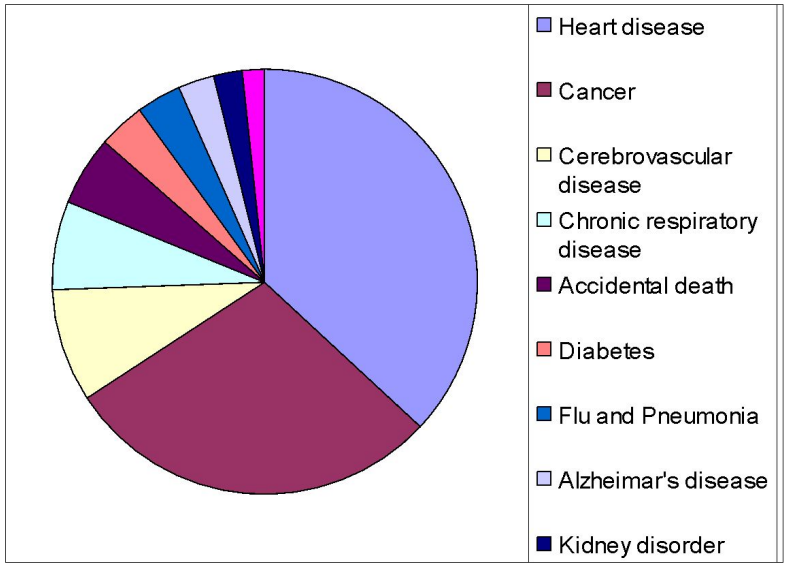
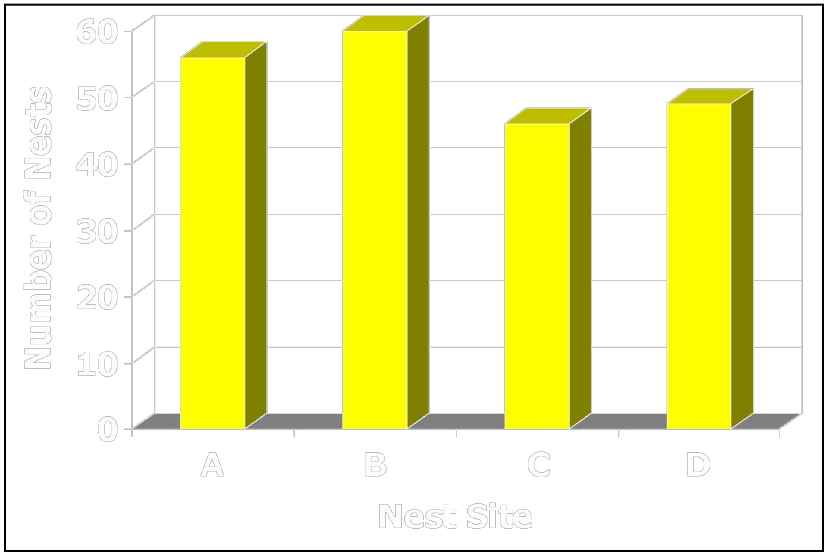
ЛЕКЦИЯ 2

# ОПИСАТЕЛЬНАЯ СТАТИСТИКА

# 2.1. Группировка данных



# Обработку данных полезно начать с их *группировки*…

**<u>Группировка</u> - это систематизация первичных данных, направленная на *извлечение заключенной в них информации и выявление закономерностей*, которым подчиняется изучаемое явление или объект.**

# Пример: медицинские сведения

- Пол (м, ж)
- Возраст (полных лет)
- Группа крови (I, II, III, IV)
- Систолическое давление (мм рт.ст.)
- Курильщик (да, нет)
- Рост (см)
- Вес (кг)
- …

