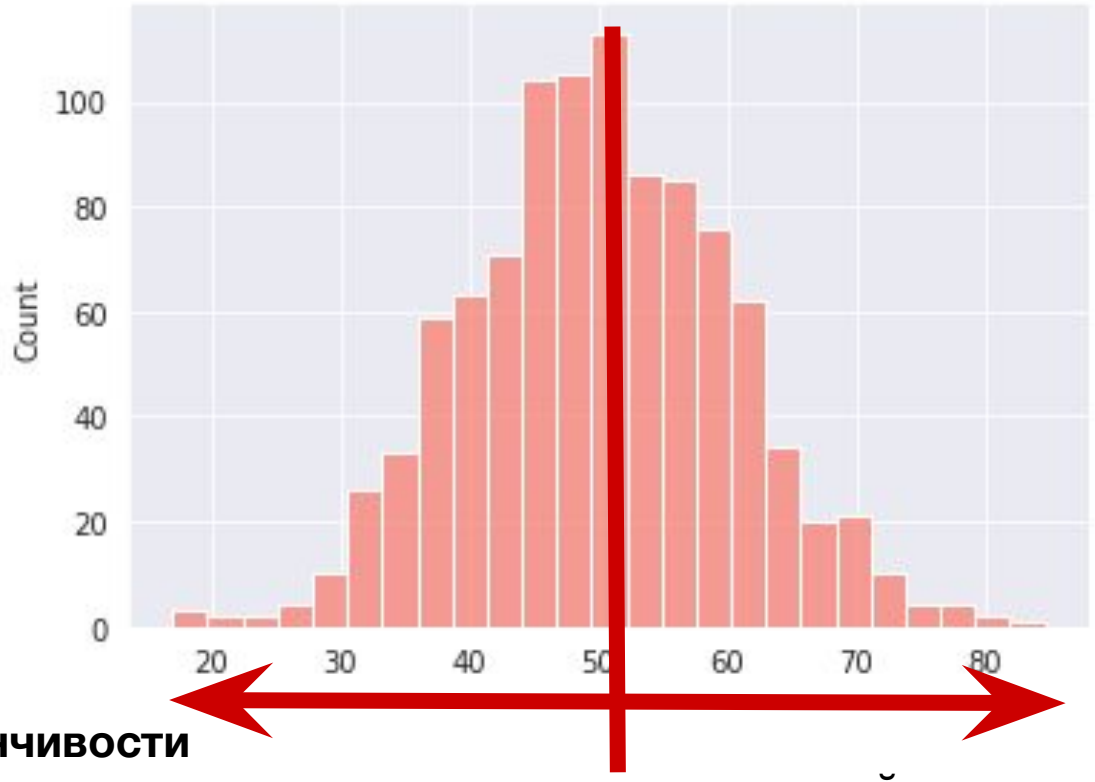
гистограммы



есть 20 наблюдений, где $0 \leq$ признак ≤ 10,

есть 40 наблюдений, где $20 \leq$ признак ≤ 30

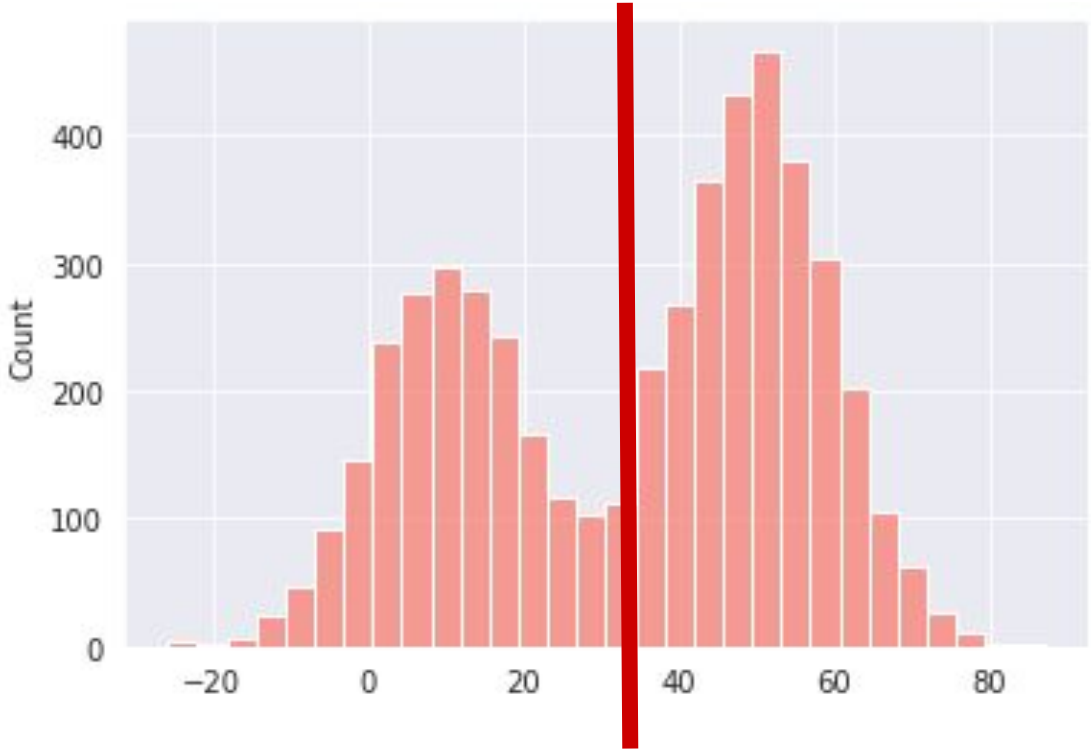
описательные статистики



меры изменчивости

меры центральной тенденции

меры центральной тенденции



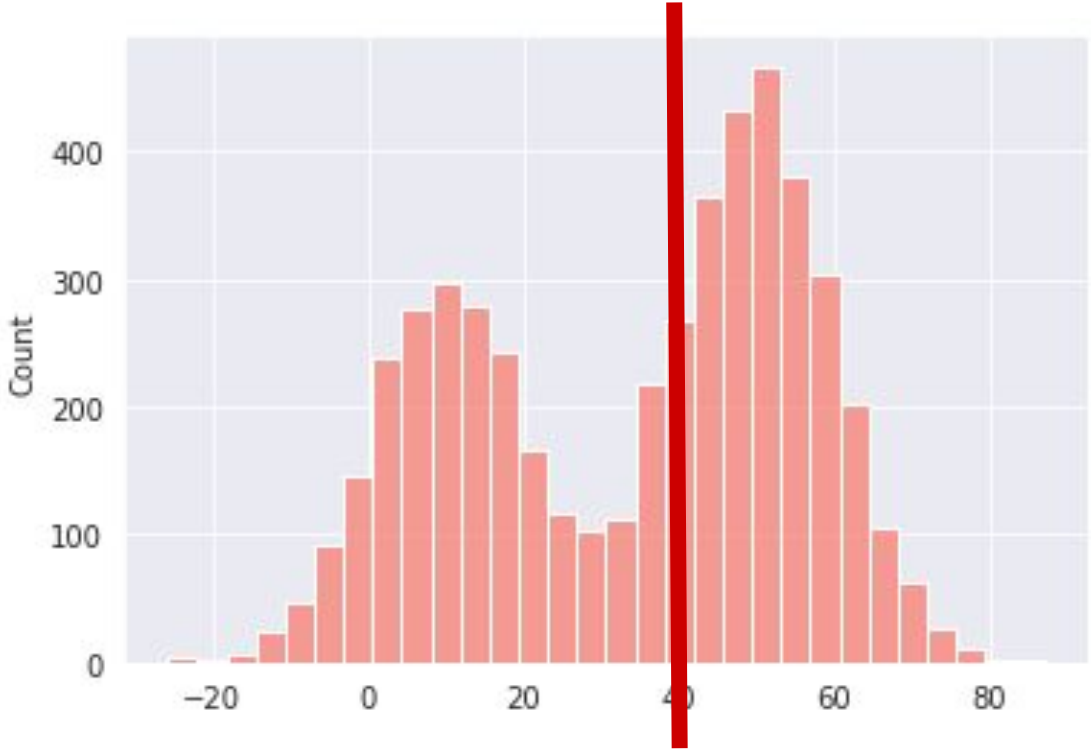
среднее значение

сумма значений, деленная на их количество

```
sum(a)/len(a)
#или np.mean(a)
```

```
33.93214742943539
```

меры центральной тенденции



медиана

