

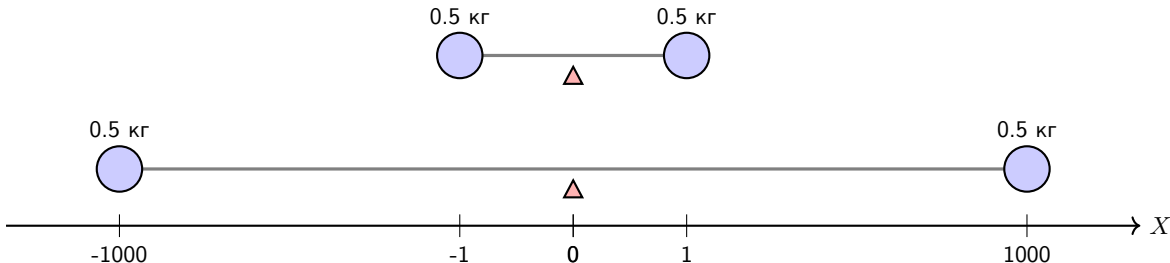
Теория вероятностей и математическая статистика

Дисперсия. Функция распределения. Бернулли, биномиальное, Пуассон.

Глеб Карпов

Программы БИ, УЦП, ТАДвБ ВШБ ВШЭ.

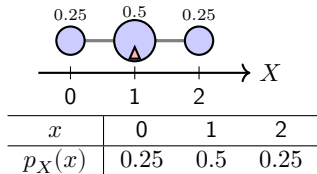
Физическое Лирическое отступление 2



- В вероятностном мире интерпретация картинки такая, что может быть две случайные величины с одинаковым математическим ожиданием, но с кардинально разными значениями.
- Представьте, что в одном случае вы кидаете монетку с привязанными численными значениями $+1$ и -1 (например, выигрыш - проигрыш), а в другом ту же монетку, но с привязанными значениями уже $+1000$ и -1000 . Математическое ожидание в обоих случаях будет 0, но сравните магнитуды возможных значений!
- Это помогает нам понять мотивацию введения новой характеристики для случайной величины - **дисперсии** - которая сможет охарактеризовать степень разброса значений случайной величины.

Примеры для интуиции

Система А: $\Omega_X = \{0, 1, 2\}$.



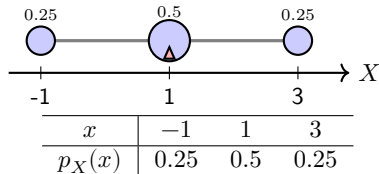
$$E[X] = 0 \cdot 0.25 + 1 \cdot 0.5 + 2 \cdot 0.25 = 1$$

$$Y = f(X) = (X - E[X])^2 = (X - 1)^2:$$

y	0	1
$p_Y(y)$	0.5	0.5

$$E[Y] = 0 \cdot 0.5 + 1 \cdot 0.5 = 0.5$$

Система В: $\Omega_X = \{-1, 1, 3\}$.



$$E[X] = (-1) \cdot 0.25 + 1 \cdot 0.5 + 3 \cdot 0.25 = 1$$

$$Y = f(X) = (X - E[X])^2 = (X - 1)^2:$$

y	0	4
$p_Y(y)$	0.5	0.5

$$E[Y] = 0 \cdot 0.5 + 4 \cdot 0.5 = 2$$

Характеристики случайной величины

Дисперсия

- Дисперсия случайной величины - это число, **константа**, которое помогает нам понять степень разброса потенциальных значений случайной величины относительно среднего.

Definition

Для случайной величины X её дисперсия обозначается $Var[X]$ (Variance) и определяется как:

$$D[X] = Var[X] = E[(X - E[X])^2]$$

- Формально говоря, это математическое ожидание новой случайной величины:

$$Y = g(X) = (X - E[X])^2,$$

описывающей квадрат расстояния между значением случайной величины X и её математическим ожиданием.

