IT-ОБРАЗОВАНИЕ
В ПЕТЕРБУРГЕ
И УДАЛЕННО

# Основы теории вероятностей и математической статистики

Преподаватель: Зубоченко Антон

# Математическая статистика

**Математическая статистика** – это область математики, которая посвящена изучению методов и техник анализа данных с целью решения практических задач. Она широко применяется в различных областях, где имеются данные, включая область искусственного интеллекта.

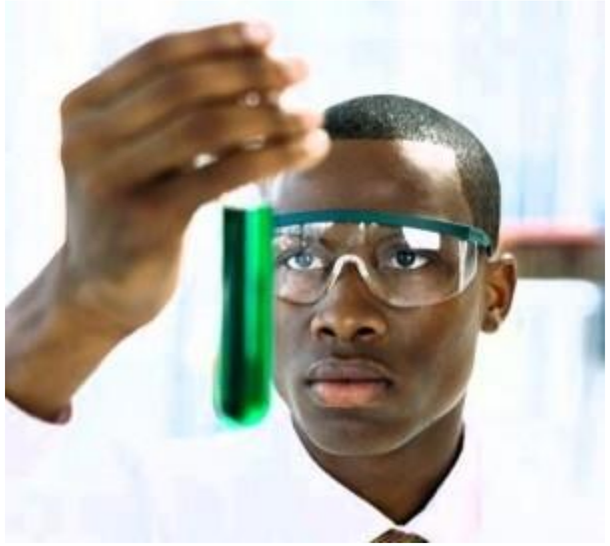
# Теория вероятностей

**Теория вероятностей** – это раздел математики, посвященный исследованию закономерностей случайных явлений, которые могут происходить многократно или в случайный момент времени.

# Основные понятия теории вероятностей

# Основные понятия теории вероятностей

- *Испытание* — эксперимент, который можно провести неограниченное количество раз



# Основные понятия теории вероятностей

- *Испытание* — эксперимент, который можно провести неограниченное количество раз
  - Бросок монетки/кубика

# Основные понятия теории вероятностей

- *Испытание* — эксперимент, который можно провести неограниченное количество раз
  - Бросок монетки/кубика
  - Вытаскивание разноцветных шаров из мешка

# Основные понятия теории вероятностей

- *Испытание* — эксперимент, который можно провести неограниченное количество раз
  - Бросок монетки/кубика
  - Вытаскивание разноцветных шаров из мешка
- *Элементарный исход* — элементарный результат испытания

# Основные понятия теории вероятностей

- *Испытание* — эксперимент, который можно провести неограниченное количество раз
  - Бросок монетки/кубика
  - Вытаскивание разноцветных шаров из мешка
- *Элементарный исход* — элементарный результат испытания
- Множество всех возможных элементарных исходов называется *пространством элементарных исходов* (обозначается Ω)
- Бросок монетки: Ω = {“орёл”, “решка”}

**Задача 1**. Ребёнок играет с 10 буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность, что при случайном расположении букв в ряд он получит слово математика?

**Задача 1**. Ребёнок играет с 10 буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность, что при случайном расположении букв в ряд он получит слово математика?

**Решение.**

**Задача 1**. Ребёнок играет с 10 буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность, что при случайном расположении букв в ряд он получит слово математика?

**Решение.**

1. Формализуем задачу

   Эксперимент — расположение букв в случайном порядке

   Ω (пр-во эл. исходов) — множество всех возможных перестановок букв

   А, А, А, Е, И, К, М, М, Т, Т

   МАТЕМАТИКА

**Требуется** найти P(буквы образовали слово математика) = $\dfrac{\text{кол-во подходящих перестановок}}{\text{общее количество перестановок}}$

**Задача 1**. Ребёнок играет с 10 буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность, что при случайном расположении букв в ряд он получит слово математика?

**Решение.**

1. Формализуем задачу
   **Требуется** найти Р(буквы образовали слово математика) = $\dfrac{\text{кол-во подходящих перестановок}}{\text{общее количество перестановок}}$

2. Найдем кол-во подходящих перестановок
   $А_1, А_2, А_3, М_1, М_2, Т_1, Т_2$, Е, И, К

М А Т Е М А Т И К А

**Задача 1**. Ребёнок играет с 10 буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность, что при случайном расположении букв в ряд он получит слово математика?

**Решение.**

1. Формализуем задачу
   **Требуется** найти Р(буквы образовали слово математика) = $\dfrac{\text{кол-во подходящих перестановок}}{\text{общее количество перестановок}}$

2. Найдем кол-во подходящих перестановок
   $А_1, А_2, А_3, М_1, М_2, Т_1, Т_2$, Е, И, К

М А Т Е М А Т И К А

Способов расставить 3 буквы А на 3 места: $3 \cdot 2 \cdot 1 = 3! = 6$

Способов расставить 2 буквы М на 2 места: $2 \cdot 1 = 2! = 2$

Способов расставить 2 буквы Т на 2 места: $2 \cdot 1 = 2! = 2$

Общее количество перестановок равно $6 \cdot 2 \cdot 2 = 24$

**Задача 1**. Ребёнок играет с 10 буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность, что при случайном расположении букв в ряд он получит слово математика?

**Решение.**

1. Формализуем задачу
   **Требуется** найти Р(буквы образовали слово математика) = $\dfrac{\text{кол-во подходящих перестановок}}{\text{общее количество перестановок}}$

2. Найдем кол-во подходящих перестановок

   М А Т Е М А Т И К А

   Общее количество перестановок равно $6 \cdot 2 \cdot 2 = 24$

3. Посчитаем размер пространства элементарных исходов: $|\Omega|$

$$|\Omega| = 10 \cdot 9 \cdot \ldots \cdot 2 \cdot 1 = 10!$$

**Задача 1**. Ребёнок играет с 10 буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность, что при случайном расположении букв в ряд он получит слово математика?

**Решение.**

1. Формализуем задачу
   **Требуется** найти P(буквы образовали слово математика) = $\dfrac{\text{кол-во подходящих перестановок}}{\text{общее количество перестановок}}$

2. Найдем кол-во подходящих перестановок

   М А Т Е М А Т И К А

   Общее количество перестановок равно $6 \cdot 2 \cdot 2 = 24$

3. Посчитаем размер пространства элементарных исходов: $|\Omega|$

   $$|\Omega| = 10 \cdot 9 \cdot \ldots \cdot 2 \cdot 1 = 10!$$

4. Ответ: $\dfrac{\text{кол-во подходящих перестановок}}{\text{общее количество перестановок}} = \dfrac{3! \cdot 2! \cdot 2!}{10!}$

# Вероятностное пространство

- Иногда вероятности элементарных исходов бывают не одинаковыми

# Вероятностное пространство

- Иногда вероятности элементарных исходов бывают не одинаковыми
- Пусть $\Omega = \{\omega_1, ..., \omega_n\}$ — пространство элементарных исходов
- Присвоим каждому элемент вероятности: $p_1, \ldots, p_n$

$$p_i = \mathsf{P}(\{\omega_i\}), \quad p_1 + \ldots + p_n = 1$$

# Вероятностное пространство

- Иногда вероятности элементарных исходов бывают не одинаковыми
- Пусть $\Omega = \{\omega_1, ..., \boxed{\omega}_n\}$ — пространство элементарных исходов
- Присвоим каждому элемент вероятности: $p_1, \ldots, p_n$

$$p_i = \mathsf{P}(\{\omega_i\}), \quad p_1 + \ldots + p_n = 1$$

# Вероятностное пространство

- Иногда вероятности элементарных исходов бывают не одинаковыми
- Пусть $\Omega = \{\omega_1, ..., \omega_n\}$ — пространство элементарных исходов
- Присвоим каждому элемент вероятности: $p_1, \ldots, p_n$

$$p_i = \mathsf{P}(\{\omega_i\}), \quad p_1 + \ldots + p_n = 1$$

# Вероятностное пространство

- Иногда вероятности элементарных исходов бывают не одинаковыми
- Пусть $\Omega = \{\omega_1, ..., \omega_n\}$ — пространство элементарных исходов
- Присвоим каждому элемент вероятности: $p_1, \ldots, p_n$

$$p_i = \mathsf{P}(\{\omega_i\}), \quad p_1 + \ldots + p_n = 1$$

- Вероятностное пространство:
  - пространство элементарных исходов
  - соответствующие вероятности

# События

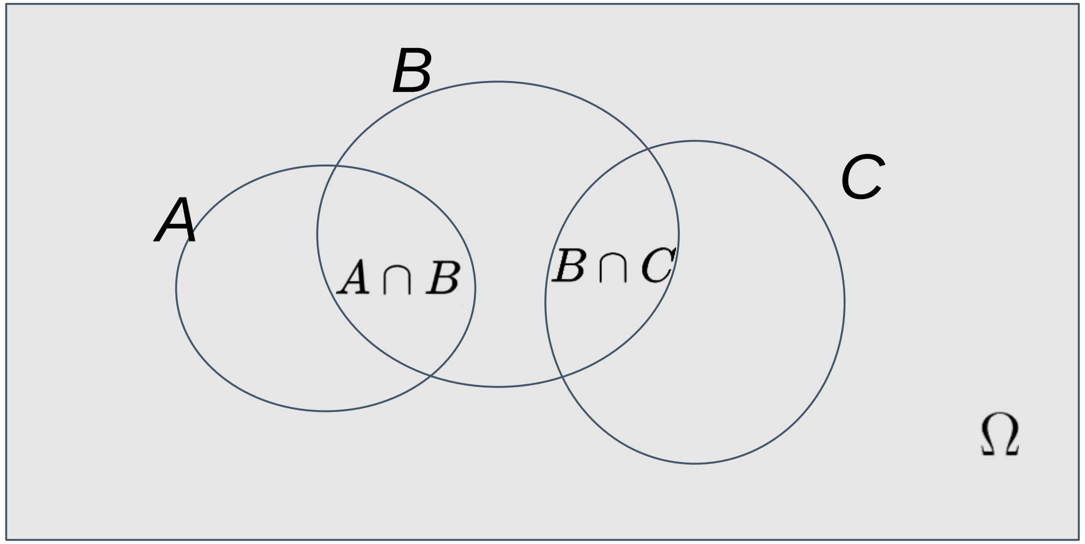
**Определение**. *Событие* — некоторое подмножество вероятностного пространства

- Событие из предыдущей задачи:
    - {буквы образуют слово “МАТЕМАТИКА”}

# События

**Определение**. *Событие* — некоторое подмножество вероятностного пространства

- Событие из предыдущей задачи:
  - {буквы образуют слово “МАТЕМАТИКА”}



# Вероятность события

- Вероятность события $A \subseteq \Omega$ обозначается как $\mathsf{P}(A)$
- $\mathsf{P}(A)$ вычисляется как *сумма вероятностей элементарных исходов* в *A*
  - Вероятность — число от 0 до 1
  - Сумма вероятностей элементарных исходов равна 1

# Свойства вероятности

1. $\mathsf{P}(\Omega) = 1$
2. Для любого события *A* ⊂ Ω справедливо $0 \leq \mathsf{P}(A) \leq 1$
3. Пусть *A*, *B* ⊂ Ω не пересекаются. Тогда
$$\mathsf{P}(A \cup B) = P(A) + P(B)$$
4. Для любых событий *A, B* ⊂ Ω справедливо
$$\mathsf{P}(A \cup B) = P(A) + P(B) - P(A \cap B)$$
5. Пусть *A* ⊂ *B* ⊂ Ω. Тогда
$$\mathsf{P}(B \setminus A) = \mathsf{P}(B) - \mathsf{P}(A)$$
$$\mathsf{P}(\overline{A}) = 1 - \mathsf{P}(A)$$

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: А и В. Нуклеотид равен А с вероятностью $p$ и В с вероятностью $1 - p$. Под воздействием облучения нуклеотид А мутирует с вероятностью $q$, а нуклеотид В — с вероятностью $r$. С какой вероятностью мутирует не более одного нуклеотида?