**Качественные переменные** – определяется принадлежность объекта к одной из нескольких категорий

**Количественные (непрерывные, дискретные) –** дают числовую величину; к ним применяют арифметические действия

# Группировка количественных данных :

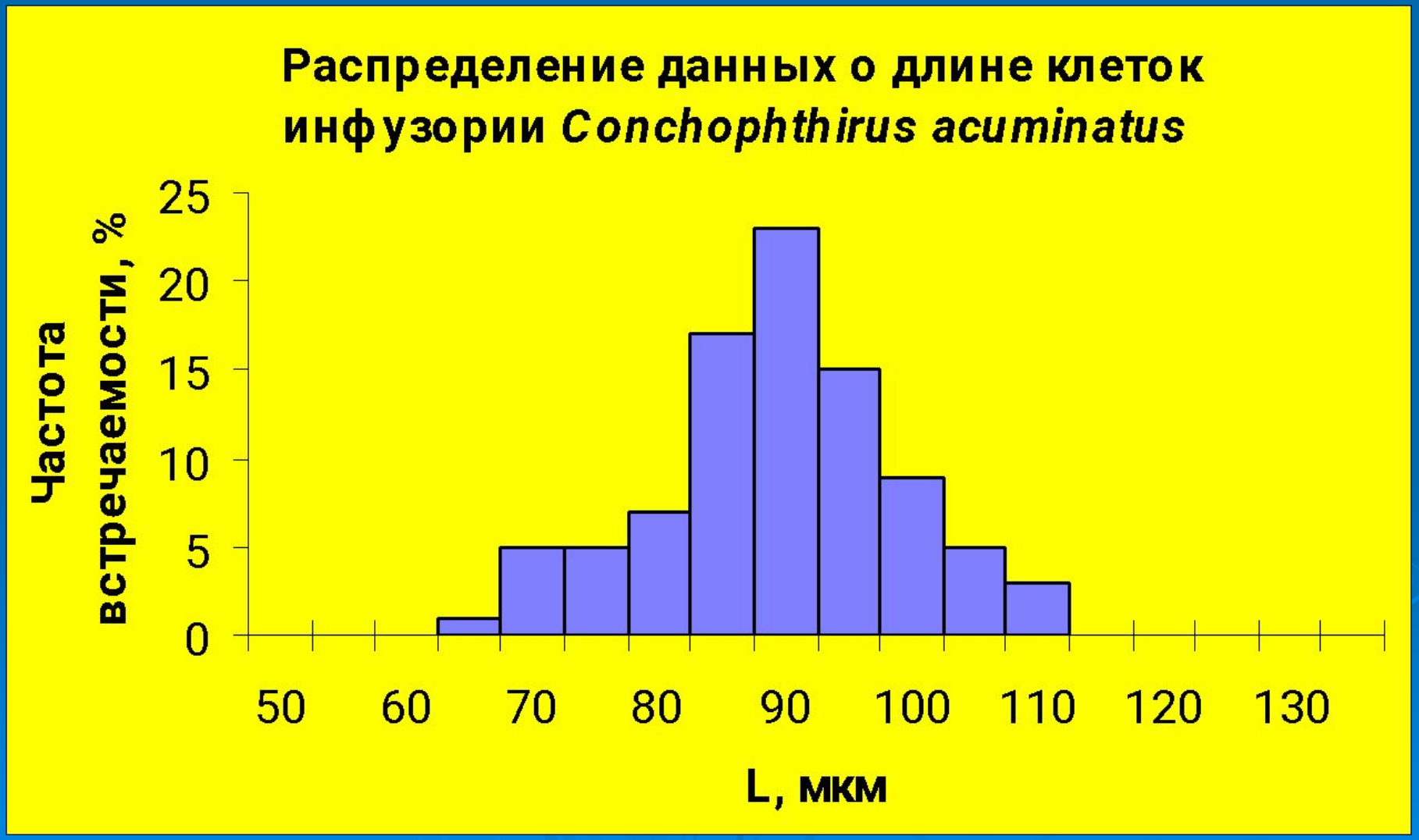
**по значениям вариант**

**по классам**

**Представление частотного распределения графически**

При небольшом *n* и незначительной вариации признака, количественные данные группируют по значениям вариант (полигон распределения)

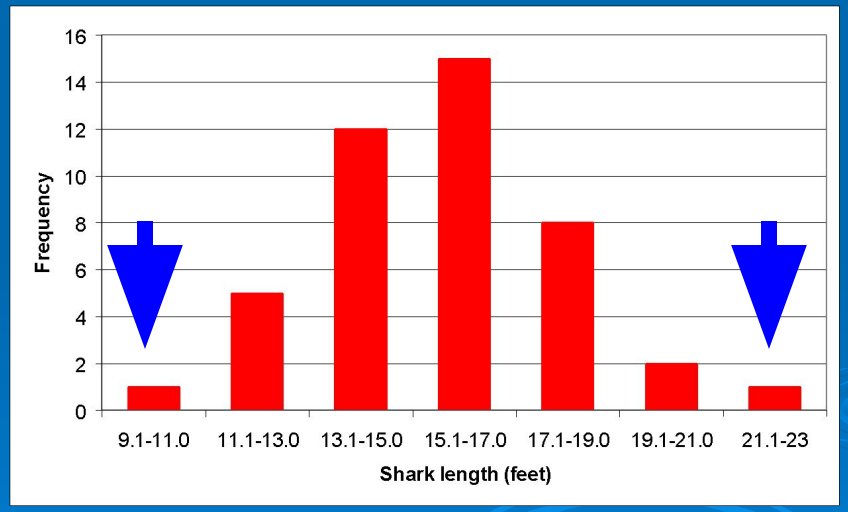
# *Гистограмма*: данные группируются по классам