середина в отсортированном списке значений

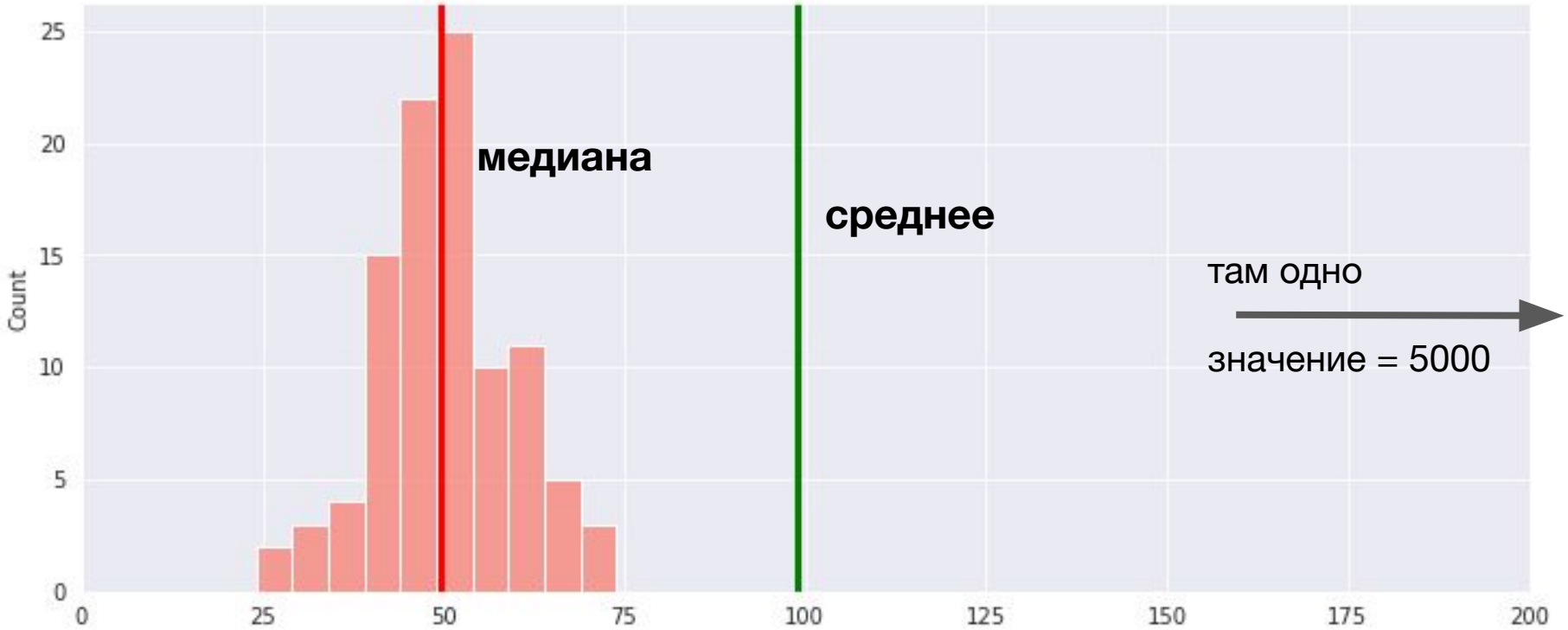
```
a.sort()
a[len(a)//2]
#или np.median(a)
```

```
40.06090011055331
```

зачем медиана, когда есть среднее?

```
plt.figure(figsize=(12,5))
a = np.random.normal(50, 10, 100)
a = np.append(a, np.ones(1)*5000)
plt.xlim(0, 200)
sns.histplot(a, color = 'salmon', bins = 1000)
median = np.median(a)
mean = np.mean(a)
plt.axvline(x=mean, color = 'g', linewidth = 3)
plt.axvline(x=median, color = 'r', linewidth = 3)
pass
```

зачем медиана, когда есть среднее?