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: A и B. Нуклеотид равен A с вероятностью $p$ и B с вероятностью $1 - p$. Под воздействием облучения нуклеотид A мутирует с вероятностью $q$, а нуклеотид B — с вероятностью $r$. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Определить пространство элементарных исходов
2. Описать событие, вероятность которого нужно найти
3. Вычислить вероятность каждого элементарного исхода
4. Вычислить ответ как сумму вероятностей элементарных исходов в нужном событии

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: A и B. Нуклеотид равен A с вероятностью *p* и B с вероятностью 1 — *p*. Под воздействием облучения нуклеотид A мутирует с вероятностью *q*, а нуклеотид B — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

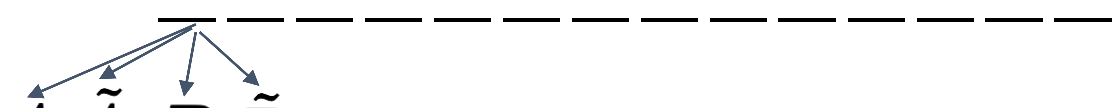
**Решение.**

1. Пространство элементарных исходов

$$ABAAABBBAAABBA$$

$$\Downarrow$$

$$A\tilde{B}AA\tilde{A}B\tilde{B}B\tilde{A}A\tilde{A}\tilde{B}BA$$



$$A\ \tilde{A}\ B\ \tilde{B}$$

Всего последовательностей: $|\Omega| = 4 \cdot 4 \cdot \ldots \cdot 4 = 4^{100}$

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: A и B. Нуклеотид равен A с вероятностью *p* и B с вероятностью 1 — *p*. Под воздействием облучения нуклеотид A мутирует с вероятностью *q*, а нуклеотид B — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов

$$ABAAABBBAAABBA$$

$$\Downarrow$$

$$A\tilde{B}AA\tilde{A}B\tilde{B}B\tilde{A}A\tilde{A}\tilde{B}BA$$

$$\_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_$$

$$A\ \tilde{A}\ B\ \tilde{B}$$

| | | | |
|---|---|---|---|
| $AA$ | $A\tilde{A}$ | $AB$ | $A\tilde{B}$ |
| $\tilde{A}A$ | $\tilde{A}\tilde{A}$ | $\tilde{A}B$ | $\tilde{A}\tilde{B}$ |
| $BA$ | $B\tilde{A}$ | $BB$ | $B\tilde{B}$ |
| $\tilde{B}A$ | $\tilde{B}\tilde{A}$ | $\tilde{B}B$ | $\tilde{B}\tilde{B}$ |

Всего последовательностей: $|\Omega| = 4 \cdot 4 \cdot \ldots \cdot 4 = 4^{100}$

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: A и B. Нуклеотид равен A с вероятностью *p* и B с вероятностью 1 — *p*. Под воздействием облучения нуклеотид A мутирует с вероятностью *q*, а нуклеотид B — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов
2. Событие: мутирует не больше одного нуклеотида
   Всего таких последовательностей:

$$101 \cdot 2^{100}$$

| | | | |
|---|---|---|---|
| $\boxed{AA}$ | $\boxed{A\tilde{A}}$ | $\boxed{AB}$ | $\boxed{A\tilde{B}}$ |
| $\boxed{\tilde{A}A}$ | $\tilde{A}\tilde{A}$ | $\boxed{\tilde{A}B}$ | $\tilde{A}\tilde{B}$ |
| $\boxed{BA}$ | $\boxed{B\tilde{A}}$ | $\boxed{BB}$ | $\boxed{B\tilde{B}}$ |
| $\boxed{\tilde{B}A}$ | $\tilde{B}\tilde{A}$ | $\boxed{\tilde{B}B}$ | $\tilde{B}\tilde{B}$ |

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: А и В. Нуклеотид равен А с вероятностью *p* и В с вероятностью 1 — *p*. Под воздействием облучения нуклеотид А мутирует с вероятностью *q*, а нуклеотид В — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов:
2. Событие: мутирует не больше одного нуклеотида
3. Задаем вероятности элементарных исходов
   a. Вероятности отдельного нуклеотида

| $A$ | $\tilde{A}$ | $B$ | $\tilde{B}$ |
|---|---|---|---|
| $p(1-q)$ | $pq$ | $(1-p)(1-r)$ | $(1-p)r$ |

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: A и B. Нуклеотид равен A с вероятностью *p* и B с вероятностью 1 — *p*. Под воздействием облучения нуклеотид A мутирует с вероятностью *q*, а нуклеотид B — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов:
2. Событие: мутирует не больше одного нуклеотида
3. Задаем вероятности элементарных исходов
    a. Вероятности отдельного нуклеотида

| $A$ | $\tilde{A}$ | $B$ | $\tilde{B}$ |
|---|---|---|---|
| $p(1-q)$ | $pq$ | $(1-p)(1-r)$ | $(1-p)r$ |

    b. Вероятность конкретной последовательности ДНК с $n_1, n_2, n_3, n_4$ букв:

$$(p(1-q))^{n_1} \cdot (pq)^{n_2} \cdot ((1-p)(1-r))^{n_3} \cdot ((1-p)r)^{n_4}$$

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: A и B. Нуклеотид равен A с вероятностью *p* и B с вероятностью 1 — *p*. Под воздействием облучения нуклеотид A мутирует с вероятностью *q*, а нуклеотид B — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов:
2. Событие: мутирует не больше одного нуклеотида
3. Вероятности отдельного нуклеотида

| $A$ | $\tilde{A}$ | $B$ | $\tilde{B}$ |
|---|---|---|---|
| $p(1-q)$ | $pq$ | $(1-p)(1-r)$ | $(1-p)r$ |

4. P(конкретный нуклеотид мутировал) = $pq + (1-p)r = \alpha$
   P(конкретный нуклеотид не мутировал) = $q - pq + pr - r + 1 = 1 - \alpha$

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: А и В. Нуклеотид равен А с вероятностью *p* и В с вероятностью 1 — *p*. Под воздействием облучения нуклеотид А мутирует с вероятностью *q*, а нуклеотид В — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов:
2. Событие: мутирует не больше одного нуклеотида
3. Вероятности отдельного нуклеотида

| $A$ | $\tilde{A}$ | $B$ | $\tilde{B}$ |
|---|---|---|---|
| $p(1-q)$ | $pq$ | $(1-p)(1-r)$ | $(1-p)r$ |

4. P(конкретный нуклеотид мутировал) = $pq + (1-p)r = \alpha$
   P(конкретный нуклеотид не мутировал) = $q - pq + pr - r + 1 = 1 - \alpha$
5. P(нет мутаций) = P($A$) = $(1-\alpha)^{100}$

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: А и В. Нуклеотид равен А с вероятностью *p* и В с вероятностью 1 — *p*. Под воздействием облучения нуклеотид А мутирует с вероятностью *q*, а нуклеотид В — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов:
2. Событие: мутирует не больше одного нуклеотида
3. Вероятности отдельного нуклеотида

| $A$ | $\tilde{A}$ | $B$ | $\tilde{B}$ |
|---|---|---|---|
| $p(1-q)$ | $pq$ | $(1-p)(1-r)$ | $(1-p)r$ |

4. P(конкретный нуклеотид мутировал) = $pq + (1-p)r = \alpha$
   P(конкретный нуклеотид не мутировал) = $q - pq + pr - r + 1 = 1 - \alpha$
5. P(нет мутаций) = P(*A*) = $(1-\alpha)^{100}$
6. P(ровно одна мутация) = P(*B*)
   $B_k$ = {произошла ровно одна мутация на *k*-ой позиции}. $\mathsf{P}(B_k) = \alpha(1-\alpha)^{99}$

$$\mathsf{P}(B) = 100 \cdot \alpha(1-\alpha)^{99}$$

**Задача 3.** ДНК Снежного человека — это последовательность из 100 нуклеотидов двух типов: А и В. Нуклеотид равен А с вероятностью *p* и В с вероятностью 1 — *p*. Под воздействием облучения нуклеотид А мутирует с вероятностью *q*, а нуклеотид В — с вероятностью *r*. С какой вероятностью мутирует не более одного нуклеотида?

**Решение.**

1. Пространство элементарных исходов:
2. Событие: мутирует не больше одного нуклеотида
3. Вероятности отдельного нуклеотида

| $A$ | $\tilde{A}$ | $B$ | $\tilde{B}$ |
|---|---|---|---|
| $p(1-q)$ | $pq$ | $(1-p)(1-r)$ | $(1-p)r$ |

4. P(конкретный нуклеотид мутировал) = $pq + (1-p)r = \alpha$
   P(конкретный нуклеотид не мутировал) = $q - pq + pr - r + 1 = 1 - \alpha$
5. P(нет мутаций) = P(*A*) = $(1-\alpha)^{100}$
6. P(ровно одна мутация) = P(*B*) = $100 \cdot \alpha(1-\alpha)^{99}$
7. P(*A* ∪ *B*) = P(*A*) + P(*B*) = $\boxed{(1-\alpha)^{100} + \alpha(1-\alpha)^{99}}$

# Резюме

- Вероятностное пространство — это множество Ω, элементам которого присвоены вероятности $p_i = \mathsf{P}(\{\omega_i\})$
- Событие — это произвольное подмножество вероятностного пространства
  - Вероятность события равна сумме вероятностей элементарных исходов в событии

# Независимость

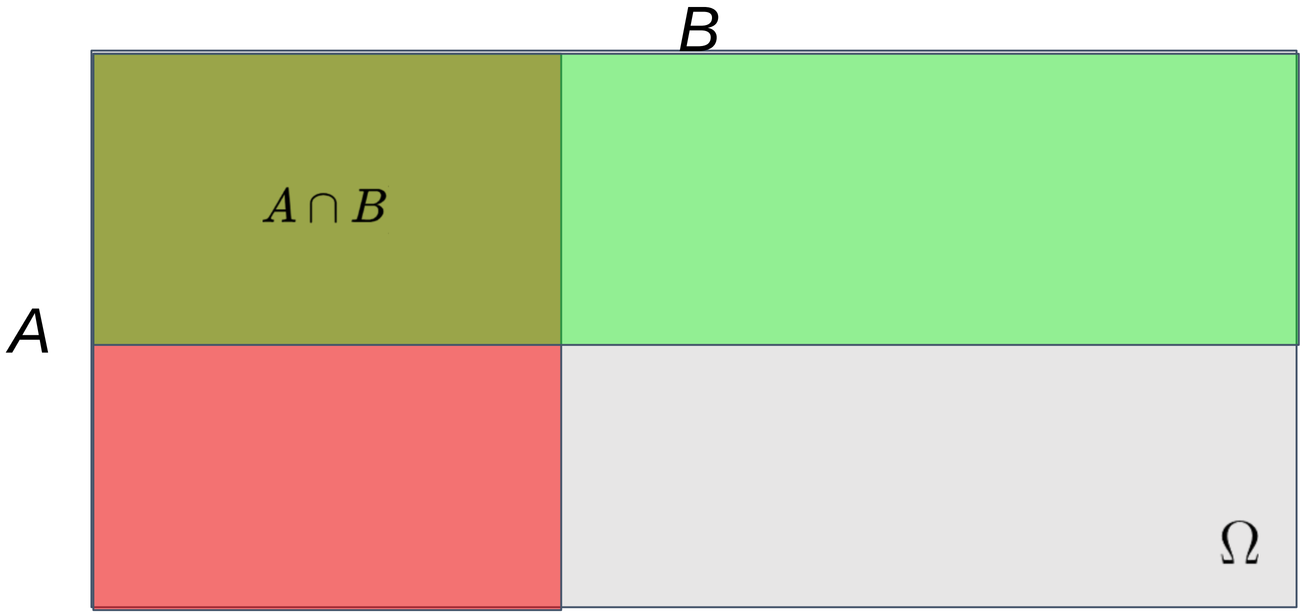
**Определение.** События $A, B \in \Omega$ называются *независимыми*, если

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B)$$

# Независимость

**Определение.** События $A, B \in \Omega$ называются *независимыми*, если

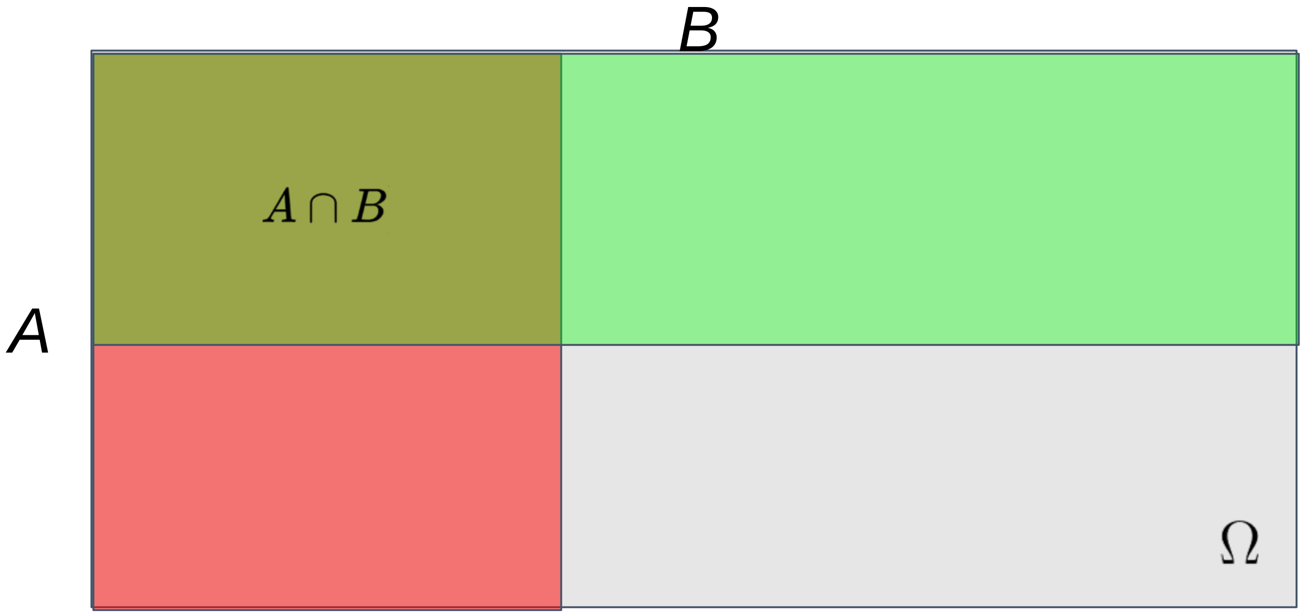
$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B)$$