# Какую информацию дает *вариационный ряд и его график*?

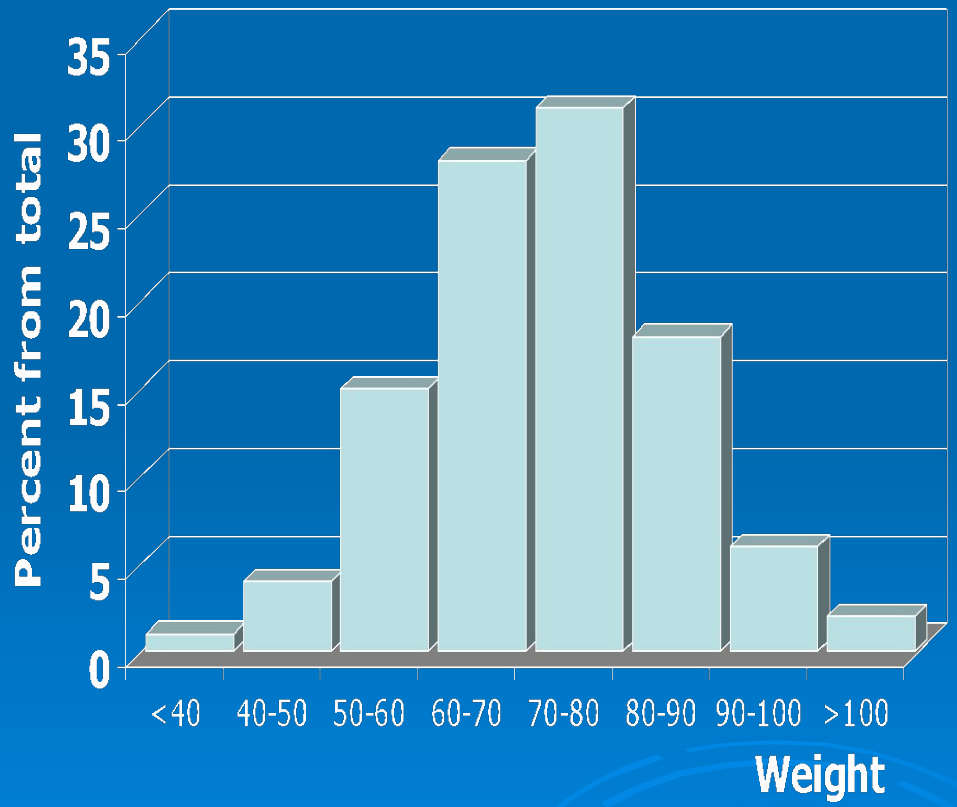
- ***Границы изменчивости признака*: минимальное и максимальное значение вариант, или *лимиты*.**

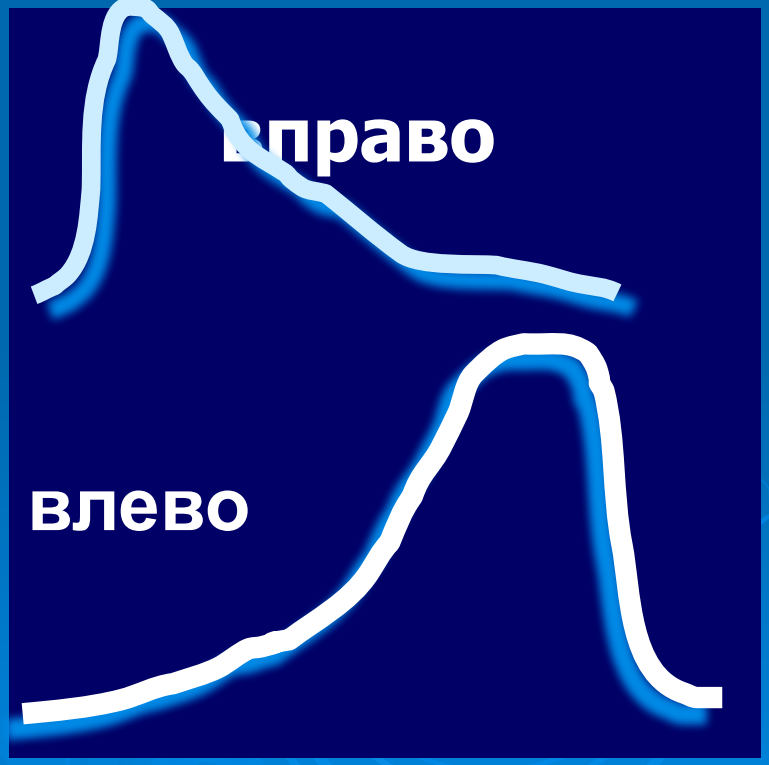
| | | | | |
|---|---|---|---|---|
| $(x_i)$: | 2 | 3 | 4 | 5 |
| $(f_i)$: | 1 | 2 | 5 | 2 |