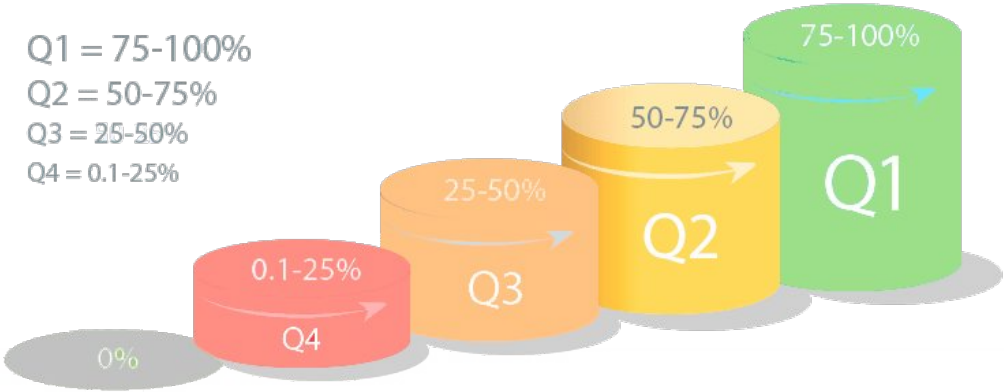
<небольшое отступление от мер центральной тенденции>

если наш отсортированный список – это отрезок, то можно делить его не пополам а, например, на 3 или на 4 части

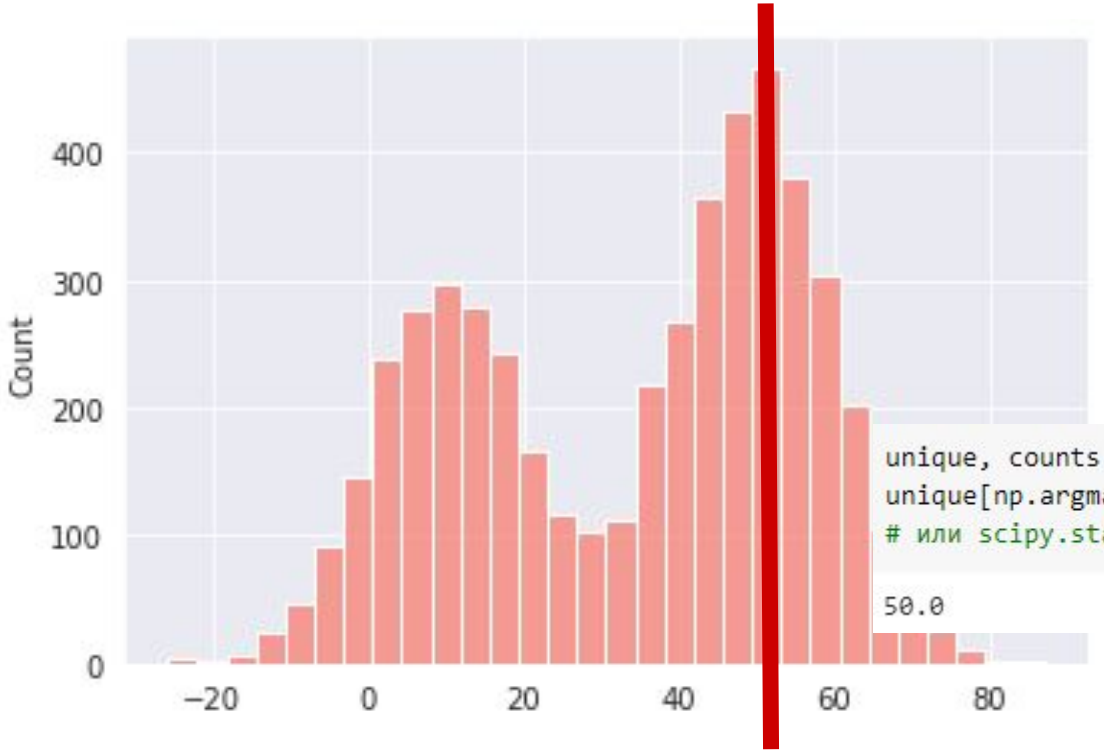
квантили - это такие значения признака, которые делят упорядоченные данные на некоторое число равных частей