# Независимость

**Определение.** События $A, B \in \Omega$ называются *независимыми*, если

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B)$$

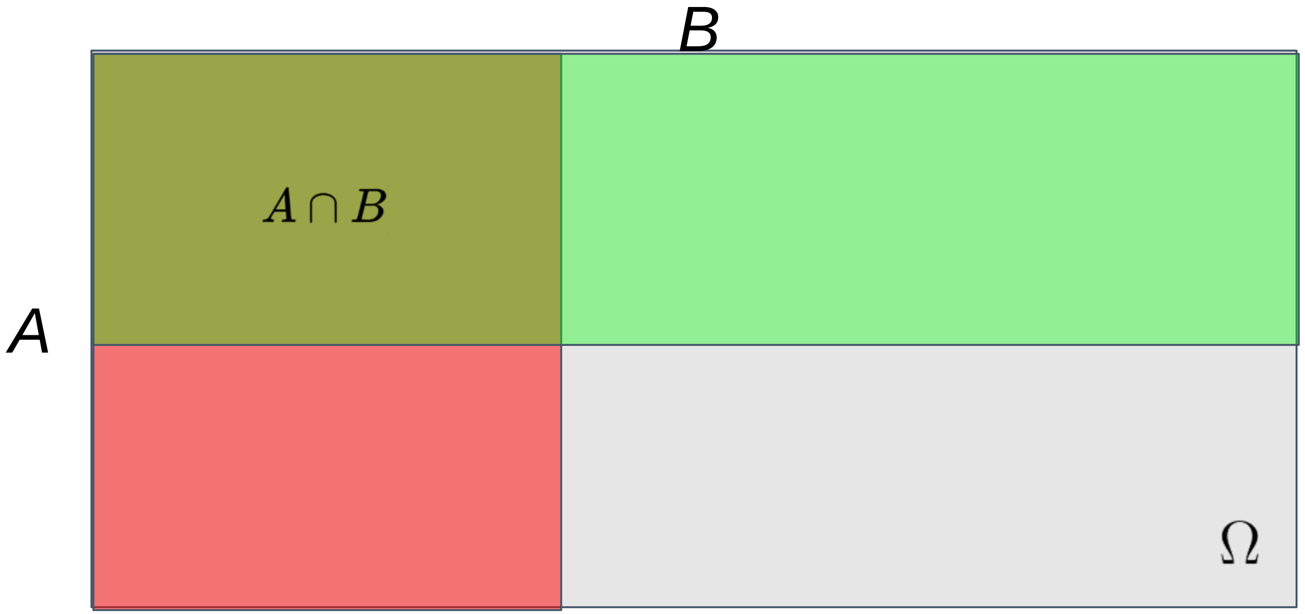



$$\frac{\mathsf{P}(A \cap B)}{P(A)} = \frac{\mathsf{P}(B)}{P(\Omega)} = P(B)$$

# Независимость

**Определение.** События $A, B \in \Omega$ называются *независимыми*, если

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B)$$




- Если *A* и *B* независимы, то знание, случилось ли *A*, не даёт нам никакой информации о том, случилось ли *B.*

# Пример: последовательные броски монетки

- Два последовательных броска монеты — независимые события.
- Вероятность «орла» в каждом отдельном эксперименте равна $\frac{1}{2}$
- Вероятность «орла» в обоих экспериментах равна $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$

# Независимость и причинно-следственная связь

- Событие “школьник поступил в сильный вуз” зависит от события “школьник хорошо сдал экзамен”
- Зависимость есть и наоборот, хотя причинно-следственная связь в нужную сторону
- $A$ не зависит от $B \Leftrightarrow B$ не зависит от $A$

**Задача 1**. A и B играют в игру из *n* раундов. В каждом раунде игрок A подбрасывает 3 игральные кости, а игрок B — 2 кости одновременно с игроком A. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

**Задача 1**. А и В играют в игру из $n$ раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за $n$ раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6).

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим *p* = P(в раунде А выбросил 6, а В не выбросил 6) и *q* = P(в раунде никто не выбросил 6).

$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2$$

P(А выбросил 6) P(В не выбросил 6)

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6).

$$p = \underbrace{\boxed{\left(1 - \left(\tfrac{5}{6}\right)^3\right)}}_{\text{P(А выбросил 6)}} \cdot \underbrace{\left(\tfrac{5}{6}\right)^2}_{\text{P(В не выбросил 6)}}$$

**Задача 1**. A и B играют в игру из *n* раундов. В каждом раунде игрок A подбрасывает 3 игральные кости, а игрок B — 2 кости одновременно с игроком A. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим *p* = P(в раунде A выбросил 6, а B не выбросил 6) и *q* = P(в раунде никто не выбросил 6).

$$p = \underbrace{\left(1 - \left(\tfrac{5}{6}\right)^3\right)}_{\text{P(A выбросил 6)}} \cdot \underbrace{\boxed{\left(\tfrac{5}{6}\right)^2}}_{\text{P(B не выбросил 6)}}$$

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6)

$$p = \left(1 - \left(\tfrac{5}{6}\right)^3\right) \cdot \left(\tfrac{5}{6}\right)^2$$

P(A выбросил 6) P(B не выбросил 6)

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим *p* = P(в раунде А выбросил 6, а В не выбросил 6) и *q* = P(в раунде никто не выбросил 6)

$$p = \left(1 - \left(\tfrac{5}{6}\right)^3\right) \cdot \left(\tfrac{5}{6}\right)^2 \qquad q = \left(\tfrac{5}{6}\right)^3 \cdot \left(\tfrac{5}{6}\right)^2 = \left(\tfrac{5}{6}\right)^5$$

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6).
$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2 \qquad q = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \left(\frac{5}{6}\right)^5$$

2. Обозначим $M_k$ = {А выиграл в $k$-ом раунде}. Найдём P($M_k$).

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6)

$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2 \qquad q = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \left(\frac{5}{6}\right)^5$$

2. Обозначим $M_k$ = {А выиграл в $k$-ом раунде}. Найдём $\mathsf{P}(M_k)$.

$$\mathsf{P}(M_k) = q^{k-1} \cdot p$$

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6).
$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2 \qquad q = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \left(\frac{5}{6}\right)^5$$

2. Обозначим $M_k$ = {А выиграл в $k$-ом раунде}. Найдём P($M_k$).
$$\mathsf{P}(M_k) = q^{k-1} \cdot p$$

3. Вероятность победы складывается из $\mathsf{P}(M_k)$ по всем k = 1, …, n. Найдём $\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n)$

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6)
$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2 \qquad q = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \left(\frac{5}{6}\right)^5$$
2. Обозначим $M_k$ = {А выиграл в $k$-ом раунде}. Найдём P($M_k$).
$$\mathsf{P}(M_k) = q^{k-1} \cdot p$$
3. Вероятность победы складывается из $\mathsf{P}(M_k)$ по всем k = 1, …, n. Найдём $\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n)$

$$\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n) =$$

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6)
$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2 \qquad q = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \left(\frac{5}{6}\right)^5$$
2. Обозначим $M_k$ = {А выиграл в $k$-ом раунде}. Найдём P($M_k$).
$$\mathsf{P}(M_k) = q^{k-1} \cdot p$$
3. Вероятность победы складывается из $\mathsf{P}(M_k)$ по всем k = 1, …, n. Найдём $\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n)$
$$\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n) = q^0 \cdot p + q^1 \cdot p + \ldots + q^{n-1} \cdot p =$$

**Задача 1**. А и В играют в игру из *n* раундов. В каждом раунде игрок А подбрасывает 3 игральные кости, а игрок В — 2 кости одновременно с игроком А. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за *n* раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде А выбросил 6, а В не выбросил 6) и $q$ = P(в раунде никто не выбросил 6)

$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2 \qquad q = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \left(\frac{5}{6}\right)^5$$

2. Обозначим $M_k$ = {А выиграл в $k$-ом раунде}. Найдём $\mathsf{P}(M_k)$.

$$\mathsf{P}(M_k) = q^{k-1} \cdot p$$

3. Вероятность победы складывается из $\mathsf{P}(M_k)$ по всем k = 1, …, n. Найдём $\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n)$

$$\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n) = q^0 \cdot p + q^1 \cdot p + \ldots + q^{n-1} \cdot p = p \cdot \frac{1-q^n}{1-q}$$

**Задача 1**. A и B играют в игру из $n$ раундов. В каждом раунде игрок A подбрасывает 3 игральные кости, а игрок B — 2 кости одновременно с игроком A. Как только один из игроков выбрасывает на одной из костей «шестёрку», игра заканчивается победой этого игрока. Если шестёрка появилась одновременно (или не появилась вообще за $n$ раундов), объявляется ничья. Какова вероятность победы первого игрока?

**Решение.**

1. Вычислим $p$ = P(в раунде A выбросил 6, а B не выбросил 6) и $q$ = P(в раунде никто не выбросил 6).
$$p = \left(1 - \left(\frac{5}{6}\right)^3\right) \cdot \left(\frac{5}{6}\right)^2 \qquad q = \left(\frac{5}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = \left(\frac{5}{6}\right)^5$$
2. Обозначим $M_k$ = {A выиграл в $k$-ом раунде}. Найдём P($M_k$).
$$\mathsf{P}(M_k) = q^{k-1} \cdot p$$
3. Вероятность победы складывается из $\mathsf{P}(M_k)$ по всем k = 1, …, n. Найдём $\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n)$
$$\mathsf{P}(M_1) + \ldots + \mathsf{P}(M_n) = q^0 \cdot p + q^1 \cdot p + \ldots + q^{n-1} \cdot p = p \cdot \frac{1-q^n}{1-q}$$
4. Ответ: $\boxed{p \cdot \frac{1-q^n}{1-q}}$, где $p$, $q$ найдены выше.

# Попарная независимость

**Определение.** События $A_1, \ldots, A_n$ называются *попарно независимыми*, если для любых *i*, *j* независимы $A_i, A_j$ .

# Независимость в совокупности

**Определение.** События $A_1, \ldots, A_n$ называются *независимыми в совокупности*, если для любого поднабора $A_{i_1}, \ldots, A_{i_k}$ справедливо равенство $\mathsf{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathsf{P}(A_{i_1}) \cdot \ldots \cdot \mathsf{P}(A_{i_k})$

# Независимость в совокупности

**Определение.** События $A_1, \ldots, A_n$ называются *независимыми в совокупности*, если для любого поднабора $A_{i_1}, \ldots, A_{i_k}$ справедливо равенство $\mathsf{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathsf{P}(A_{i_1}) \cdot \ldots \cdot \mathsf{P}(A_{i_k})$

- Попарная независимость следует из независимости в совокупности.

# Независимость в совокупности

**Определение.** События $A_1, \ldots, A_n$ называются *независимыми в совокупности*, если для любого поднабора $A_{i_1}, \ldots, A_{i_k}$ справедливо равенство

$$\mathsf{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathsf{P}(A_{i_1}) \cdot \ldots \cdot \mathsf{P}(A_{i_k})$$

- Попарная независимость следует из независимости в совокупности.
- Справедливо ли обратное?