**Разница между лимитами называется *размахом* выборки**

- ***Характер вариации признака***: исследователь может установить симметричность распределения



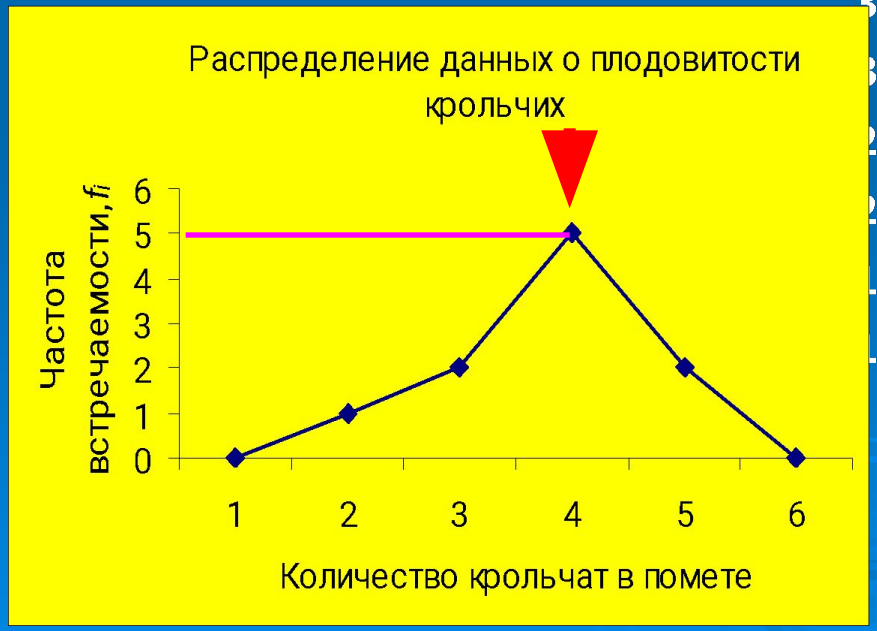



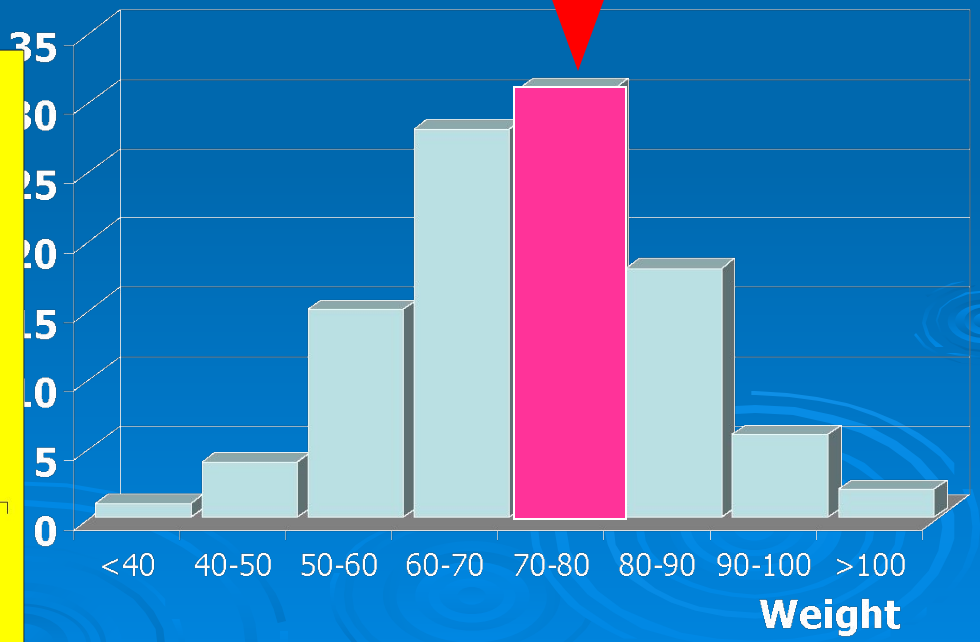
**а также** моду (наиболее часто встречающееся значение)