Характеристики С.В.: Дисперсия

Вычисление дисперсии

Существует упрощенная формула:

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2XE[X] + (E[X])^2] = \\ &= E[X^2] - E[2E[X]X] + E[(E[X])^2] = E[X^2] - 2E[X]E[X] + (E[X])^2 = \\ &\quad \boxed{E[X^2] - (E[X])^2} \end{aligned}$$

В процессе мы используем свойство линейности математического ожидания, а также знание того, что $E[X]$ - это константа, при этом $E[c] = c$, то есть $E[E[X]] = E[X]$.

Свойства дисперсии

1. **Неотрицательность:** Для любой случайной величины X :

$$\text{Var}[X] \geq 0$$

2. **Дисперсия константы:** Если c - константа, то:

$$\text{Var}[c] = 0$$

3. **Масштабирование:** Для любой константы c :

$$\text{Var}[cX] = c^2 \text{Var}[X]$$

Дисперсия: пример

Посчитаем дисперсию на примере нашей задачи. Считаем уже известным, что $E[X] = 1.2$.

x	0	-1	-2	1	2	3
$p_X(x)$	0.1	0.1	0.1	0.2	0.2	0.3

- Нам нужно вычислить $E[Y]$, где $Y = (X - E[X])^2$. Сразу воспользуемся подходом, позволяющим миновать шаг построения таблицы вероятности для Y :

$$\begin{aligned} E[Y] &= \sum_{i=1}^6 g(x_i) P_X(X = x_i) = \sum_{i=1}^6 (x_i - 1.2)^2 P_X(X = x_i) = \\ &= 1.44 \cdot 0.1 + 4.84 \cdot 0.1 + 10.24 \cdot 0.1 + 0.04 \cdot 0.2 + 0.64 \cdot 0.2 + 3.24 \cdot 0.3 = 2.76 \end{aligned}$$

- Но для дисперсии у нас есть ещё более умный подход, описанный на одном из слайдов выше:

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 = 4.2 - (1.2)^2 = 2.76 \\ E[X^2] &= \sum_{i=1}^6 (x_i)^2 P_X(X = x_i) = 0.1 + 4 \cdot 0.1 + 1 \cdot 0.2 + 4 \cdot 0.2 + 9 \cdot 0.3 = 4.2 \end{aligned}$$

- Voilà!

Функция распределения (кумулятивная/интегральная)

i Definition

Функция распределения (кумулятивная/интегральная) (CDF) случайной величины X определяется как:

$$F_X(x) = P(X \leq x), \quad F_X(x) : \mathbb{R} \rightarrow [0, 1] \subset \mathbb{R}$$

Для дискретной случайной величины:

$$F_X(x) = \sum_{i: x_i \leq x} P_X(X = x_i)$$

- $F_X(x)$ показывает вероятность того, что случайная величина X примет значение, не превышающее x . Иными словами, сколько вероятности “скопилось”, аккумулировалось, до значения x .

Свойства CDF для дискретных случайных величин

1. **Монотонность:** $F_X(x)$ неубывающая функция

2. **Граничные значения:**

- $\lim_{x \rightarrow -\infty} F_X(x) = 0$

- $\lim_{x \rightarrow +\infty} F_X(x) = 1$

3. **Ступенчатая функция:** CDF дискретной с.в. имеет скачки в точках, где $p_X(x) > 0$

Пример 1: CDF для распределения X

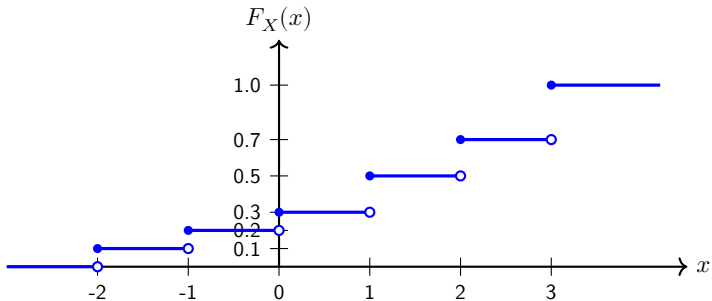
Вспомним наше первое распределение:

x	0	-1	-2	1	2	3
$p_X(x)$	0.1	0.1	0.1	0.2	0.2	0.3

Вычислим CDF:

1. $F_X(x) = 0$ при $x < -2$
2. $F_X(x) = 0.1$ при $-2 \leq x < -1$
3. $F_X(x) = 0.2$ при $-1 \leq x < 0$
4. $F_X(x) = 0.3$ при $0 \leq x < 1$
5. $F_X(x) = 0.5$ при $1 \leq x < 2$
6. $F_X(x) = 0.7$ при $2 \leq x < 3$
7. $F_X(x) = 1.0$ при $x \geq 3$

График CDF для распределения X



Использование CDF для вычисления вероятностей

CDF позволяет легко вычислять вероятности попадания в интервалы:

💡 Формулы для вычисления вероятностей

1. $P(X \leq a) = F_X(a)$
2. $P(X < a) = F_X(a^-) = \lim_{x \rightarrow a^-} F_X(x)$
3. $P(X > a) = 1 - F_X(a)$
4. $P(X \geq a) = 1 - F_X(a^-)$
5. $P(a < X \leq b) = F_X(b) - F_X(a)$
6. $P(a \leq X \leq b) = F_X(b) - F_X(a^-)$

Распределение Бернулли

Начнем с простейшего случая - **одного случайного эксперимента** с двумя возможными исходами.

i Definition

Распределение Бернулли - это распределение случайной величины X , которая принимает только два значения:

$$X = \begin{cases} 1 & \text{с вероятностью } p \text{ ("успех")} \\ 0 & \text{с вероятностью } q = 1 - p \text{ ("неудача")} \end{cases}$$

Обозначается как $X \sim \text{Bern}(p)$, где $p \in [0, 1]$ - параметр распределения.

Характеристики распределения Бернулли

💡 Основные характеристики

Для случайной величины $X \sim \text{Bern}(p)$:

- Математическое ожидание:

$$E[X] = 0 \cdot (1 - p) + 1 \cdot p = p$$

- Дисперсия:

$$\text{Var}[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$$

- Стандартное отклонение:

$$\sigma = \sqrt{p(1 - p)}$$

Примеры распределения Бернулли

1. Подбрасывание монеты: "Орел" = 1, "Решка" = 0, $p = 0.5$ (для честной монеты)
2. Производственный контроль: "Дефектное изделие" = 1, "Качественное" = 0
3. Медицинский тест: "Положительный результат" = 1, "Отрицательный" = 0
4. Маркетинг: "Покупка товара" = 1, "Отказ от покупки" = 0

От одного испытания к процессу

- Что если мы хотим провести несколько независимых испытаний Бернулли?
- Например: подбросить монету 10 раз, проверить 100 изделий на дефекты, провести 50 медицинских тестов.
- Это приводит нас к понятию процесса Бернулли - последовательности независимых испытаний Бернулли.

Процесс Бернулли

i Definition

Процесс Бернулли - это последовательность независимых одинаково распределенных случайных величин $X_1, X_2, \dots, X_n$, где каждая $X_i \sim \text{Bern}(p)$:

$$X_i = \begin{cases} 1 & \text{с вероятностью } p \text{ ("успех")} \\ 0 & \text{с вероятностью } 1 - p \text{ ("неудача")} \end{cases}$$

Ключевое отличие от одного испытания: теперь у нас есть **последовательность** результатов.

Свойства процесса Бернулли

1. **Независимость:** Результат каждого испытания X_i не зависит от результатов других испытаний
2. **Два исхода:** Каждое испытание может закончиться только успехом (1) или неудачей (0)
3. **Постоянная вероятность:** Вероятность успеха p не изменяется от испытания к испытанию

Примеры процесса Бернулли

1. **Серия подбрасываний монеты:** Последовательность $(X_1, X_2, \dots, X_{10})$ для 10 подбрасываний
2. **Контроль качества:** Проверка партии из n изделий, где каждое может быть дефектным независимо от других
3. **Клинические испытания:** Тестирование лекарства на n пациентах, где каждый может показать положительную реакцию

Биномиальное распределение

- Теперь, когда у нас есть процесс Бернулли, возникает естественный вопрос: **“Сколько успехов мы получим в n испытаниях?”**

i Definition

Биномиальное распределение описывает количество успехов в n независимых испытаниях Бернулли с вероятностью успеха p . Если $X_1, X_2, \dots, X_n$ - процесс Бернулли с параметром p , то случайная величина:

$$Y = X_1 + X_2 + \dots + X_n$$

имеет биномиальное распределение и обозначается $Y \sim \text{Bin}(n, p)$. Y может принимать значения $0, 1, 2, \dots, n$ (от 0 до n успехов).