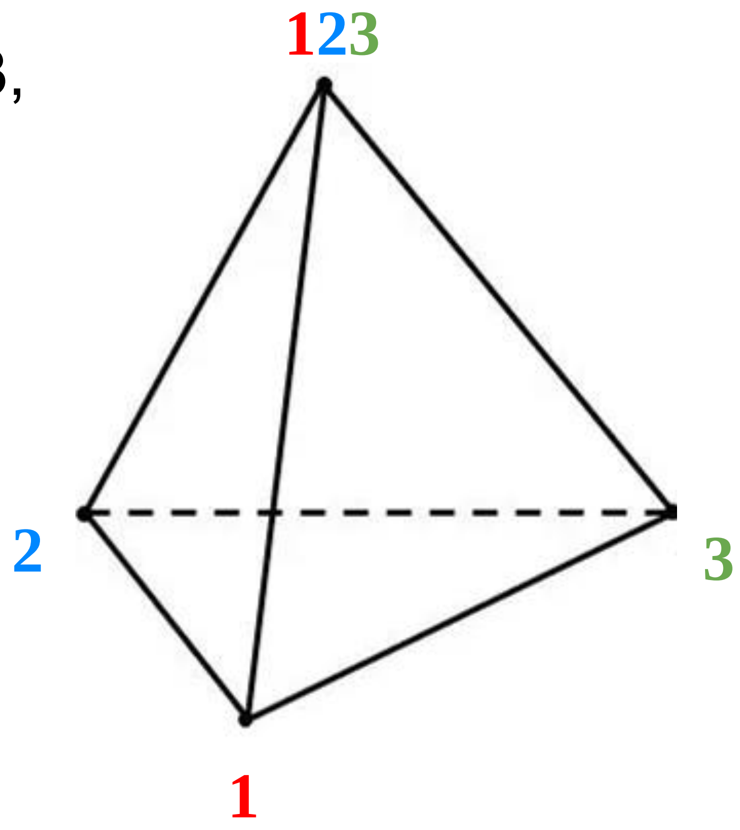
# Пример с тетраэдром

**Утверждение.** Из попарной независимости не следует независимость в совокупности.

# Пример с тетраэдром

**Утверждение.** Из попарной независимости не следует независимость в совокупности.

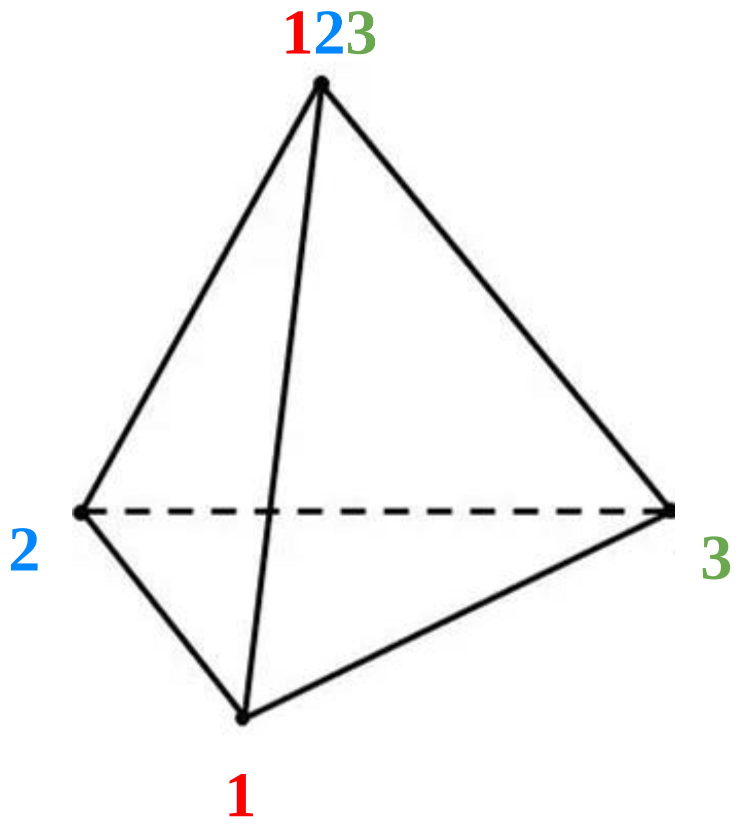
- Напишем на вершинах тетраэдра цифры: 1, 2, 3, 123



# Пример с тетраэдром

**Утверждение.** Из попарной независимости не следует независимость в совокупности.

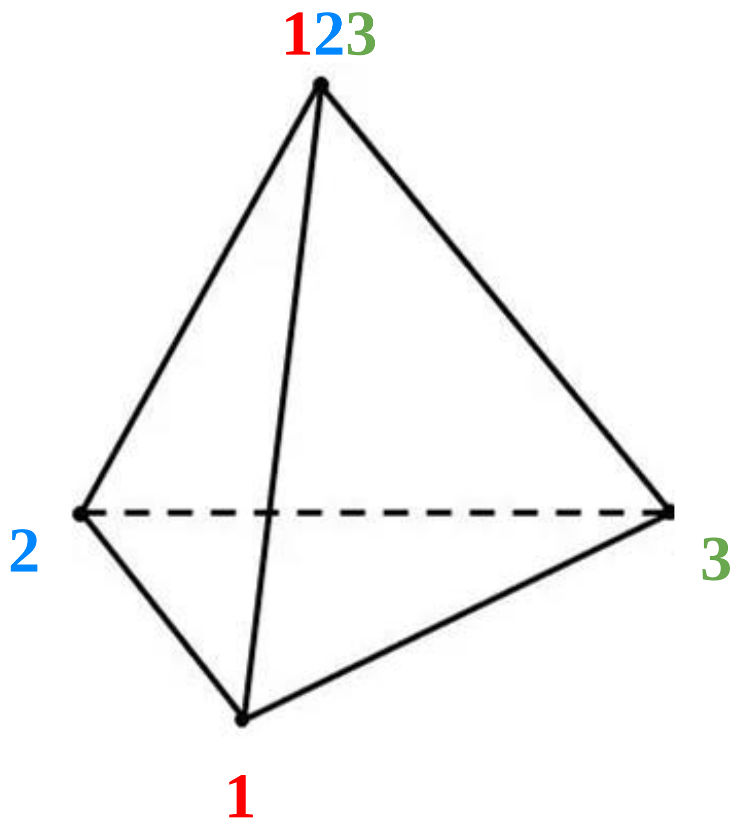
- Напишем на вершинах тетраэдра цифры:
  1, 2, 3, 123.
- Выбираем случайную вершину тетраэдра.



# Пример с тетраэдром

**Утверждение.** Из попарной независимости не следует независимость в совокупности.

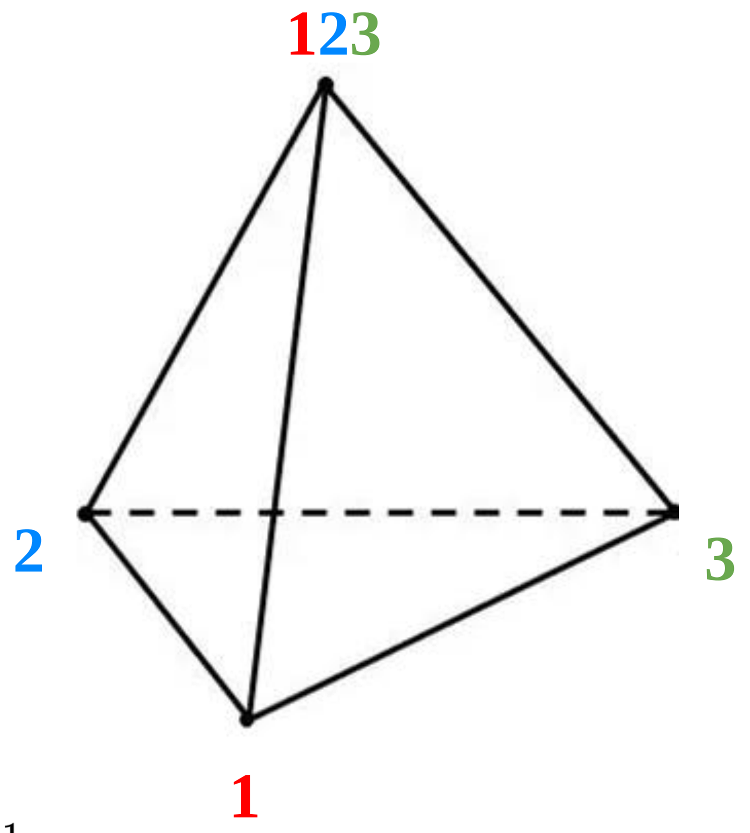
- Напишем на вершинах тетраэдра цифры: 1, 2, 3, 123.
- Выбираем случайную вершину тетраэдра.
- Рассмотрим события:
  - $A$ = {в вершине есть 1},
  - $B$ = {в вершине есть 2},
  - $C$ = {в вершине есть 3}.
- $A$, $B$ и $C$ независимы попарно, но зависимы в совокупности.



# Пример с тетраэдром

**Утверждение.** Из попарной независимости не следует независимость в совокупности.

- Напишем на вершинах тетраэдра цифры: 1, 2, 3, 123.
- Выбираем случайную вершину тетраэдра.
- Рассмотрим события:
  - $A$ = {в вершине есть 1},
  - $B$ = {в вершине есть 2},
  - $C$ = {в вершине есть 3}.




$$\mathsf{P}(A) = \mathsf{P}(B) = \mathsf{P}(C) = \frac{1}{2}$$

$$\mathsf{P}(A \cap B) = \mathsf{P}(A \cap C) = \mathsf{P}(B \cap C) = \frac{1}{4} \qquad \mathsf{P}(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8}$$

# Итоги

- События A и B называются независимыми, если вероятность пересечения событий A и B равна произведению вероятностей событий A и B.
- Независимость событий помогает свести вычисление вероятностей сложных событий к вычислению простых.
- Попарная независимость не влечёт независимости в совокупности.

# Случайная величина

- *Случайная величина* — численное описание результата эксперимента

# Случайная величина

- *Случайная величина* — численное описание результата эксперимента

# Случайная величина

- *Случайная величина* — численное описание результата эксперимента

# Случайная величина

- *Случайная величина* — численное описание результата эксперимента
- *Дискретная случайная величина* принимает конечное количество значений с различными вероятностями
- Мы будем рассматривать только дискретные случайные величины

# Распределение случайной величины

- *Распределение случайной величины* — соответствие между возможными значениями этой величины и их вероятностями
- Распределение можно описать таблицей

| | | | | | |
|---|---|---|---|---|---|
| $X$ | $X_1$ | $X_2$ | $X_3$ | $\dots$ | $X_n$ |
| | $P_1$ | $P_2$ | $P_3$ | $\dots$ | $P_n$ |

$$p + p_2 + p_3 + \dots + p_n = 1$$

# Пример: бросок кубика

- Бросаем игральный кубик, каждая грань выпадает с равной вероятностью
- Случайная величина *X* описывает результат эксперимента:

$$\mathsf{P}(X = 1) = \mathsf{P}(X = 2) = \ldots = \mathsf{P}(X = 6) = \frac{1}{6}$$

- *X* принимает значения 1, 2, …, 6 с вероятностью ⅙.

| *X* | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| | ⅙ | ⅙ | ⅙ | ⅙ | ⅙ | ⅙ |

**Задача 1.** Пусть $X$ — случайная величина, равная результату броска кубика. Найдите распределение случайной величины $(X - 2)^2$.

**Задача 1.** Пусть $X$ — случайная величина, равная результату броска кубика. Найдите распределение случайной величины $(X - 2)^2$.

| $X$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| | ⅙ | ⅙ | ⅙ | ⅙ | ⅙ | ⅙ |
| $(X\text{-}2)^2$ | 1 | 0 | 1 | 4 | 9 | 16 |

**Задача 1.** Пусть $X$ — случайная величина, равная результату броска кубика. Найдите распределение случайной величины $(X - 2)^2$.

| | | | | | | |
|---|---|---|---|---|---|---|
| $X$ | 1 | 2 | 3 | 4 | 5 | 6 |
| | ⅙ | ⅙ | ⅙ | ⅙ | ⅙ | ⅙ |
| $(X-2)^2$ | 1 | 0 | 1 | 4 | 9 | 16 |

| | | | | | |
|---|---|---|---|---|---|
| $(X-2)^2$ | 0 | 1 | 4 | 9 | 16 |
| | ⅙ | 2/6 | ⅙ | ⅙ | ⅙ |

# Пример: распределение роста

- Закон распределения роста:

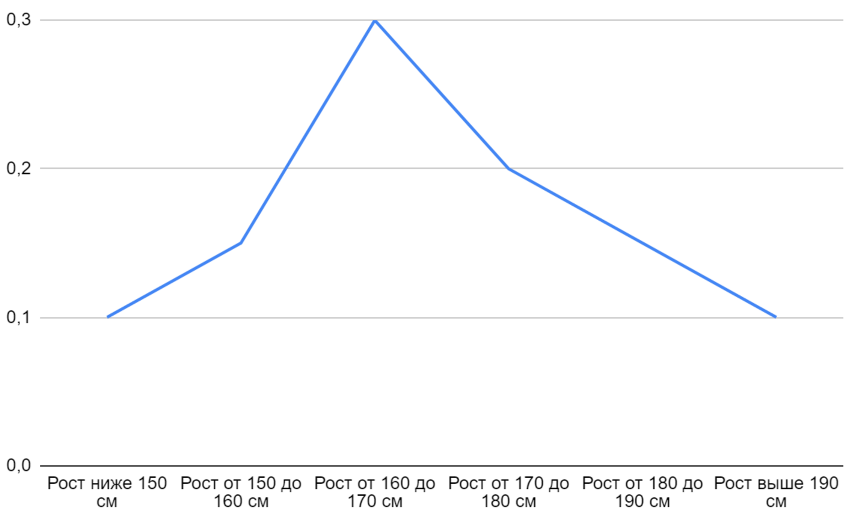
| Рост ниже 150 см | Рост от 150 до 160 см | Рост от 160 до 170 см | Рост от 170 до 180 см | Рост от 180 до 190 см | Рост выше 190 см |
|---|---|---|---|---|---|
| 0,1 | 0,15 | 0,3 | 0,2 | 0,15 | 0,1 |

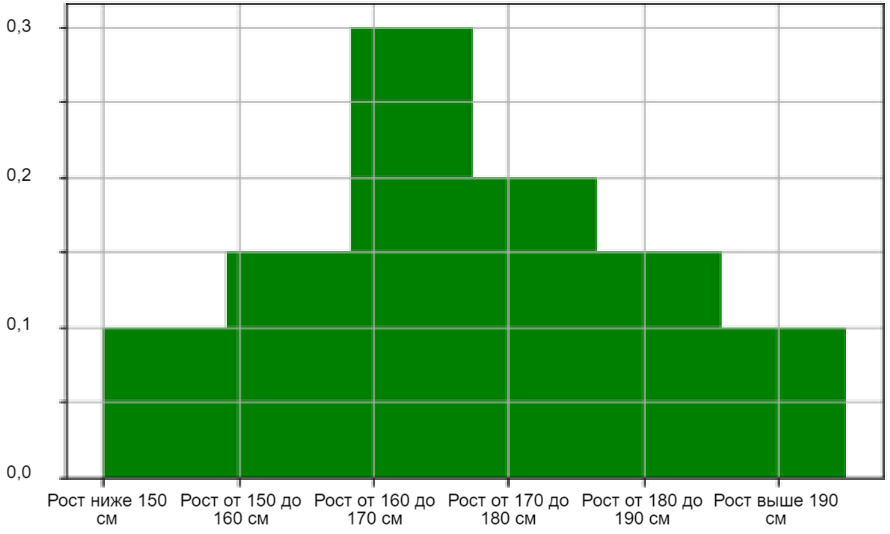
# Пример: распределение роста

- Закон распределения роста:

| Рост ниже 150 см | Рост от 150 до 160 см | Рост от 160 до 170 см | Рост от 170 до 180 см | Рост от 180 до 190 см | Рост выше 190 см |
|---|---|---|---|---|---|
| 0,1 | 0,15 | 0,3 | 0,2 | 0,15 | 0,1 |