или ***модальный*** **класс**

$(x_i)$: 2 3 4 5

$(f_i)$: 1 2 5 2



# Круговые диаграммы (*Pie chart*)

## (для качественных признаков)

**Включают <u>все</u> категории которые <u>формируют совокупность</u>**

Используют, чтобы изобразить <u>вклад каждой категории</u>

| Top 10 causes of death | Counts of deaths | % of top 10 case |
|---|---|---|
| Heart disease | 700,142 | 37% |
| Cancer | 553,768 | 29% |
| Cerebrovascular disease | 163,538 | 9% |
| Chronic respiratory disease | 123,013 | 6% |
| Accidental death | 101,537 | 5% |
| Diabetes | 71,537 | 4% |
| Flu and Pneumonia | 62,034 | 3% |
| Alzheimer's disease | 53,852 | 3% |
| Kidney disorder | 39,480 | 2% |
| Septicemia | 32,238 | 2% |
| | 1,901,139 | 100% |

# 2.2. Среднее значение и стандартное отклонение

Любое нормальное распределение можно описать с помощью всего двух параметров:

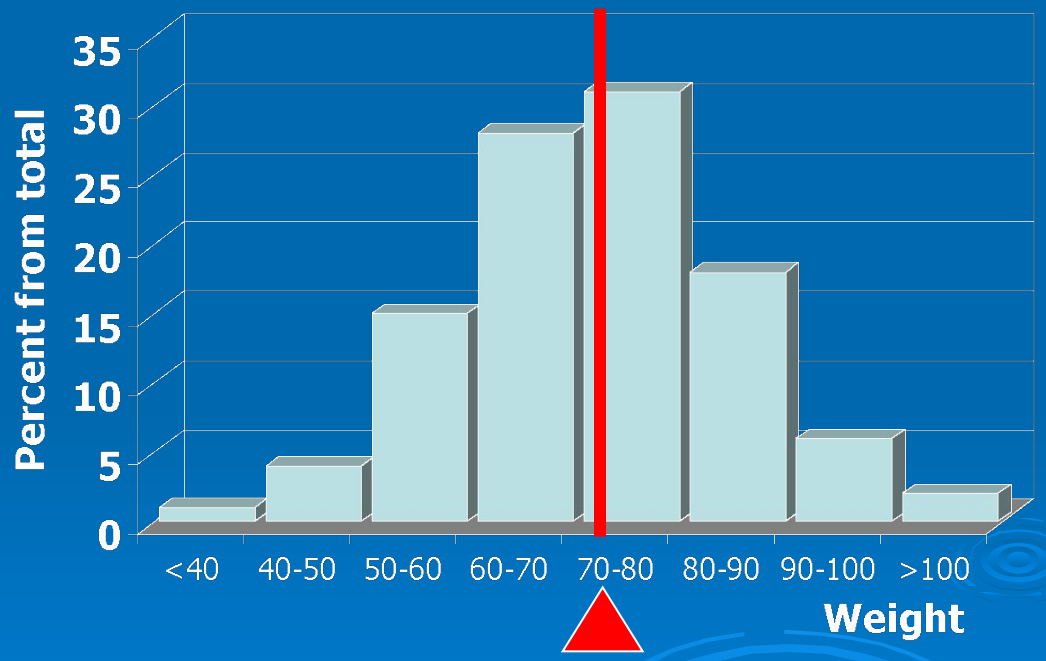
*среднего значения ($\mu$)* и *стандартного отклонения ($\sigma$)*

# ВЫБОРОЧНАЯ СРЕДНЯЯ

(англ.: *sample mean*)

(= средняя арифметическая)

