Занятие 22. 4 апреля

Иван Высоцкий

Дисперсия случайной величины

На этом занятии было немного задач. Основное время было посвящено рассказу про дисперсию и стандартное отклонение случайной величины.

Обычно найти математическое ожидание случайной величины недостаточно, чтобы понять, какие ее значения следует считать вполне вероятными, а какие – маловероятными. Важно понять, каким может быть отклонение значений от среднего. Например, можно найти среднее отклонение: $\mathrm{E}|X-\mathrm{E}X|$. Это совсем не плохо, но свойства этой величины оказываются не слишком удобными – во многом мешает модуль.

Зато, если рассматривать не средний модуль, а средний квадрат отклонения, получается ***дисперсия***:

$$\mathrm{D}X=\mathrm{E}(X-\mathrm{E}X)^2, \tag{1}$$

которая обладает целым рядом удобных и полезных свойств. Например, есть более простая формула:

$$\mathrm{D}X=\mathrm{E}X^2-\mathrm{E}^2X \tag{2}$$

– дисперсия равна среднему квадрату без квадрата среднего. Получается эта формула простыми преобразованиями: нужно возвести в квадрат скобку $(X-\mathrm{E}X)$ и аккуратно привести подобные, ничего не перепутав.

Вторая особенность дисперсии: часто ее можно находить по слагаемым. А именно:

$$\mathrm{D}(X+Y)=\mathrm{D}X+\mathrm{D}Y, \tag{3}$$

если только случайные величины X и Y независимы. Это свойство также несложно получить с помощью формулы (2). В ходе выкладок нужно воспользоваться свойством $\mathrm{E}(XY)=\mathrm{E}X\cdot\mathrm{E}Y$, которое верно только для независимых величин. Отсюда и требование независимости.

Пример 1. Биномиальное распределение. Проведено n независимых и одинаковых испытаний. Вероятность успеха в каждом равна p. Найти математическое ожидание и дисперсию числа наступивших успехов.

Решение. Пусть S – число успехов, а I_k – индикатор успеха в испытании номер k.

$$I_k\sim\begin{pmatrix}0 & 1\\ q & p\end{pmatrix},$$

тогда $\mathrm{E}I_k=p$ для всех индикаторов, и

$$\mathrm{E}S=\mathrm{E}I_1+\mathrm{E}I_2+\ldots+\mathrm{E}I_n=np.$$

Индикаторы независимы, поскольку результаты испытаний по условию не связаны. Следовательно, в силу (3)

$$\mathrm{D}S=\mathrm{D}(I_1+I_2+\ldots+I_n)=\mathrm{D}I_1+\mathrm{D}I_2+\ldots+\mathrm{D}I_n=n\mathrm{D}I_1.$$

Чтобы найти дисперсию одного индикатора, воспользуемся формулой (2):

$$\mathrm{D}I_1=\mathrm{E}I_1^2-\mathrm{E}^2I_1=\mathrm{E}I_1-\mathrm{E}^2I_1=p-p^2=pq.$$

Здесь буквой q, как обычно, обозначена вероятность неудачи: $q=1-p$. В выкладках использовано, что $I_1^2=I_1$ (поскольку $0^2=0$, $1^2=1$).

Следовательно, $\mathrm{D}S=npq$.

Недостаток дисперсии в том, что она измеряется в квадратных единицах. Например, если X – расстояние в метрах, то $\mathrm{D}X$ измеряется в квадратных метрах, с чем можно смириться. Но вот если X – угол в градусах, то $\mathrm{D}X$ выражается в квадратных градусах, что очень загадочно. Чтобы получить меру рассеивания, выраженную в тех же единицах, что и исходная величина, можно извлечь квадратный корень из дисперсии и получить так называемое ***стандартное отклонение***. Если вдуматься, то понятно, что стандартное отклонение всегда не меньше, чем средний модуль отклонения $\mathrm{E}|X-\mathrm{E}X|$, но все равно оно может использоваться в качестве типичного отклонения. Часто стандартное отклонение используют в качестве единицы для измерения отклонений.

В серии независимых испытаний из примера 1 стандартное отклонение равно[1]

$$\sigma=\sqrt{\mathrm{D}S}=\sqrt{npq}.$$

Например, если монету бросают 10 раз, то стандартное отклонение числа орлов равно

$$\sigma=\sqrt{10\cdot 0,5\cdot 0,5}\approx 1,58.$$

Разумно договориться, что в однократном испытании не следует рассчитывать на отклонения, превышающие 3σ, поскольку такие отклонения слишком маловероятны (и это соглашение почти в любом случае имеет смысл, как мы увидим в дальнейшем). Значит, бросая 10 монет, разумно рассчитывать на число орлов от

$$\mathrm{E}S-3\sigma=5-3\cdot 1,58\approx 0,26 \quad \text{до} \quad \mathrm{E}S+3\sigma=5+3\cdot 1,58\approx 9,74,$$

то есть в реальности нужно взять в расчет все возможные значения от 1 до 9, а вот в 0 или в 10 орлов верить не следует (это случается слишком редко). Таким образом, имеется большая неопределенность: примерно 82% всех возможных исходов вероятны.

А что, если монет 100? Диапазон разумного ожидания, конечно, расширяется, но мало по сравнению с ростом числа бросков. При том же соглашении будем ждать от

$$\mathrm{E}S-3\sigma=50-3\cdot 5=35 \quad \text{до} \quad \mathrm{E}S+3\sigma=50+3\cdot 5=65$$

орлов. Длина диапазона теперь $65-35+1=31$, что составляет уже 31% от всех теоретически возможных исходов.