- Распределение можно представить в виде графика и в виде гистограммы:



# Пример: нормальное распределение

- Нормальное распределение моделирует многие физические процессы

# Как узнать распределение случайной величины?

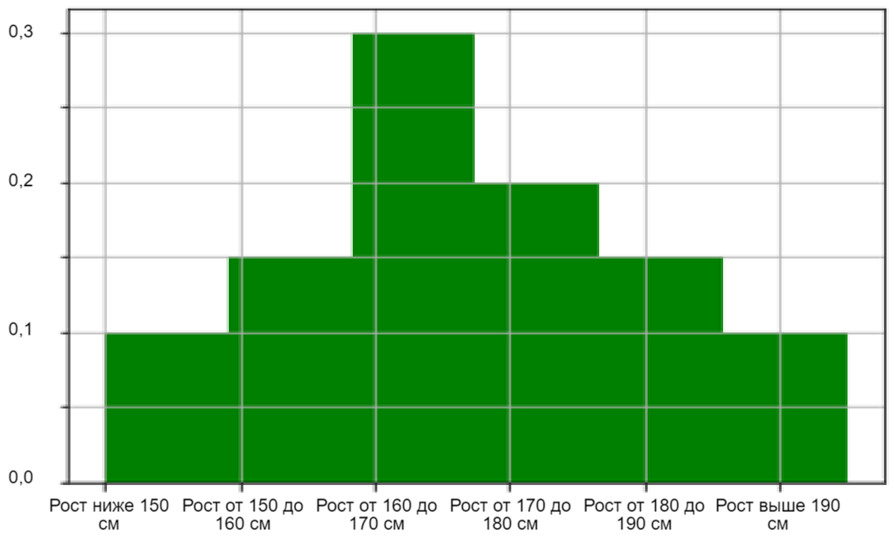
- Теоретически (как в задаче про кубик)
- Собрать данные в ходе эксперимента
  - Распределение роста
  - Масса одинаковых грузов
  - Отклонение стрелка от центра мишени
- Набор данных, соответствующих одной случайной величине, называется *выборкой*
  - Считается, что элементы в выборке независимы друг от друга

# Резюме

- Случайная величина — численная характеристика результата эксперимента
- Случайная величина задаётся своим распределением
- На практике часто распределение случайной величины вычисляется экспериментально

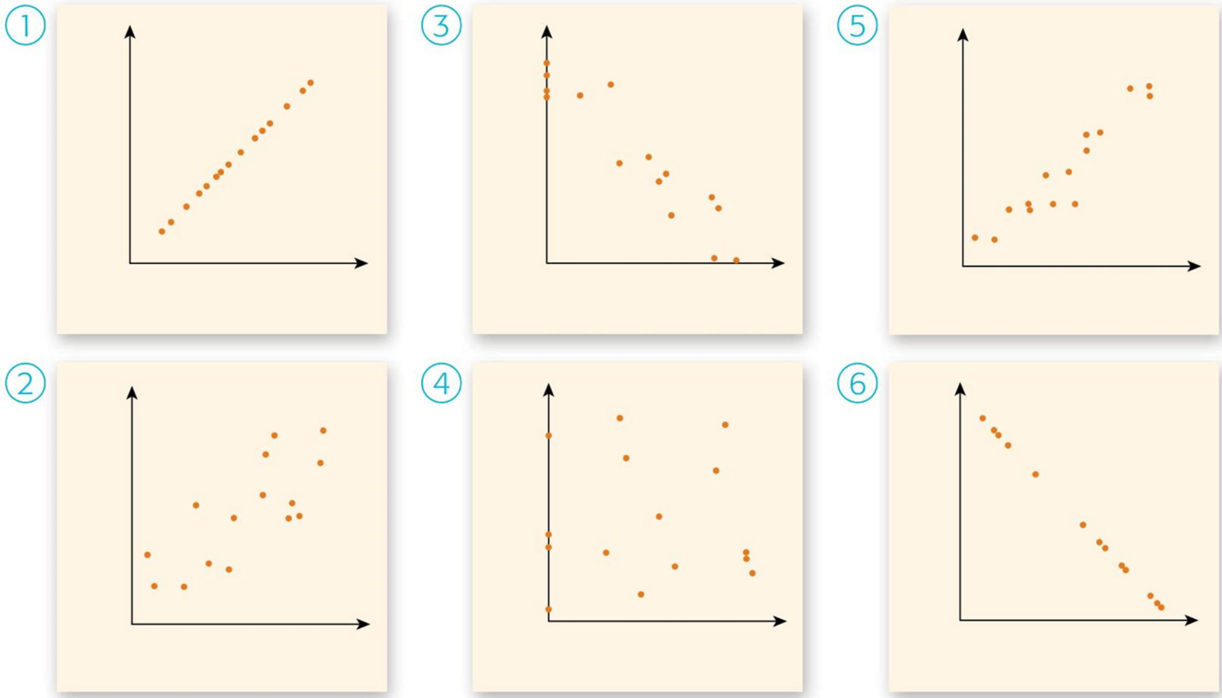
# Многомерное распределение вероятностей

- Для описания одной случайной величины мы ввели распределение вероятностей
- Для описания реальных процессов недостаточно одной случайной величины
- Закон, по которому распределены взаимосвязанные случайные величины (в рамках одного эксперимента), называется *многомерным распределением вероятностей*



# Двумерное распределение вероятностей

- В $n$-мерном случае описать распределение вероятностей сложно, мы этого делать не будем
- В двумерном случае можно визуализировать точки выборки на графике
- Примеры распределений:



# Применения независимости

- В машинном обучении: если признак $X$ и целевая переменная $Y$ независимы, то $X$ вряд ли поможет при предсказании $Y$

Что же такое независимость случайных величин?

# Независимость случайных величин

**Определение.** Случайные величины $X$ и $Y$ называются *независимыми*, если для любых $x$, $y$ выполняется равенство

$$\mathsf{P}(X = x, Y = y) = \mathsf{P}(X = x) \cdot \mathsf{P}(Y = y)$$

# Независимость случайных величин

**Определение.** Случайные величины $X$ и $Y$ называются *независимыми*, если для любых $x$, $y$ выполняется равенство

$$\mathsf{P}(X = x, Y = y) = \mathsf{P}(X = x) \cdot \mathsf{P}(Y = y)$$

Сравним с определением независимости событий:

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B)$$

Иными словами, $X$ и $Y$ независимы, если события $\{X = x\}$ и $\{Y = y\}$ независимы для любых $x$, $y$

# Существуют ли независимые случайные величины?

- Полностью независимых событий в природе почти не бывает
- Независимые случайные величины — это всегда “идеализация”

# Примеры независимых случайных величин

- Результаты последовательных бросков монетки/кубика
- Рост двух случайных жителей планеты
  - Значения двух случайных элементов выборки

# Независимость на графиках двумерных распределений

В каких случаях случайные величины могут быть независимы?

# Независимость на графиках двумерных распределений

В каких случаях случайные величины могут быть независимы?

**Ответ**: только в 4

# Распределение суммы случайных величин

Если известны распределения двух случайных величин по отдельности, то не всегда можно найти распределение суммы

- Пусть случайные величины $X$ и $Y$ — это результаты последовательных бросков одного и того же кубика.
  - Вероятность того, что сумма бросков равна 12, составляет
$$\mathrm{P}(X + Y = 12) = \mathrm{P}(X = 6,\ Y = 6) = \mathrm{P}(X = 6) \cdot \mathrm{P}(Y = 6) = \tfrac{1}{6} \cdot \tfrac{1}{6} = 1/36.$$
- Пусть случайная величина $X$ — это результат броска кубика, а случайная величина $Y$ — результат *того же броска того же кубика*.
  - $X$ и $Y$ всегда принимают равные значения. Вероятность того, что сумма бросков равна 12, равна
$$\mathrm{P}(X + Y = 12) = \mathrm{P}(X = 6) = \tfrac{1}{6}.$$

# Распределение суммы независимых случайных величин

Если случайные величины независимы, то можно вычислить распределение их суммы, зная их распределения по отдельности.

| $X$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | $1/2$ | $1/3$ | $1/4$ | $1/12$ |

| $Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | $1/5$ | $1/10$ | $1/2$ | $1/5$ |

# Распределение суммы независимых случайных величин

Если случайные величины независимы, то можно вычислить распределение их суммы, зная их распределения по отдельности.

| | | | | |
|---|---|---|---|---|
| $X$ | 1 | 2 | 3 | 4 |
| | ½ | ⅓ | ¼ | 1/12 |

| | | | | |
|---|---|---|---|---|
| $Y$ | 1 | 2 | 3 | 4 |
| | ⅕ | 1/10 | ½ | ⅕ |

| | $X$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $Y$ | | ½ | ⅓ | ¼ | 1/12 |
| 1 | ⅕ | 1/10 | 1/15 | 1/20 | 1/60 |
| 2 | 1/10 | 1/20 | 1/30 | 1/40 | 1/120 |
| 3 | ½ | 1/4 | 1/6 | 1/8 | 1/24 |
| 4 | ⅕ | 1/10 | 1/15 | 1/20 | 1/60 |

# Распределение суммы независимых случайных величин

Если случайные величины независимы, то можно вычислить распределение их суммы, зная их распределения по отдельности.

| $X$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | ½ | ⅓ | ¼ | 1/12 |

| $Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | ⅕ | 1/10 | ½ | ⅕ |

| | $X$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $Y$ | | ½ | ⅓ | ¼ | 1/12 |
| 1 | ⅕ | 1/10 | 1/15 | 1/20 | 1/60 |
| 2 | 1/10 | 1/20 | 1/30 | 1/40 | 1/120 |
| 3 | ½ | 1/4 | 1/6 | 1/8 | 1/24 |
| 4 | ⅕ | 1/10 | 1/15 | 1/20 | 1/60 |

$P(Z = 5) = P(X = 1, Y = 4) + P(X = 2, Y = 3) +$

$+ P(X = 3, Y = 2) + P(X = 4, Y = 1) =$

$= 1/10 + 1/6 + 1/40 + 1/60 = 37/120.$

# Распределение суммы независимых случайных величин

Если случайные величины независимы, то можно вычислить распределение их суммы, зная их распределения по отдельности.

| $X$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | ½ | ⅓ | ¼ | $1/12$ |

| $Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | ⅕ | $1/10$ | ½ | ⅕ |

| $Z$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| | $1/10$ | $7/60$ | $1/3$ | $37/120$ | $7/60$ | $11/120$ | $1/60$ |

# Резюме

- Независимость — отношение между случайными величинами, когда значение одной величины “не даёт информации” о значении другой
    - Отношение независимости симметрично
- Если случайные величины независимы, можно легко вычислить распределение их суммы

# Характеристики среднего

# Основные характеристики выборки

- Распределение — слишком сложная характеристика выборки
- Важными показателями распределения являются
  - среднее значение
  - “типичное” отклонение от среднего

# Математическое ожидание

**Определение.** Пусть случайная величина *X* принимает значения $x_1, \ldots, x_n$.
*Математическое ожидание* случайной величины *X* — это сумма

$$\mathsf{E}X = x_1 \cdot \mathsf{P}(X = x_1) + \ldots + x_n \cdot \mathsf{P}(X = x_n)$$

- Математическое ожидание — это среднее значение случайной величины

# Математическое ожидание

**Определение.** Пусть случайная величина *X* принимает значения $x_1, \ldots, x_n$
*Математическое ожидание* случайной величины *X* — это сумма

$$\mathsf{E}X = x_1 \cdot \mathsf{P}(X = x_1) + \ldots + x_n \cdot \mathsf{P}(X = x_n)$$

- Математическое ожидание — это среднее значение случайной величины
- Если у нас есть выборка данных $X_1, \ldots, X_N$ , то математическое ожидание соответствует среднему

# Свойства математического ожидания

- Можно выносить множитель: $\mathsf{E}(\mathrm{c} \cdot X) = \mathrm{c} \cdot \mathsf{E}X$
- Линейность: $\mathsf{E}(X + Y) = \mathsf{E}X + \mathsf{E}Y$
- Матожидание константы равно константе: $\mathsf{E}c = c$