$$\bar{x} = \frac{1}{n}\sum x_i$$



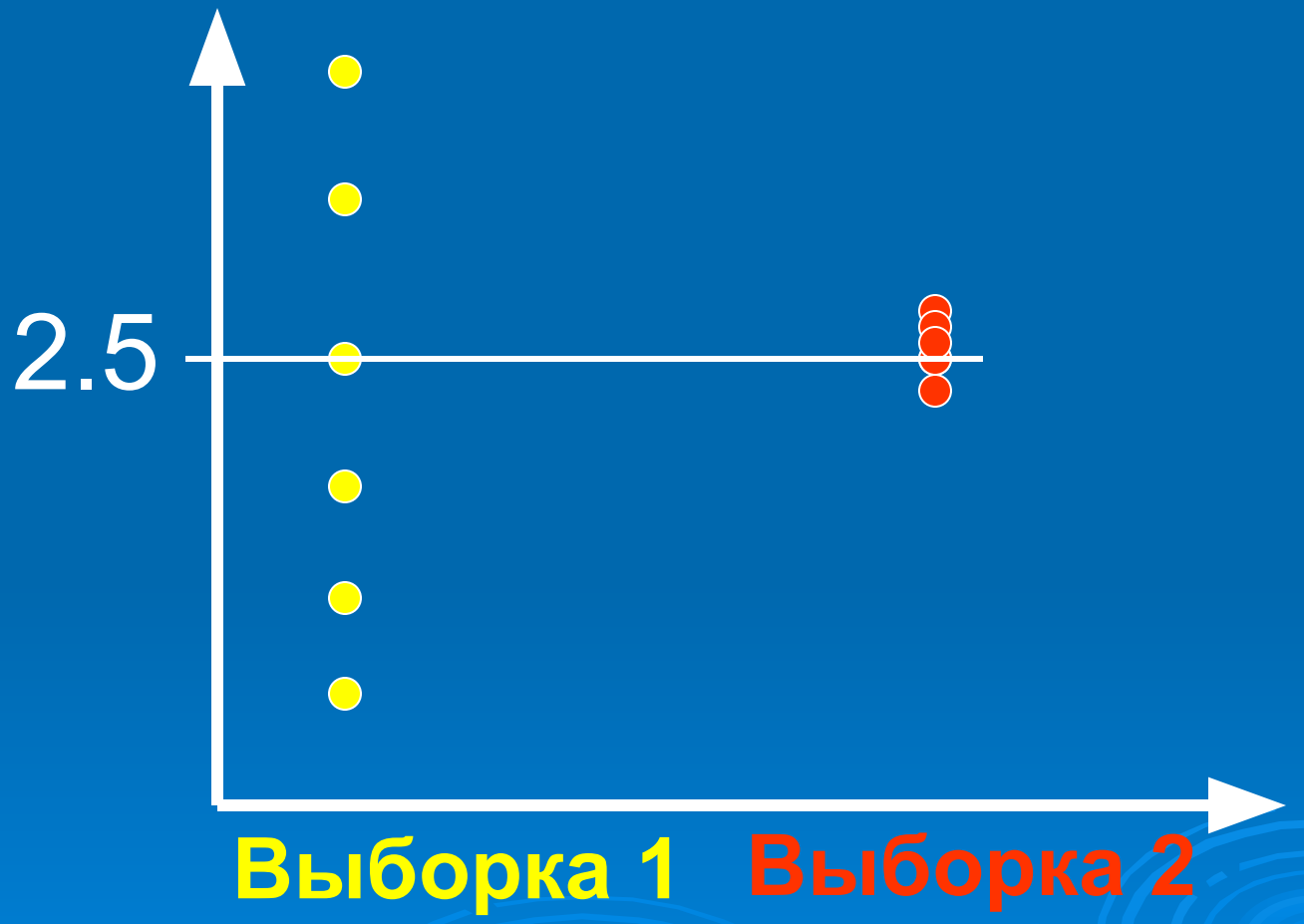
# ВЗВЕШЕННАЯ СРЕДНЯЯ (англ.: *Weighted mean*):

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \ldots + \bar{x}_k n_k}{\sum n_k}$$

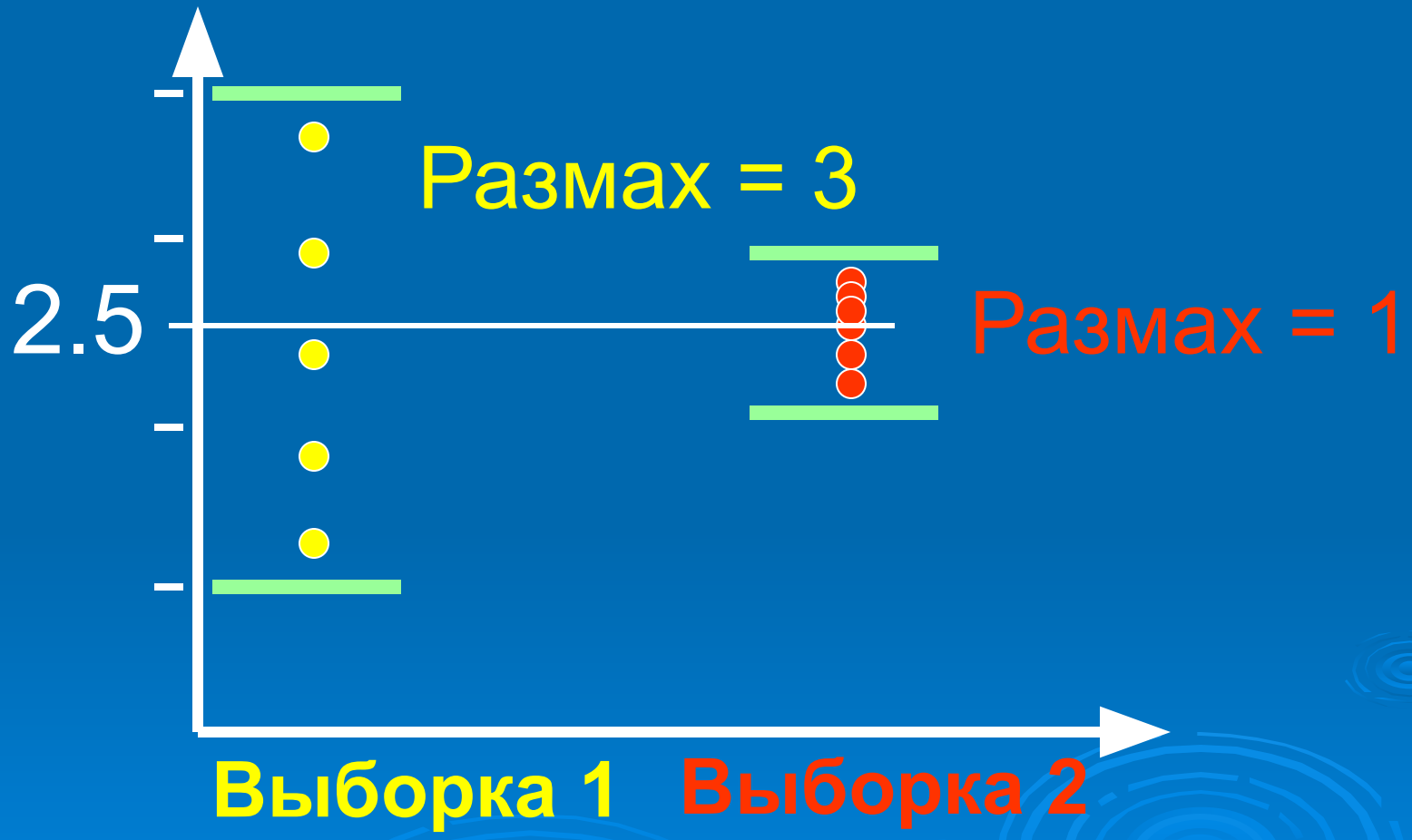
# СРЕДНЯЯ ГЕОМЕТРИЧЕСКАЯ (англ.: *Geometric mean*):