если таких частей 4, мы имеем дело с **квартилями (Q1, Q2, Q3)**

если же 100 – с **перцентилями**



меры центральной тенденции



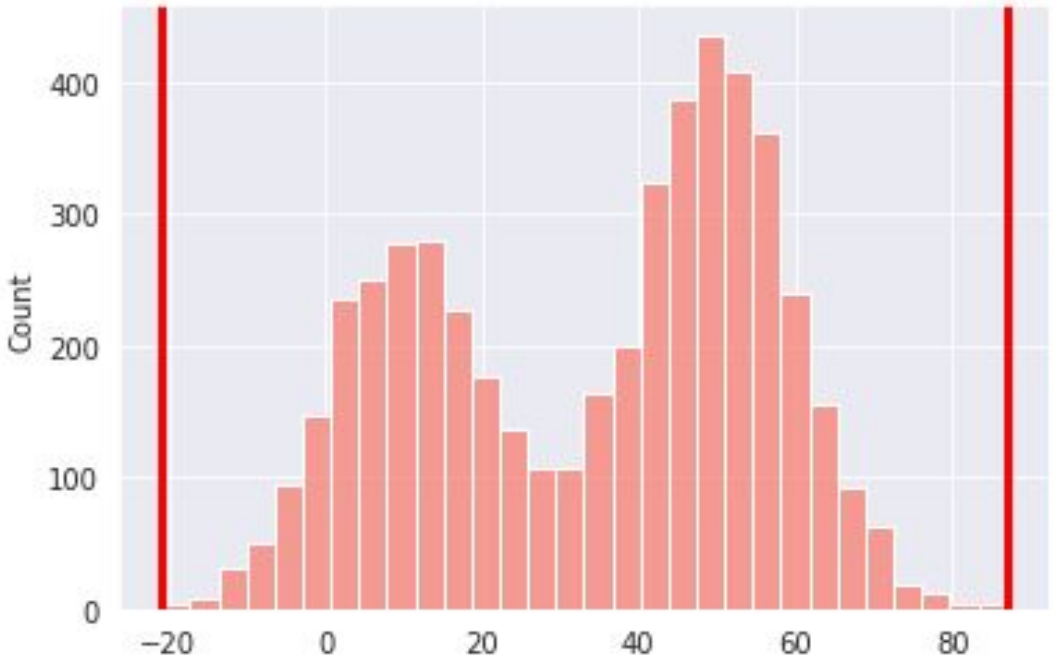
мода

значение,
встречавшееся
чаще всего

```
unique, counts = np.unique(a.round(), return_counts=True)
unique[np.argmax(counts)]
# или scipy.stats.mode(a.round())
```

```
50.0
```

меры изменчивости



размах

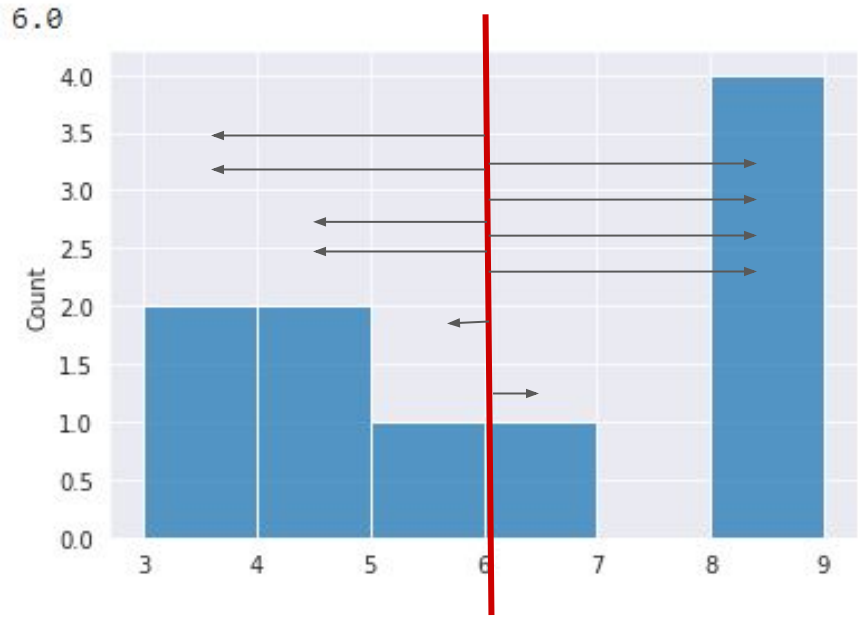
разница между минимальным и максимальным значением

```
np.max(a)-np.min(a)
```

```
107.63015268314605
```

меры изменчивости

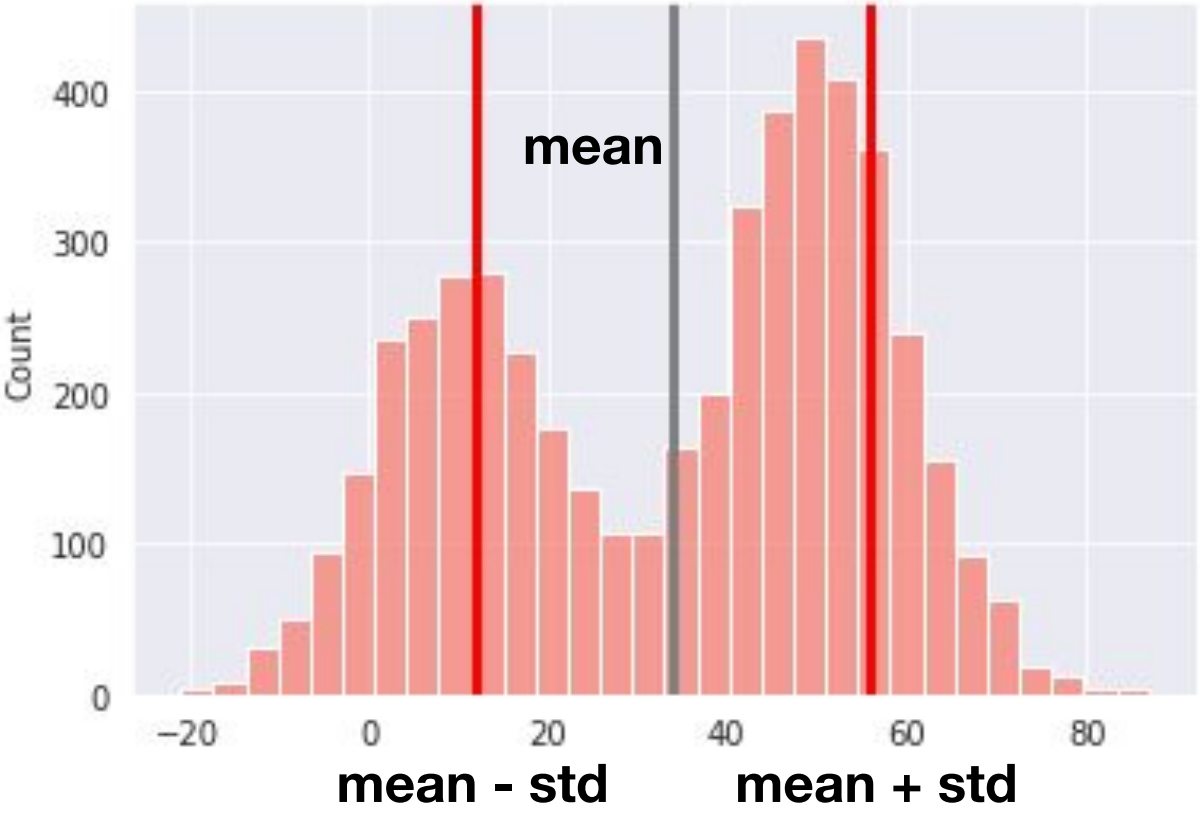
```
k = np.random.randint(3, 10, 10)
sns.histplot(x = k, bins = 10-3-1)
print(np.mean(k))
```