**Задача 1**. Кидается 5 шестигранных кубиков. Найдите математическое ожидание суммы выпавших очков.

**Задача 1**. Кидается 5 шестигранных кубиков. Найдите математическое ожидание суммы выпавших очков.

**Решение.**

1. $X_1, X_2, X_3, X_4, X_5$ — случайные величины, равные результатам бросков 1, 2, 3, 4, 5 кубиков
   $X = X_1 + X_2 + X_3 + X_4 + X_5$

$$\mathsf{E}X = \mathsf{E}(X_1 + X_2 + X_3 + X_4 + X_5) =$$
$$\mathsf{E}X_1 + \mathsf{E}X_2 + \mathsf{E}X_3 + \mathsf{E}X_4 + \mathsf{E}X_5 = 5\mathsf{E}X_1$$

**Задача 1**. Кидается 5 шестигранных кубиков. Найдите математическое ожидание суммы выпавших очков.

**Решение.**

1. $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ — случайные величины, равные результатам бросков 1, 2, 3, 4, 5 кубиков

   $X = X_1 + X_2 + X_3 + X_4 + X_5$

$$\mathsf{E}X = \mathsf{E}(X_1 + X_2 + X_3 + X_4 + X_5) =$$
$$\mathsf{E}X_1 + \mathsf{E}X_2 + \mathsf{E}X_3 + \mathsf{E}X_4 + \mathsf{E}X_5 = 5\mathsf{E}X_1$$

2. Считаем математическое ожидание

$$\mathsf{E}X_1 = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \ldots + \frac{1}{6} \cdot 6 = 3.5$$

**Задача 1**. Кидается 5 шестигранных кубиков. Найдите математическое ожидание суммы выпавших очков.

**Решение.**

1. $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ — случайные величины, равные результатам бросков 1, 2, 3, 4, 5 кубиков
   $X = X_1 + X_2 + X_3 + X_4 + X_5$

$$\mathsf{E}X = \mathsf{E}(X_1 + X_2 + X_3 + X_4 + X_5) =$$
$$\mathsf{E}X_1 + \mathsf{E}X_2 + \mathsf{E}X_3 + \mathsf{E}X_4 + \mathsf{E}X_5 = 5\mathsf{E}X_1$$

2. Считаем математическое ожидание

$$\mathsf{E}X_1 = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \ldots + \frac{1}{6} \cdot 6 = 3.5$$

3. Ответ: $\mathsf{E}X = 5 \cdot \mathsf{E}X_1 = 17.5$

# Выборочное среднее

Как найти математическое ожидание случайной величины по выборке?

# Выборочное среднее

Как найти математическое ожидание случайной величины по выборке?

**Определение.** Пусть дана выборка $X = \{X_1, X_2, \ldots, X_n\}$ Выборочным средним выборки *X* называется величина

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

# Выборочное среднее

Как найти математическое ожидание случайной величины по выборке?

**Определение.** Пусть дана выборка $X = \{X_1, X_2, \ldots, X_n\}$. Выборочным средним выборки *X* называется величина

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- С ростом *n* выборочное среднее стремится к математическому ожиданию

# Проблема: выбросы в данных

- В данных могут быть выбросы, которые сместят среднее значение

# Проблема: выбросы в данных

- В данных могут быть выбросы, которые сместят среднее значение

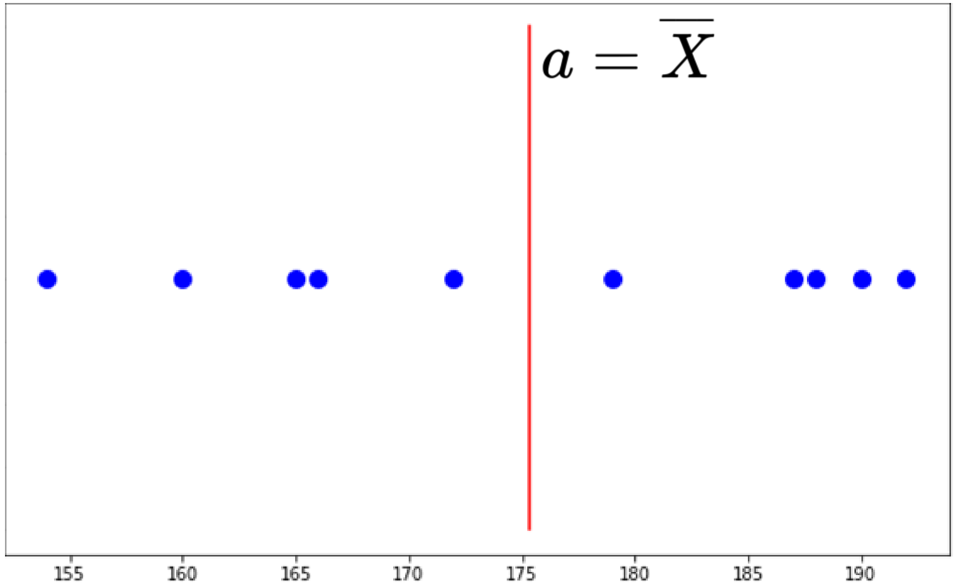
# Проблема: выбросы в данных

- В данных могут быть выбросы, которые сместят среднее значение
- Медиана — это число, которое делит упорядоченный ряд значений пополам
- Если в данных много выбросов, лучше считать медиану, а не среднее

# Характеристики разброса

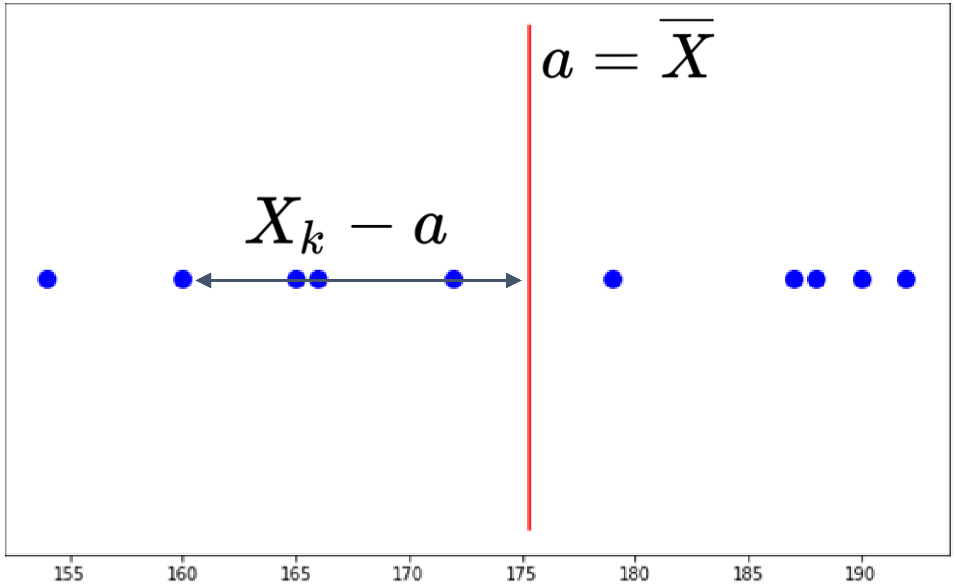
# Выборочная дисперсия

- Разброс определяет, как сильно элементы выборки отклоняются от среднего
- Пусть дана выборка $X = \{X_1, X_2, \ldots, X_n\}$
- Обозначим $a = \overline{X}$ — среднее значение выборки



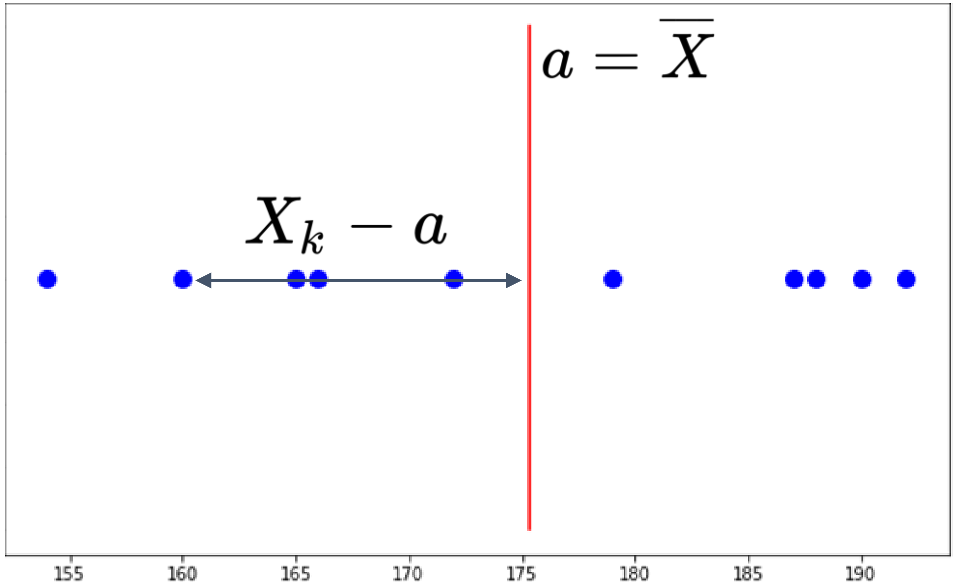
# Выборочная дисперсия

- Разброс определяет, как сильно элементы выборки отклоняются от среднего
- Пусть дана выборка $X = \{X_1, X_2, \ldots, X_n\}$
- Обозначим $a = \overline{X}$ — среднее значение выборки
- $X_k - a$ — отклонение *k*-ого объекта от среднего



# Выборочная дисперсия

- Разброс определяет, как сильно элементы выборки отклоняются от среднего
- Пусть дана выборка $X = \{X_1, X_2, \ldots, X_n\}$
- Обозначим $a = \overline{X}$ — среднее значение выборки
- $X_k - a$ — отклонение *k*-ого объекта от среднего
- $\overline{(X-a)^2} = \overline{(X-\overline{X})^2}$ — выборочная дисперсия выборки (обозначается $\hat{\mathrm{D}} X$ )




$$\hat{\mathrm{D}} X = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n}$$

# Стандартное отклонение

- Стандартное отклонение (std) — это корень из выборочной дисперсии:

$$\sigma(X) = \sqrt{\hat{\mathrm{D}}X} = \sqrt{\frac{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}}$$

- Стандартное отклонение имеет ту же размерность, что и дисперсия

# Пример: расчёт стандартного отклонения

- Выборка:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Петя | Маша | Саша | Витя | Рома | Настя | Коля | Миша | Катя | Наташа |
| 166 | 187 | 172 | 192 | 188 | 165 | 154 | 179 | 190 | 160 |

# Пример: расчёт стандартного отклонения

- Выборка:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Петя | Маша | Саша | Витя | Рома | Настя | Коля | Миша | Катя | Наташа |
| 166 | 187 | 172 | 192 | 188 | 165 | 154 | 179 | 190 | 160 |

- Средний рост равен 175.3 см

# Пример: расчёт стандартного отклонения

- Выборка:

| Петя | Маша | Саша | Витя | Рома | Настя | Коля | Миша | Катя | Наташа |
|---|---|---|---|---|---|---|---|---|---|
| 166 | 187 | 172 | 192 | 188 | 165 | 154 | 179 | 190 | 160 |

- Средний рост равен 175.3 см
- Отклонения:

| Петя | Маша | Саша | Витя | Рома | Настя | Коля | Миша | Катя | Наташа |
|---|---|---|---|---|---|---|---|---|---|
| -9,3 | 11,7 | -3,3 | 16,7 | 12,7 | -10,3 | -21,3 | 3,7 | 14,7 | -15,3 |

# Пример: расчёт стандартного отклонения

- Выборка:

| Петя | Маша | Саша | Витя | Рома | Настя | Коля | Миша | Катя | Наташа |
|---|---|---|---|---|---|---|---|---|---|
| 166 | 187 | 172 | 192 | 188 | 165 | 154 | 179 | 190 | 160 |

- Средний рост равен 175.3 см
- Отклонения:

| Петя | Маша | Саша | Витя | Рома | Настя | Коля | Миша | Катя | Наташа |
|---|---|---|---|---|---|---|---|---|---|
| -9,3 | 11,7 | -3,3 | 16,7 | 12,7 | -10,3 | -21,3 | 3,7 | 14,7 | -15,3 |

- Выборочная дисперсия:

$$D = \frac{(-9.3)^2+11.7^2+(-3.3)^2+16.7^2+12.7^2+(-10.3)^2+(-21.3)^2+3.7^2+14.7^2+(-15.3)^2}{10} = \frac{1698.1}{10} = 169,81$$