$$\bar{x}_g = \sqrt[n]{x_1 \times x_2 \ldots \times x_n}$$

# Одинаковы ли выборки ???????



# Размах



# Размах одинаковый

10 15 20 25 30 35 40 45 50
10 28 28 30 30 30 32 32 50

X = 30; размах = 40
X = 30, размах = 40

**Выборки различаются!**

Находим расстояние, на котором находится каждая единица изучаемой выборки от среднего значения:

$$(x_i - \bar{x})$$

**Избавляемся от отрицательных значений**

$$(x_i - \bar{x})^2$$

# Усредняем вычисленные расстояния и получаем *дисперсию* (англ.: *variance*):

**SS (sum of squares) – сумма квадратов**

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

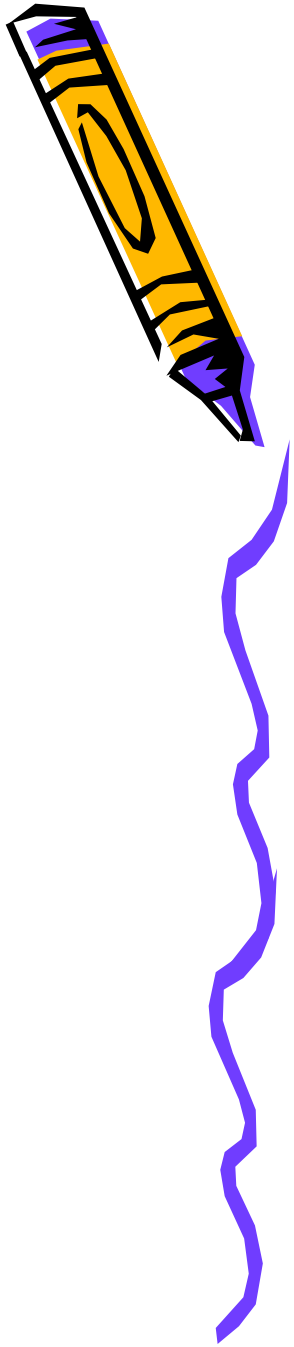
Извлекая корень из дисперсии, получаем *стандартное отклонение* (англ.: *standard deviation*; SD):

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

# Несмещенные оценки дисперсии и стандартного отклонения (для малых *n*):

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

**ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ (*df*)**



# 2.3. Медиана и процентили

**Медиана (*Me; англ.: Median*) - значение, которое делит распределение ровно пополам.**

**Для нахождения:**

- выстроить данные min ⟶ max
- если ***n*** нечетное, ищем центральное значение (n+1)/2
- если ***n*** четное, находим среднее между двумя центральными значениями