дисперсия

$$Var = \frac{\sum_{i=1}^{n} (X_{mean} - X_i)^2}{n-1}$$

по размерности дисперсия – квадрат величины. поэтому чтобы вернуться в наши единицы измерения, надо взять из дисперсии корень. полученная величина – **стандартное отклонение (std)**




```
2*std
```

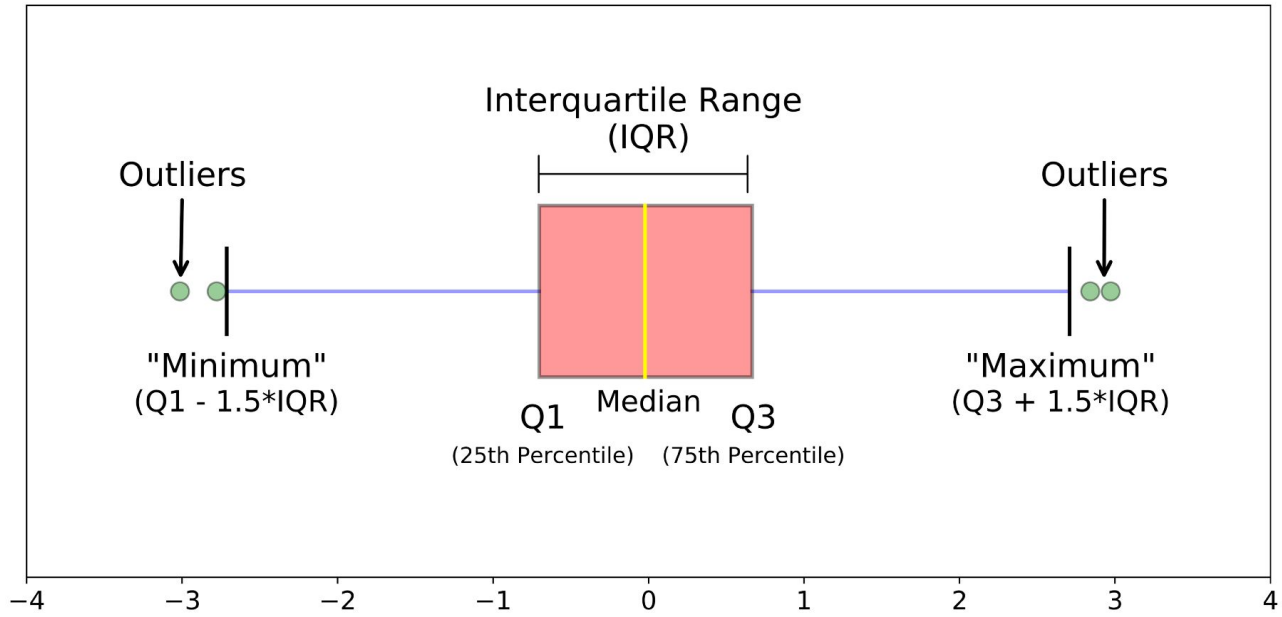
```
107.63015268314605
```

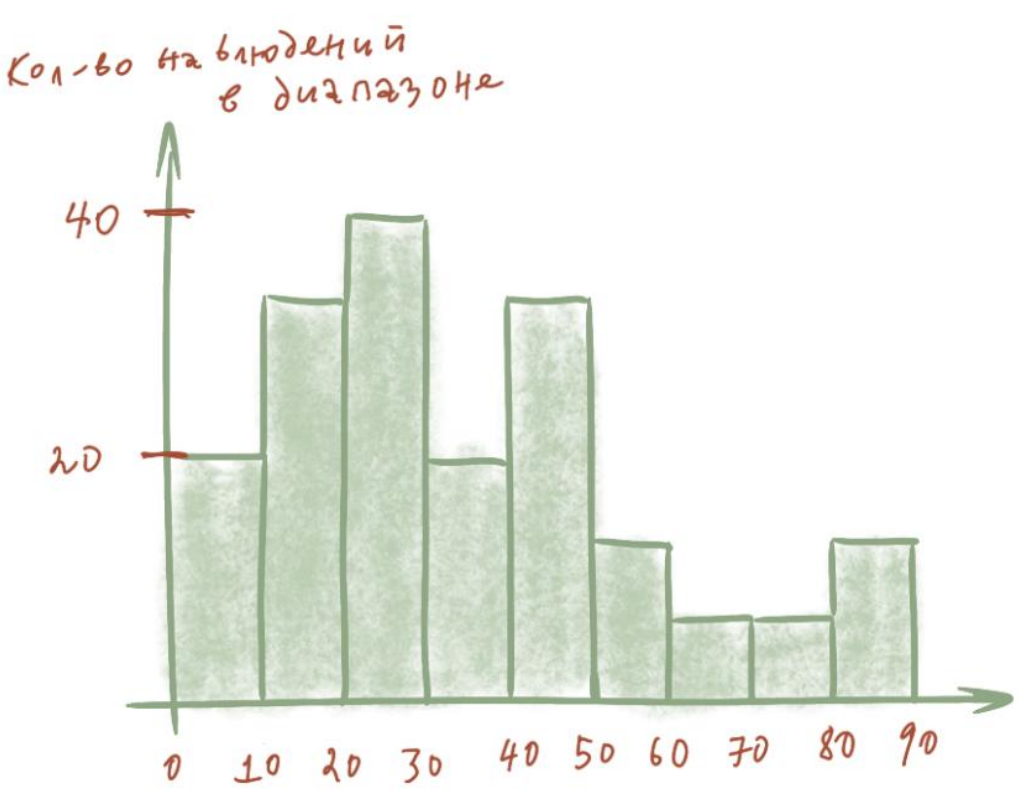
меры изменчивости

интерквартильное расстояние

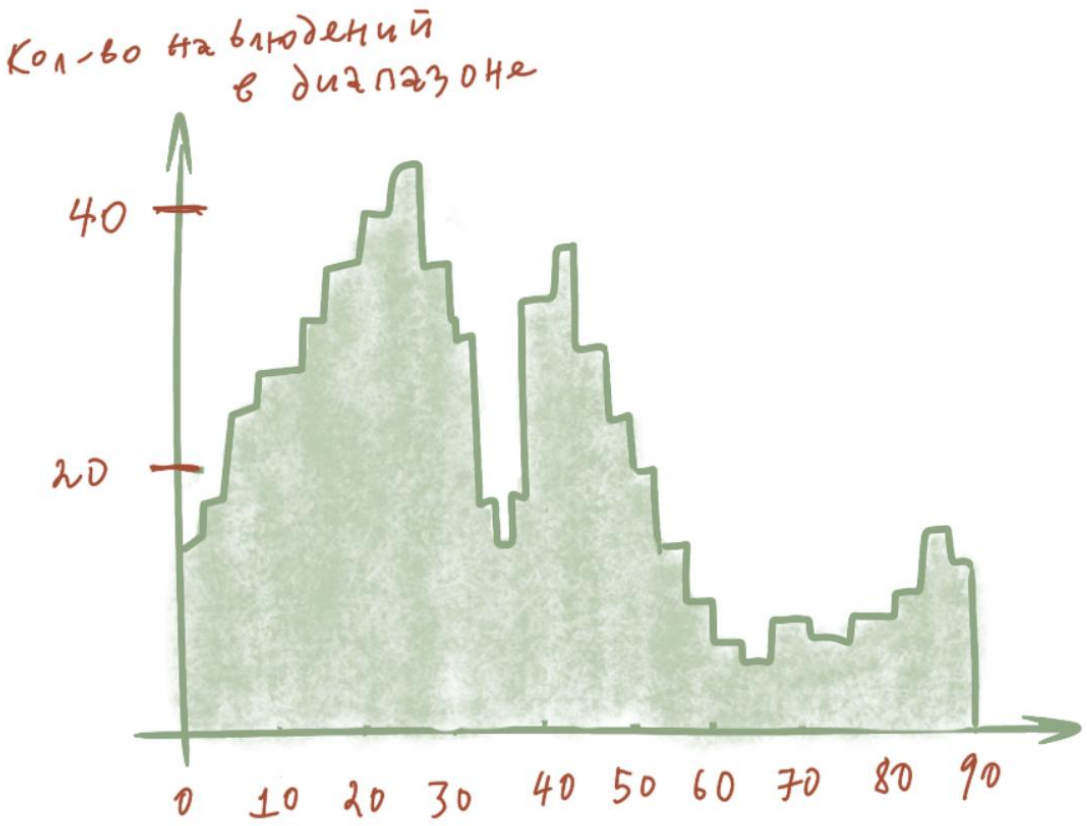
расстояние между первым и третьим квартилем

