

Статистика: Справочник по теории

Профильный уровень · Необходимый минимум для решения задач

Математическое ожидание

Математическое ожидание $E(X)$ дискретной случайной величины — это её среднее ожидаемое значение при многократном повторении эксперимента. Физически его можно представить как «центр масс» распределения вероятностей.

Формула для вычисления — это сумма произведений всех возможных значений величины на их вероятности:

$$E(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \dots + x_n \cdot p_n = \sum_{i=1}^n x_i p_i$$

Пример 1 (Игровой): Вы играете в игру, бросая кубик. Если выпадает 6, вы получаете 300 рублей. В остальных случаях вы платите 60 рублей. Вероятность выигрыша $p_1 = 1/6$, вероятность проигрыша $p_2 = 5/6$.

$$E(X) = 300 \cdot \frac{1}{6} + (-60) \cdot \frac{5}{6} = 50 - 50 = 0 \text{ руб.}$$

Игра является «справедливой», так как средний ожидаемый результат равен нулю.

Пример 2 (Бытовой): Таксист анализирует свой дневной заработок. Из опыта он знает, что «отличный день» (доход 5000 руб) бывает с вероятностью 0,2, «обычный день» (3000 руб) — с вероятностью 0,5, а «плохой день» (1000 руб) — с вероятностью 0,3. Его средний ожидаемый доход за каждый день работы в долгом периоде составит:

$$E(X) = 5000 \cdot 0,2 + 3000 \cdot 0,5 + 1000 \cdot 0,3 = 1000 + 1500 + 300 = 2800 \text{ руб.}$$

Ожидание серии успехов

Часто в статистике мы имеем дело с серией независимых испытаний, в каждом из которых событие («успех») происходит с одинаковой вероятностью p .

До первого успеха:

Если мы повторяем испытания до тех пор, пока успех не произойдет *один раз*, ожидаемое количество нужных попыток вычисляется по формуле:

$$E(X) = \frac{1}{p}$$

Почему формула именно такая? Всё очень логично. Представьте, что вероятность забросить баскетбольный мяч в кольцо равна $p = \frac{1}{5}$ (то есть вы забиваете 1 мяч из 5). Разумеется, чтобы забить этот 1 мяч, вам в среднем потребуется сделать 5 бросков. Формула $\frac{1}{p}$ просто переворачивает дробь вероятности: $\frac{1}{1/5} = 5$.

До k -го успеха:

Так как испытания полностью независимы, ожидание складывается. Чтобы дожидаться ровно k успехов (не обязательно подряд), нам в среднем понадобится в k раз больше попыток:

$$E(X) = \frac{k}{p}$$

Пример: Вероятность того, что начинающий биатлонист попадет в мишень, равна $p = 0,2$ (или $\frac{1}{5}$). Чтобы попасть *хотя бы один раз*, ему в среднем потребуется $\frac{1}{0,2} = 5$ выстрелов. Для поражения *трёх мишеней* ему потребуется сделать в среднем $3 \cdot 5 = 15$ выстрелов.

Несмещённая дисперсия

Дисперсия $D(X)$ показывает, насколько сильно значения случайной величины разбросаны относительно её математического ожидания.

Как это трактовать? Дисперсия — это мера нестабильности или разброса:

- **Чем больше дисперсия**, тем сильнее реальные значения «прыгают» вокруг среднего. Это признак сильной изменчивости и нестабильности (например, резкие перепады температур зимой, скачки курса валют или игрок, который то забивает 5 голов, то ни одного).
- **Чем меньше дисперсия**, тем кучнее и плотнее результаты лежат к среднему. Это признак стабильности и предсказуемости (например, станок, который всегда вытачивает деталь строго одного размера с минимальной погрешностью).

На практике у нас часто есть лишь **выборка** данных. Обычная формула выборочной дисперсии занижает реальный разброс, так как элементы выборки группируются вокруг своего собственного среднего $\bar{x}$. Для исправления этого смещения применяется **поправка Бесселя** — мы делим сумму квадратов отклонений не на n , а на $n - 1$. Такая оценка называется **несмещённой** (S^2):

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Пример: Дана выборка из 4 чисел: 2, 4, 4, 6.

- Находим выборочное среднее: $\bar{x} = \frac{2+4+4+6}{4} = 4$.
- Считаем сумму квадратов отклонений элементов от $\bar{x}$:
 $(2-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 = 4 + 0 + 0 + 4 = 8$.
- Несмещённая дисперсия: $S^2 = \frac{8}{4-1} = \frac{8}{3} \approx 2,67$.

Неравенство Чебышёва

Неравенство Чебышёва — это «правило защиты от аномалий». Оно простыми словами говорит: *вероятность того, что величина сильно отойдёт от нормы, ограничена и падает при увеличении отступа*.

Если известна дисперсия $D(X)$, то вероятность того, что отклонение от среднего $E(X)$ составит величину ε (эпсилон) или больше, задается формулой:

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$$

Как это понимать? Обратите внимание на ε^2 в знаменателе. Чем дальше мы отступаем от нормы (чем больше «прыжок» ε), тем стремительнее — в квадрате! — падает вероятность встретить такое значение. А если дисперсия $D(X)$ маленькая (то есть процесс сам по себе стабильный), то шансы на возникновение аномалии становятся совсем крошечными.

Пример: Автомат наливает кофе. Известно, что средний объем чашки $E(X) = 200$ мл, а дисперсия $D(X) = 9$. Оценим вероятность того, что автомат сломается и нальет объем, отличающийся от нормы на 15 мл и более.

Отклонение от нормы (среднего) составляет $\varepsilon = 15$. Подставляем в формулу:

$$P(|X - 200| \geq 15) \leq \frac{9}{15^2} = \frac{9}{225} = 0,04$$

То есть вероятность получить такую аномально пустую или переполненную чашку не превышает 4%.