Функция вероятности биномиального распределения

Вероятность получить ровно k успехов в n испытаниях:

$$P(Y = k) = C_n^k p^k (1 - p)^{n-k}$$

Интуиция формулы

- p^k - вероятность k успехов
- $(1 - p)^{n-k}$ - вероятность $(n - k)$ неудач
- C_n^k - количество способов выбрать k позиций из n для размещения успехов

Необходимость подсчета комбинаций

Пример: $n = 7$ испытаний, ровно $k = 2$ успеха. Зелёный круг — успех, белый — неудача. Несколько возможных серий:



- Для каждой уникальной серии испытаний есть уникальный набор **координат** успехов.
- Таких наборов координат много: (1, 2), (1, 3), ..., (6, 7).
- Чтобы посчитать вероятность ровно 2 успехов, нужно знать, сколько таких наборов координат возможно. Иными словами, сколько множеств мощности 2 можно извлечь из множества мощности 7.
- Это и есть число комбинаций:

$$C_7^2 = \binom{7}{2} = \frac{7!}{2!5!} = 21.$$

Пусть $p = 0.2$, $Y \sim \text{Bin}(7, 0.2)$. Тогда

$$P(Y = 2) = C_7^2 p^2 (1 - p)^5 = 21 \cdot 0.2^2 \cdot 0.8^5 = 21 \cdot 0.04 \cdot 0.32768 \approx 0.275,$$

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= C_7^0 0.8^7 + C_7^1 0.2 \cdot 0.8^6 + C_7^2 p^2 (1 - p)^5 \approx 0.852. \end{aligned}$$

Биномиальное распределение

Основные характеристики

Для случайной величины $Y \sim \text{Bin}(n, p)$:

Математическое ожидание:

$$E[Y] = np$$

Дисперсия:

$$\text{Var}[Y] = np(1 - p)$$

Стандартное отклонение:

$$\sigma = \sqrt{np(1 - p)}$$

Наивероятнейшее число успехов k_0 (мода распределения) — целое число, для которого $P(Y = k_0)$ максимальна. При $q = 1 - p$ оно удовлетворяет

$$np - q \leq k_0 \leq np + p.$$

Эквивалентная запись: $(n + 1)p - 1 \leq k_0 \leq (n + 1)p$. Если $(n + 1)p$ целое, наивероятнейших значений два: $k_0 = (n + 1)p - 1$ и $k_0 = (n + 1)p$.

Визуализация биномиального распределения

Форма биномиального распределения зависит от параметров n и p . Рассмотрим, как влияет значение p на распределение при фиксированном $n = 20$:

Сравнение различных значений p

Функция вероятности биномиального распределения
 $n = 20$ испытаний

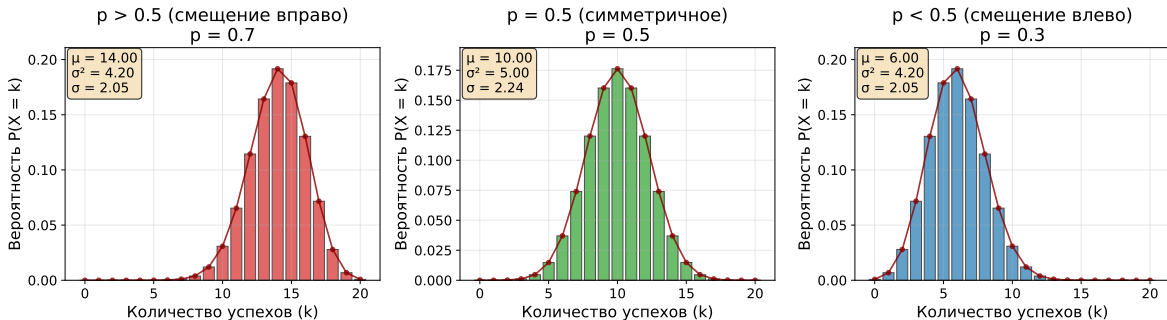


Рис. 1: Сравнение биномиального распределения для разных значений p

Визуализация биномиального распределения

i Наблюдения

$p < 0.5$ (например, $p = 0.3$): Распределение смещено влево, большинство значений сосредоточено в области малых k