боксплоты

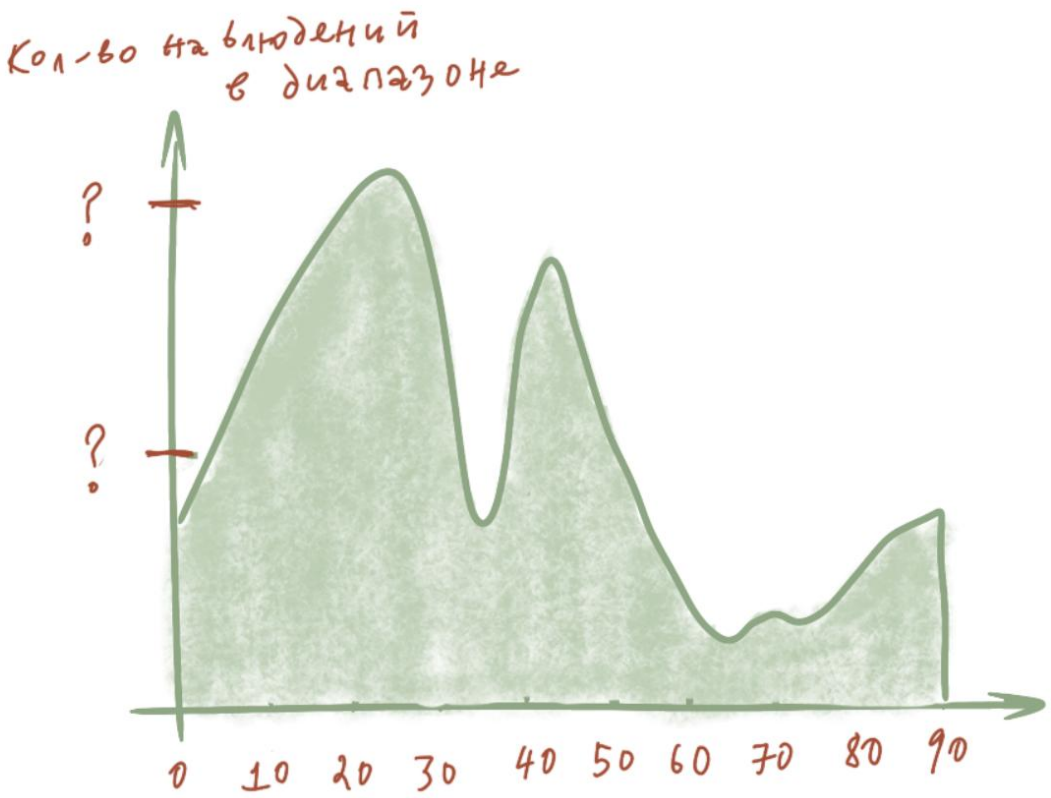

Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4



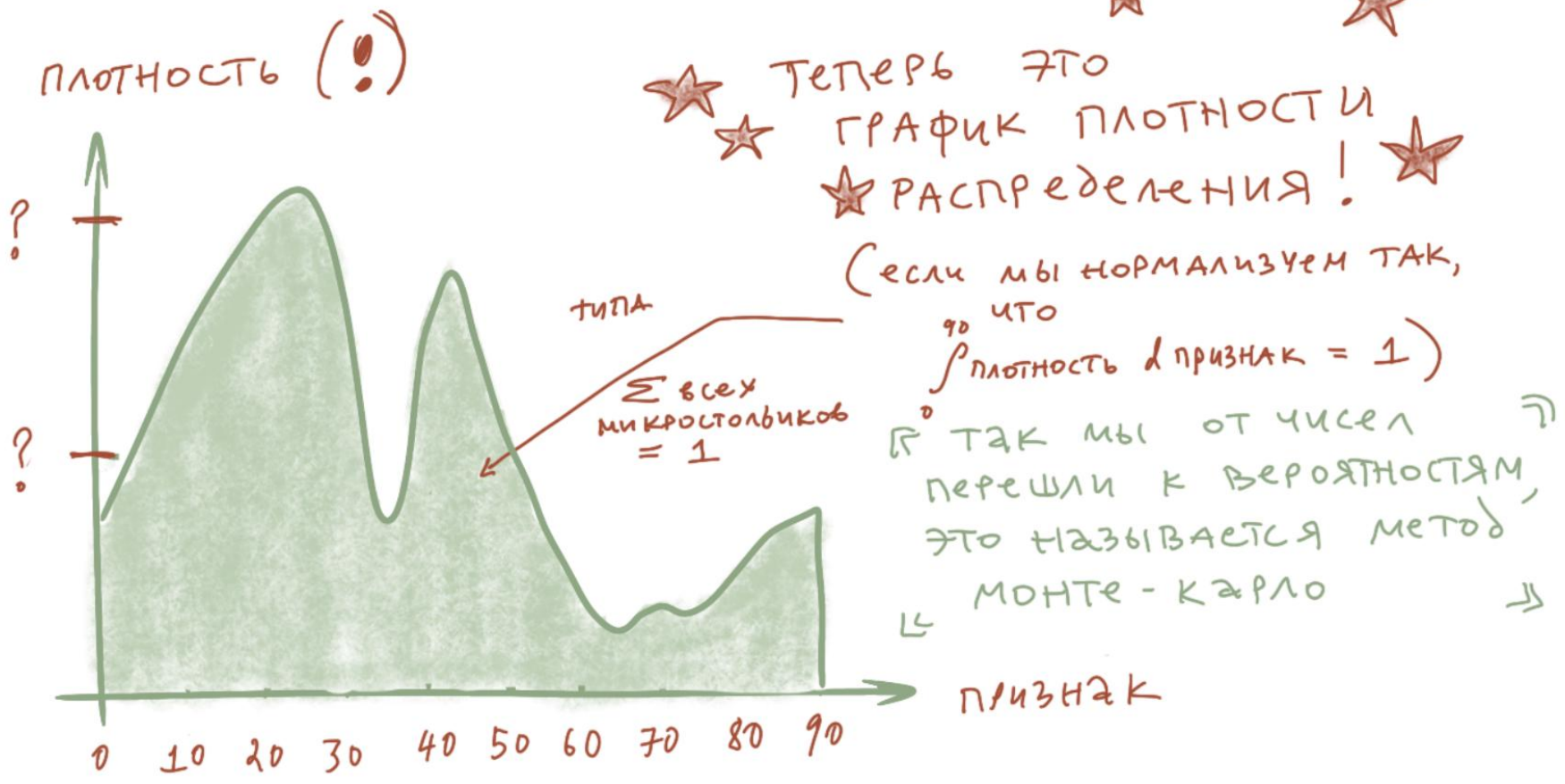
Что будет, если увеличить количество наблюдений?



что будет, если увеличить количество наблюдений?