# Пример: расчёт стандартного отклонения

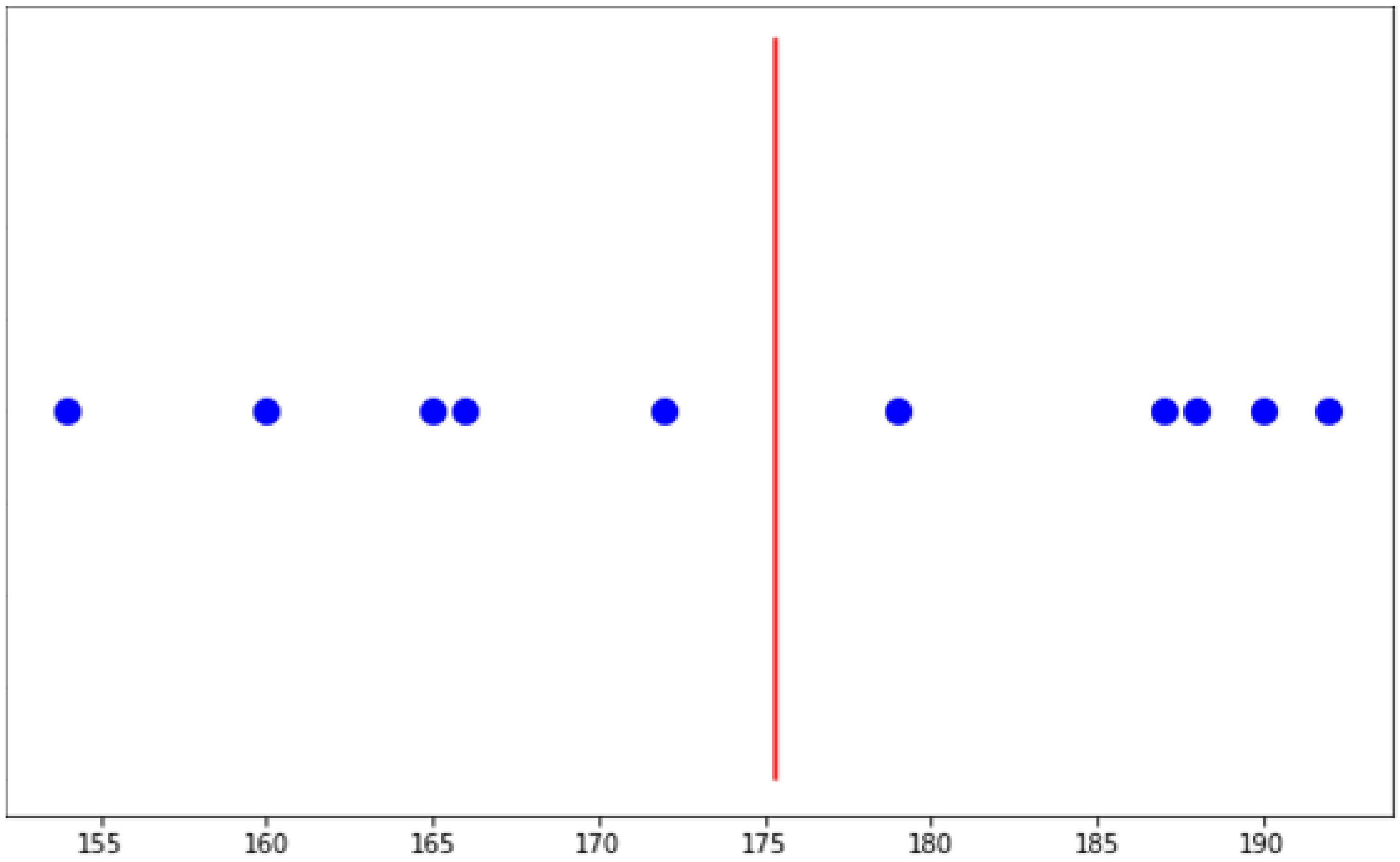
- Выборочная дисперсия:

$$D = \frac{(-9.3)^2+11.7^2+(-3.3)^2+16.7^2+12.7^2+(-10.3)^2+(-21.3)^2+3.7^2+14.7^2+(-15.3)^2}{10} = \frac{1698.1}{10} = 169,81$$

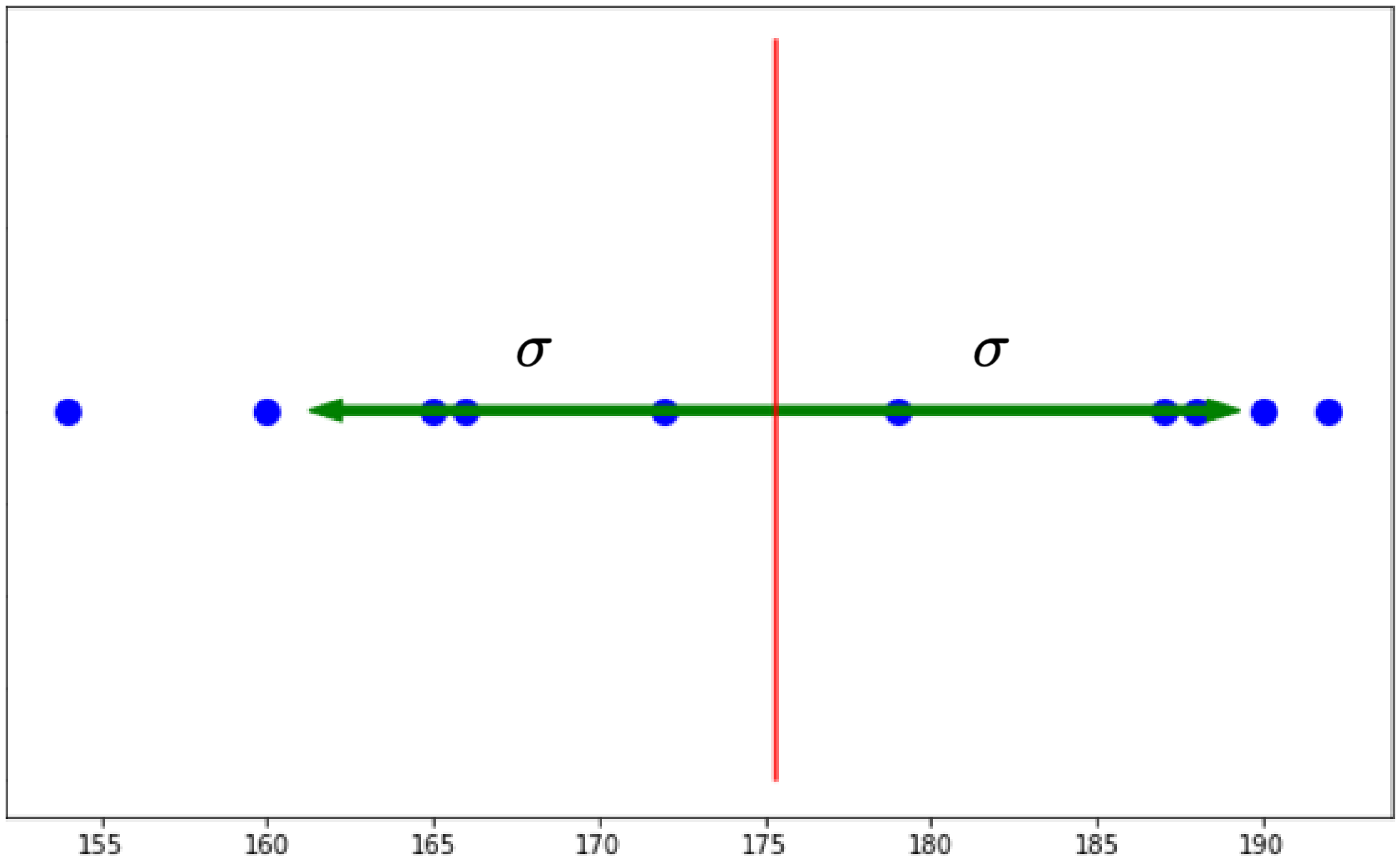
- Стандартное отклонение:

$$\sigma = \sqrt{D} \approx 13,03$$

# Визуализация стандартного отклонения



# Визуализация стандартного отклонения



# Дисперсия случайной величины

- Пусть $X$ — случайная величина, $a = \mathsf{E}X$
- $X - a$ — отклонение случайной величины от среднего

**Определение**. Дисперсией случайной величины $X$ называется число

$$\mathsf{E}[(X - \mathsf{E}X)^2]$$

# Дисперсия случайной величины

- Пусть *X* — случайная величина, $a = \mathsf{E}X$
- $X - a$ — отклонение случайной величины от среднего

**Определение**. Дисперсией случайной величины *X* называется число

$$\mathsf{E}[(X - \mathsf{E}X)^2]$$

Дисперсия аналогична среднему квадратичному отклонению выборки

# Дисперсия случайной величины

- Пусть $X$ — случайная величина, $a = \mathsf{E}X$
- $X - a$ — отклонение случайной величины от среднего

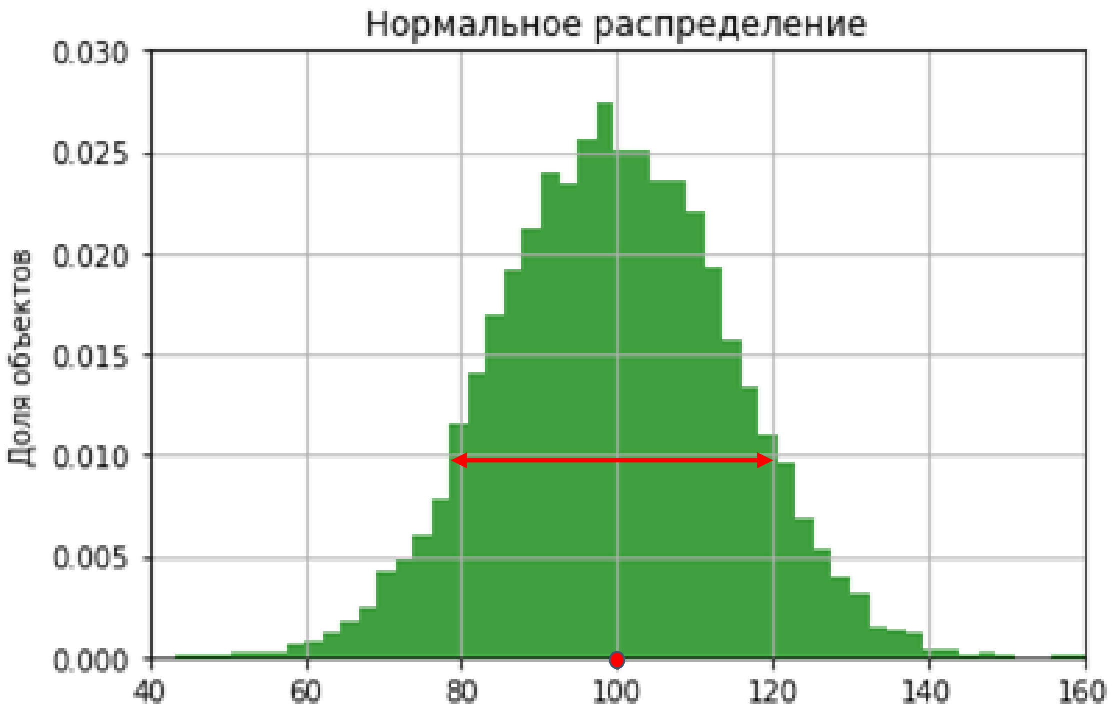
**Определение**. Дисперсией случайной величины $X$ называется число

$$\mathsf{E}[(X - \mathsf{E}X)^2]$$

Дисперсия аналогична среднему квадратичному отклонению выборки

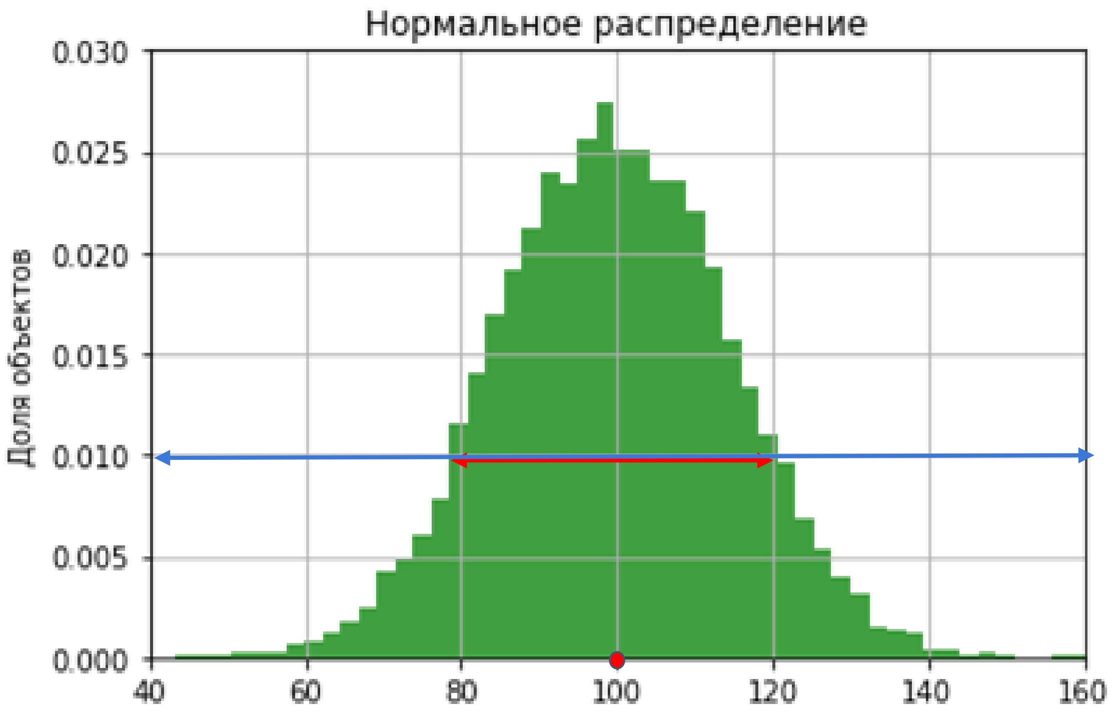
# Визуализация стандартного отклонения

- Рассмотрим нормальное распределение
- Диапазон [*a* - 3*σ*, *a* - 3*σ*] содержит 99,73% всех объектов



# Визуализация стандартного отклонения

- Рассмотрим нормальное распределение
- Диапазон [$a$ - 3$\sigma$, $a$ - 3$\sigma$] содержит 99,73% всех объектов



# Свойства дисперсии

- $DX$ неотрицательна

# Свойства дисперсии

- DX неотрицательна
- Множитель выносится с квадратом

# Свойства дисперсии

- DX неотрицательна
- Множитель выносится с квадратом
- $DX = E(X^2) - (EX)^2$

Доказательство…

# Свойства дисперсии

- DX неотрицательна
- Множитель выносится с квадратом
- $DX = E(X^2) - (EX)^2$
- Для независимых случайных величин дисперсия суммы равна сумме дисперсий

# Резюме

- Математическое ожидание — среднее значение случайной величины
  - Его можно приближённо вычислить как среднее арифметическое выборки
  - Чтобы снизить значение выбросов, используют медиану
- Дисперсия и среднее квадратичное отклонение — показатели разброса
  - Необходимы для оценки отклонения значения случайной величины от среднего

# Интуитивный смысл корреляции

- “Уровень образования коррелирует со снижением уровня преступности”
- “Курс рубля коррелирует с ценой на нефть”

**Определение** (интуитивное). Величины коррелируют, если чем выше одна величина, тем, скорее всего, будет выше другая величина.