$p = 0.5$: Распределение симметрично относительно среднего значения

$p > 0.5$ (например, $p = 0.7$): Распределение смещено вправо, большинство значений сосредоточено в области больших k

Пример: подбрасывание монеты

Подбросим честную монету 10 раз. Какова вероятность получить ровно 6 орлов?

- $n = 10, p = 0.5, k = 6, Y \sim \text{Bin}(10, 0.5)$

$$P(Y = 6) = C_{10}^6 (0.5)^6 (0.5)^4 = \frac{10!}{6! \cdot 4!} (0.5)^{10}$$

$$= \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} \cdot \frac{1}{1024} = 210 \cdot \frac{1}{1024} \approx 0.205$$

Пример: контроль качества

На заводе производят детали, 3% из которых бракованные. Проверяют партию из 100 деталей.

- $n = 100$, $p = 0.03$ (вероятность брака)
- $Y \sim \text{Bin}(100, 0.03)$ - количество бракованных деталей

Характеристики:

- $E[Y] = 100 \cdot 0.03 = 3$ (ожидаем 3 бракованные детали)
- $\text{Var}[Y] = 100 \cdot 0.03 \cdot 0.97 = 2.91$
- $\sigma = \sqrt{2.91} \approx 1.71$

Вероятность не найти ни одной бракованной детали:

$$P(Y = 0) = C_{100}^0 (0.03)^0 (0.97)^{100} = (0.97)^{100} \approx 0.048$$

Применения биномиального распределения

1. **Медицина:** Количество пациентов, для которых лекарство окажется эффективным
2. **Маркетинг:** Количество покупателей, совершивших покупку после просмотра рекламы
3. **Социология:** Количество респондентов, давших положительный ответ в опросе
4. **Спорт:** Количество побед команды в серии матчей

От биномиального к распределению Пуассона

- Биномиальное удобно, когда число испытаний n известно и невелико.
- На практике часто n огромно, а p мало: редкие события на большом потоке попыток.
- Примеры: число аварий за день, звонков в колл-центр за минуту, опечаток на странице.
- Тогда удобнее модель **Пуассона**: считаем среднее число событий λ за интервал (времени или участка), а не пару (n, p) .

Распределение Пуассона

i Definition

Распределение Пуассона описывает число наступлений редкого события за фиксированный промежуток времени (или на фиксированном участке). Случайная величина $X \sim Pois(\lambda)$ принимает значения $0, 1, 2, \dots$ с вероятностями

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

где $\lambda > 0$ — параметр: среднее число событий за выбранный интервал.

Интуиция формулы Пуассона

💡 Предельный переход от биномиального

Возьмём $Y \sim \text{Bin}(n, p)$ и устремим $n \rightarrow \infty$, $p \rightarrow 0$ так, что $np = \lambda$ (фиксируем среднее). Подставим $p = \lambda/n$:

$$\begin{aligned} P(Y = k) &= C_n^k p^k (1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \cdot \underbrace{\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n}}_{\rightarrow 1} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\rightarrow e^{-\lambda}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\rightarrow 1}. \end{aligned}$$

При $n \rightarrow \infty$ получаем $P(Y = k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}$.