# Медиана

Значение, половина данных в совокупности больше которого, а половина – меньше

**n – нечетное:**

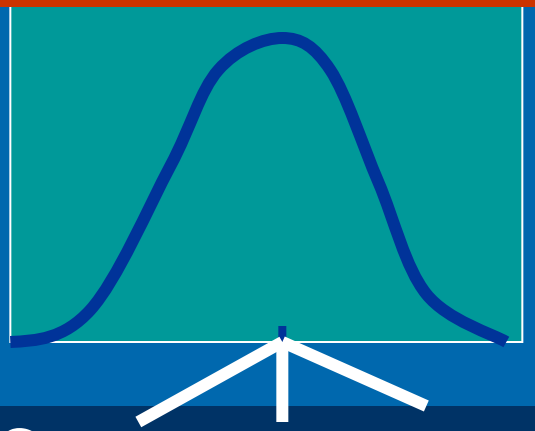
34 36 37 39 40 41 42 43 79

n=9

$Me=X_{(n+1)/2}=X_{(9+1)/2}=X_5=40$

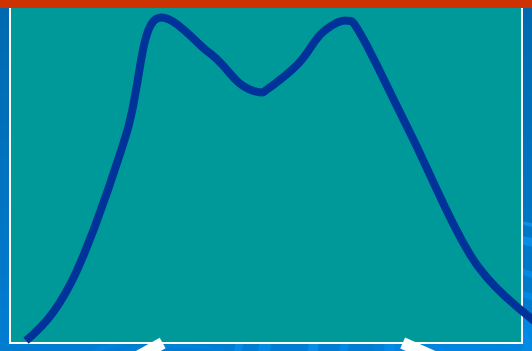
X=43.4

Симметричное унимодальное



Средняя, мода, медиана

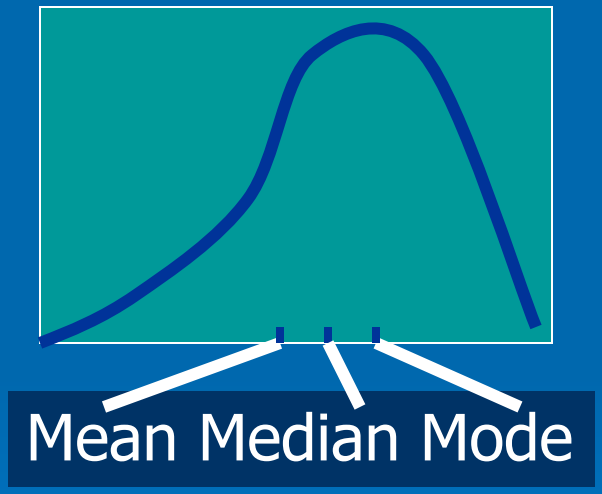
Симметричное бимодальное



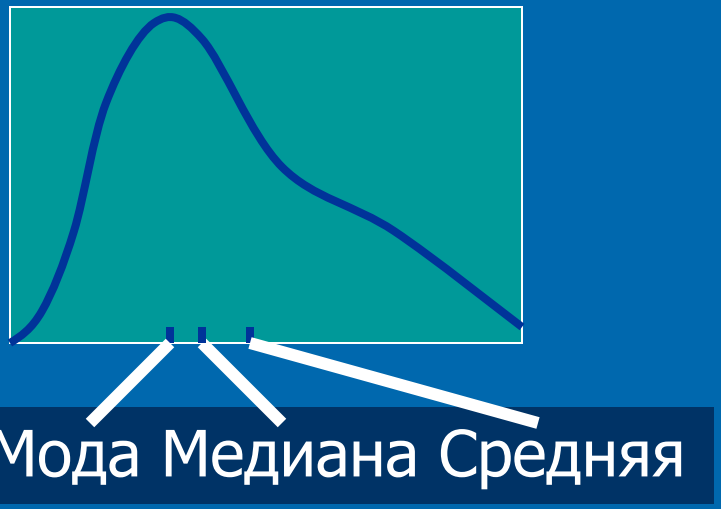
Mode Mean Median Mode

# Медиана

Скошенное влево распределение




Скошенное вправо распределение




**n – четное:**

30 33 34 37 40 41 42 43 44 45

n=10

Me= $X_{(n+1)/2}=X_{(9+1)/2}=X_{5.5}=$

$(X_5+X_6)/2 = (40+41)/2 =$ **40.5**

X = 38.9

# ВЫВОДЫ:

- Если известно, что выборка скорее всего принадлежит к совокупности с нормальным распределением, для ее описания лучше использовать выборочное ***среднее*** и выборочное ***стандартное отклонение***.

# ВЫВОДЫ:

- Если же известно, что распределение в совокупности отличается от нормального, следует использовать ***медиану***, *25-й* и *75-й* ***процентили***.