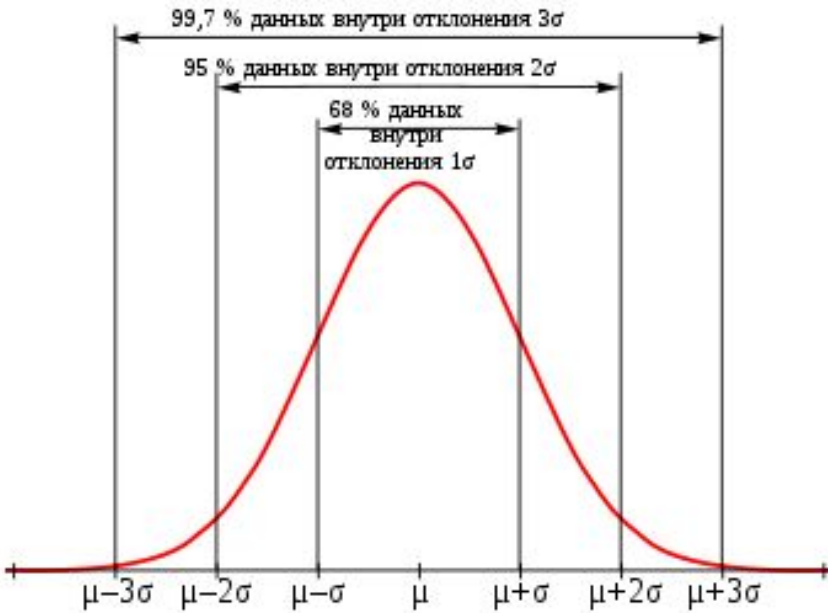


что будет, если увеличить количество наблюдений?


плотность (!)
?
?
0
10
20
30
40
50
60
70
80
90
признак
типа
Σ всех микростолбиков = 1
Теперь это
график плотности
распределения!
(если мы нормализуем так, что
$\int_0^{90}$ плотность d признак = 1)
так мы от чисел перешли к вероятностям, это называется метод монте-карло

оказывается, очень много вещей распределено одинаково…

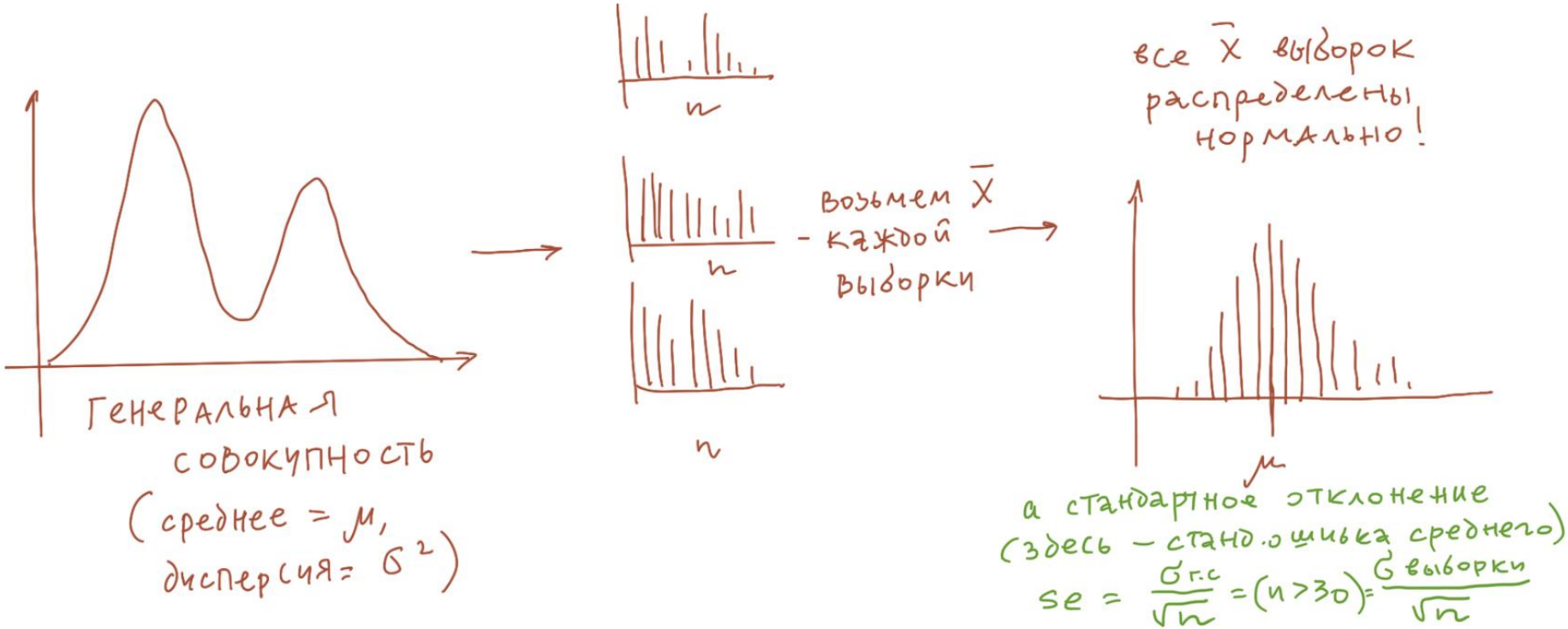
нормальное распределение



- унимодально
- симметрично
- прикольная штука с процентами: значения отклоняются от среднего в соответствии с неким вероятностным законом

если средний рост человека = 171 см
с sd = 12, какой процент людей
выше 195 см?

центральная предельная теорема



Уровень экспрессии некоторого гена измерялся в эксперименте. Ниже представлены результаты 64 наблюдений.

102 91 99 100 103 98 99 101 106 88 103 97 103 101
101 91 104 105 105 100 101 91 99 98 107 102 100 97
98 104 100 98 102 99 95 103 104 97 99 102 98 107 101
93 98 101 93 91 107 102 96 93 100 105 103 107 99 102
106 102 94 104 103 102

$\bar{X}=100$ $sd=4$

какова средняя экспрессия гена по всей популяции?