Характеристики распределения Пуассона

Основные характеристики

Для случайной величины $X \sim Pois(\lambda)$:

Математическое ожидание:

$$E[X] = \lambda$$

Дисперсия:

$$Var[X] = \lambda$$

Стандартное отклонение:

$$\sigma = \sqrt{\lambda}$$

Когда применима модель Пуассона

Модель $X \sim Pois(\lambda)$ осмысленна, если одновременно выполняются условия:

1. **Счёт за фиксированный интервал.** Считаем число наступлений события на заданном отрезке времени / участке пространства.
2. **Редкость.** За малый кусок интервала вероятность двух и более событий пренебрежимо мала (события не «слипаются»).
3. **Постоянная интенсивность.** Среднее число событий пропорционально длине интервала: за время t параметр λt (однородность; нет пиков, если это не заложено в модель).
4. **Независимость на непересекающихся кусках.** Число событий на разных непересекающихся участках независимы.

Простой поток событий

Термин пришёл из смежной области — **теории массового обслуживания** (очереди, заявки, вызовы). Это очень прикладная история: здесь те же условия, при которых хорошо работает Пуассон, поэтому мы можем моделировать ситуации из реального мира.

Простой поток — модель потока однотипных событий во времени с тремя свойствами. Справа от стрелки — отсылка к соответствующему условию применимости Пуассона:

1. Вероятностные характеристики зависят только от длины интервала, не от того, «где» он лежит на оси времени.
↔ постоянная интенсивность (п. 3 выше).
2. За малый промежуток почти наверняка не больше одного события.
↔ редкость (п. 2 выше).
3. Будущее не зависит от прошлого: число событий на непересекающихся интервалах независимы.
↔ независимость на непересекающихся кусках (п. 4 выше).

Следствие: если поток простой с интенсивностью λ , то число событий за время t имеет распределение $Pois(\lambda t)$.

Пример: звонки в колл-центр

Звонки поступают с интенсивностью $\lambda = 4$ звонка в минуту. Это пример простого потока событий: редкость на мелкой шкале, постоянная интенсивность, независимость на непересекающихся кусках времени.

1. Какова вероятность, что за минуту не будет ни одного звонка?
2. Какова вероятность ровно двух звонков за минуту?
3. Найдите $E[X]$ и σ для числа звонков за минуту.
4. Какова вероятность, что за полминуты будет не больше одного звонка?
5. Какова вероятность хотя бы одного звонка за 15 секунд?
6. Чему равно среднее число звонков за две минуты?

Пример: решение

За минуту $X \sim Pois(4)$.

1. $P(X = 0) = e^{-4} \approx 0.018$.

2. $P(X = 2) = \frac{4^2}{2!} e^{-4} = 8e^{-4} \approx 0.147$.

3. $E[X] = 4$, $\sigma = \sqrt{4} = 2$.

4. За полминуты $t = \frac{1}{2}$, поэтому $Y \sim Pois(2)$:

$$P(Y \leq 1) = e^{-2} + 2e^{-2} = 3e^{-2} \approx 0.406.$$

5. За 15 секунд $t = \frac{1}{4}$ минуты, $W \sim Pois(1)$:

$$P(W \geq 1) = 1 - P(W = 0) = 1 - e^{-1} \approx 0.632.$$

6. За две минуты $Z \sim Pois(8)$, значит $E[Z] = 8$.

Пример: брак на участке ленты

Брак на конвейере — пространственный аналог простого потока («время» заменено длиной ленты): в среднем $\lambda = 1.5$ бракованных деталей на заданный участок длины 1. Подходит, если брак не идёт «пачками» и интенсивность по длине примерно постоянна.

1. Какова вероятность, что на участке длины 1 будет не больше одного брака?
2. Какова вероятность, что участок длины 1 без брака?
3. Чему равно среднее число браков на участке вдвое длиннее?

Пример: решение

На участке длины 1 пусть $X \sim Pois(1.5)$.

1. $P(X \leq 1) = e^{-1.5} + 1.5 e^{-1.5} \approx 0.557$.

2. $P(X = 0) = e^{-1.5} \approx 0.223$.

3. На участке длины 2 параметр $1.5 \cdot 2 = 3$, поэтому $Y \sim Pois(3)$ и $E[Y] = 3$.

Где еще можно встретить распределение Пуассона?

1. **Страхование:** число страховых случаев за месяц (если нет сезонных волн и цепных аварий)
2. **Биология:** число мутаций на участке ДНК (при постоянной и малой интенсивности)