# Ковариация

**Определение**. Пусть $X$ и $Y$ — две случайные величины с математическими ожиданиями $a$ и $b$. Их **ковариацией** называется величина

$$\mathrm{cov}(X, Y) = \mathrm{E}[(X - a) \cdot (Y - b)].$$

Ковариация — это матожидание случайной величины

$$Z = (X - a) \cdot (Y - b).$$

# Пример вычисления ковариации

$X$ — рост случайного человека

$Y$ — вес того же случайного человека

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 167 | 155 | 182 | 196 | 180 | 185 | 177 | 173 | 165 | 181 |
| $Y$ | 55 | 60 | 83 | 95 | 105 | 88 | 76 | 65 | 60 | 80 |

Средние значения:

$$a = \frac{167 + 155 + 182 + 196 + 180 + 185 + 177 + 173 + 165 + 181}{10} = 176.1$$

$$b = \frac{55 + 60 + 83 + 95 + 105 + 88 + 76 + 65 + 60 + 80}{10} = 76.7.$$

# Пример вычисления ковариации

Таблицы для случайных величин $X$ - $a$ и $Y$ - $b$:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ - $a$ | -9.1 | -21.1 | 5.9 | 19.1 | 3.9 | 8.9 | 0.9 | -3.1 | -11.1 | 4.9 |
| $Y$ - $b$ | -21.7 | -16.7 | 6.3 | 18.3 | 28.3 | 11.3 | -0.7 | -11.7 | -16.7 | 3.3 |

# Пример вычисления ковариации

Таблицы для случайных величин *X - a* и *Y - b*:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ - $a$ | -9.1 | -21.1 | 5.9 | 19.1 | 3.9 | 8.9 | 0.9 | -3.1 | -11.1 | 4.9 |
| $Y$ - $b$ | -21.7 | -16.7 | 6.3 | 18.3 | 28.3 | 11.3 | -0.7 | -11.7 | -16.7 | 3.3 |

- (*X - a*) больше нуля в том случае, если значение случайной величины *X* выше своего среднего.
- Аналогично с выражением (*Y - b*).
- Поскольку *X* и *Y* коррелируют, то, как правило, выражения *X - a* и *Y - b* имеют один и тот же знак.

# Пример вычисления ковариации

Таблицы для случайных величин $X - a$, $Y - b$, $Z = (X - a)(Y - b)$:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ - $a$ | -9.1 | -21.1 | 5.9 | 19.1 | 3.9 | 8.9 | 0.9 | -3.1 | -11.1 | 4.9 |
| $Y$ - $b$ | -21.7 | -16.7 | 6.3 | 18.3 | 28.3 | 11.3 | -0.7 | -11.7 | -16.7 | 3.3 |
| $Z$ | 197.47 | 352.37 | 37.17 | 349.53 | 110.37 | 100.57 | -0.63 | 36.27 | 185.37 | 16.17 |

E$Z$ = cov($X$, $Y$)

- велико, когда $X$ и $Y$ коррелируют
- мало, когда $X$ и $Y$ отрицательно коррелируют

# Ковариация независимых случайных величин

**Утверждение.** Если $X$ и $Y$ — независимые случайные величины, то $\operatorname{cov}(X, Y) = 0$.

- Обратное утверждение неверно: можно придумать две случайные величины $X$ и $Y$, для которых $\operatorname{cov}(X, Y) = 0$, но независимыми они не являются
- Условие независимости гораздо более мощное и описывает все совместное распределение случайных величин.

# Корреляция

**Определение**. Корреляцией случайных величин *X*, *Y* называется число

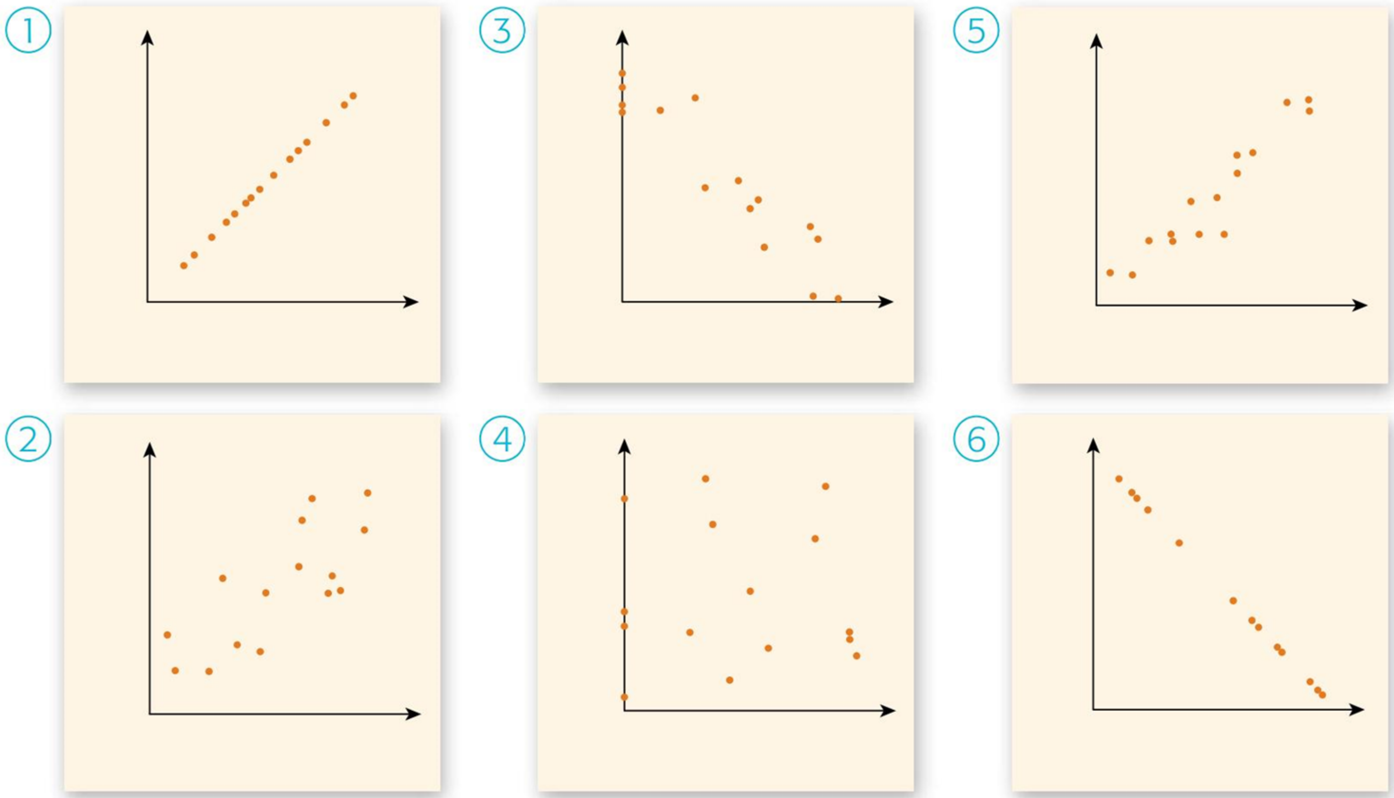
$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{D}X\mathrm{D}Y}}$$

# Свойства корреляции

- Корреляция всегда находится в промежутке [-1, 1].
- Если corr($X$, $Y$) = 1, то $X$ и $Y$ линейно зависимы с положительным коэффициентом (например, как градусы Цельсия и Фаренгейта)
- Если corr($X$, $Y$) близка к 1, то $X$ и $Y$ высоко скоррелированы
- Если corr($X$, $Y$) = 0, то $X$ и $Y$ не скоррелированы (но могут быть зависимы!).
- Если corr($X$, $Y$) близка к -1, то $X$ и $Y$ отрицательно скоррелированы.
- Если corr($X$, $Y$) равна -1, то $X$ и $Y$ линейно зависимы с отрицательным коэффициентом (например, $Y = -2X + 5$).

# Визуальное определение корреляции

**Упражнение**. Оцените коэффициенты корреляции выборок по данным рисункам



# Корреляция в анализе данных

**Вопрос**. Какие переменные будут наиболее важны при попытке предсказать целевую переменную Y для решения задачи кредитного скоринга?

- Необходимо посчитать величины corr($X_i$, $Y$) для всех i.
- Высокие (по модулю) значения корреляции будут указывать на то, что признак нам точно важен.
- Значение корреляции, близкое к 0, вообще говоря не означает, что признак нам точно будет не важен, но позволяют сделать такое предположение.
    - Такие признаки можно выкинуть, чтобы не усложнять себе задачу предсказания.

# Матрица корреляций

- Пусть у каждого объекта в выборке нам известно *n* признаков (у нас *n* = 6)
- Построим таблицу *n* x *n*, в клетку с координатами (*m*, *k*) которой запишем значение корреляции признаков $X_m$ и $X_k$.

Такая таблица называется *матрицей корреляций*.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| месячная зарплата ($X_1$) | 1 | 0.9 | -0.2 | 0.1 | -0.3 | 0.7 |
| средние месячные траты ($X_2$) | 0.9 | 1 | -0.15 | 0.25 | -0.35 | 0.65 |
| количество текущих кредитов ($X_3$) | -0.2 | -0.15 | 1 | 0.6 | 0.8 | -0.15 |
| количество погашенных кредитов ($X_4$) | 0.1 | 0.25 | 0.6 | 1 | 0.3 | -0.25 |
| кол-во просрочек по выплатам ($X_5$) | -0.3 | -0.35 | 0.8 | 0.3 | 1 | -0.4 |
| наличие собственного жилья ($X_6$) | 0.7 | 0.65 | -0.15 | -0.25 | -0.4 | 1 |

# Свойства матрицы корреляций

- Матрица симметричная (следует из симметричности корреляции).
- На главной диагонали стоят единицы (см. задачу из предыдущего параграфа).
- Число на пересечении первой строки и последнего столбца равно 0.7.
    - Это значит, что месячная зарплата высоко коррелирует с наличием собственного жилья.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| месячная зарплата ($X_1$) | 1 | 0.9 | -0.2 | 0.1 | -0.3 | 0.7 |
| средние месячные траты ($X_2$) | 0.9 | 1 | -0.15 | 0.25 | -0.35 | 0.65 |
| количество текущих кредитов ($X_3$) | -0.2 | -0.15 | 1 | 0.6 | 0.8 | -0.15 |
| количество погашенных кредитов ($X_4$) | 0.1 | 0.25 | 0.6 | 1 | 0.3 | -0.25 |
| кол-во просрочек по выплатам ($X_5$) | -0.3 | -0.35 | 0.8 | 0.3 | 1 | -0.4 |
| наличие собственного жилья ($X_6$) | 0.7 | 0.65 | -0.15 | -0.25 | -0.4 | 1 |

# Свойства матрицы корреляций

- $X_1$ и $X_2$ коррелируют с коэффициентом 0.9
  - Следовательно, второй признак почти не дает дополнительной информации по сравнению с первым. Его иногда целесообразно выкидывать из выборки.

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| месячная зарплата ($X_1$) | 1 | 0.9 | -0.2 | 0.1 | -0.3 | 0.7 |
| средние месячные траты ($X_2$) | 0.9 | 1 | -0.15 | 0.25 | -0.35 | 0.65 |
| количество текущих кредитов ($X_3$) | -0.2 | -0.15 | 1 | 0.6 | 0.8 | -0.15 |
| количество погашенных кредитов ($X_4$) | 0.1 | 0.25 | 0.6 | 1 | 0.3 | -0.25 |
| кол-во просрочек по выплатам ($X_5$) | -0.3 | -0.35 | 0.8 | 0.3 | 1 | -0.4 |
| наличие собственного жилья ($X_6$) | 0.7 | 0.65 | -0.15 | -0.25 | -0.4 | 1 |

IT-ОБРАЗОВАНИЕ
В ПЕТЕРБУРГЕ
И УДАЛЕННО

# Спасибо за внимание!