Если монет 1000? Посчитайте самостоятельно: разумно ждать от 453 до 547 орлов, то есть всего 9,5% от всего возможного множества значений.

Видим, что доля вероятных значений уменьшается с ростом числа опытов. Долю вероятных значений можно сделать сколь угодно малой, увеличивая число испытаний, и это – очень важное свойство, которое лежит в основе почти всех статистических измерений. Так происходит оттого, что стандартное отклонение пропорционально не числу испытаний n, а величине $\sqrt{n}$, которая растет намного медленнее, чем само n.

Вычислим дисперсию и стандартное отклонение в другом важном опыте – стрельба до первого попадания.

Пример 2. Геометрическое распределение. Производятся одинаковые независимые испытания с вероятностью p успеха в каждом до тех пор, пока не наступит первый успех. Найти математическое ожидание, дисперсию и стандартное отклонение.

[1] Мы использовали греческую букву σ «сигма», которой часто обозначают стандартное отклонение в разных задачах, но это всего лишь наше соглашение, а не правило.

Решение. Пусть X – число испытаний, а I_k – индикатор неудачи в испытании номер k. Тогда $X = 1 + I_1 + I_2 + I_3 + \ldots$.

Распределения индикаторов найти легко:

$$I_k \sim \begin{pmatrix} 0 & 1 \\ 1-q^k & q^k \end{pmatrix},$$

откуда $\mathrm{E}\, I_k = q^k$. Следовательно,

$$\mathrm{E}\, X = 1 + \mathrm{E}\, I_1 + \mathrm{E}\, I_2 + \mathrm{E}\, I_3 + \ldots 1 + q + q^2 + q^3 + \ldots = \frac{1}{1-q} = \frac{1}{p}.$$

Результат согласуется с интуицией: если вероятность везения 1/10, то в среднем требуется 10 попыток, чтобы добиться успеха. Но вот что с дисперсией и отклонениями? Так ли уж тут хорошо работает интуиция?

К сожалению, индикаторы в этой задаче зависимы: если успех не случился в какой-то попытке, то и во всех предыдущих успеха тоже не было. Поэтому воспользоваться формулой (3) суммирования дисперсий нельзя. Придется идти более сложным путем. Применим формулу (2). Величина $\mathrm{E}\, X$ нам известна. Осталось найти $\mathrm{E}\, X^2$:

$$X^2 = \left(1^2 + I_1^2 + I_2^2 + I_3^2 + \ldots\right) + \left(2I_1 + 2I_2 + 2I_3 + \ldots\right) +$$
$$+ \left(2I_1I_2 + 2I_1I_3 + 2I_1I_4 + \ldots\right) + \left(2I_2I_3 + 2I_2I_4 + 2I_2I_5 + \ldots\right) + \ldots + 2I_mI_n + \ldots$$

Получается бесконечная сумма всех квадратов, потом удвоенных индикаторов и, наконец, удвоенных произведений всех пар индикаторов вида $2I_mI_n$, где $m < n$. Заметим, что $I_k^2 = I_k$ и что $I_mI_n = I_n$ при $m < n$ (уже отмечалось, что если в n-м испытании неудача, то во всех предыдущих тоже). Тогда

$$\begin{aligned} X^2 &= \left(1 + I_1 + I_2 + I_3 + \ldots\right) + \left(2I_1 + 2I_2 + 2I_3 + \ldots\right) + \\ &\quad + \left(2I_2 + 2I_3 + 2I_4 + \ldots\right) + \\ &\quad + \left(2I_3 + 2I_4 + 2I_5 + \ldots\right) + \ldots = \\ &= \left(1 + I_1 + I_2 + I_3 + \ldots\right) + 2\left(I_1 + 2I_2 + 3I_3 + \ldots\right). \end{aligned}$$

Перейдем к ожиданию:

$$\mathrm{E}\, X^2 = \left(1 + q + q^2 + q^3 + \ldots\right) + 2\left(q + 2q^2 + 3q^3 + \ldots\right) = \frac{1}{p} + 2\frac{q}{p}\left(p + 2qp + 3q^2p + \ldots\right).$$

Выражение в скобках – математическое ожидание $\mathrm{E}\, X$. Поэтому

$$\mathrm{E}\, X^2 = \frac{1}{p} + 2\frac{q}{p}\mathrm{E}\, X = \frac{1}{p} + 2\frac{q}{p^2} = \frac{p+2q}{p^2} = \frac{2-p}{p^2}.$$

Тогда

$$\mathrm{D}\, X = \mathrm{E}\, X^2 - \mathrm{E}^2\, X = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}.$$

Стандартное отклонение равно $\sigma = \sqrt{\mathrm{D}\, X} = \dfrac{\sqrt{q}}{p}$.

При малых вероятностях успеха p число $\sqrt{q}$ близко к единице. Поэтому вполне вероятно, что ждать первого успеха придется долго. В этой задаче математическое ожидание – весьма ненадежный ориентир. Например, если мы захотим узнать, сколько дать патронов стрелку, чтобы он с большой вероятностью поразил цель при вероятности попадания 0,1 при каждом выстреле, нужно вычислить величину $\mathrm{E}\,X+3\sigma$ (как прежде, считаем вероятными значения, удаленные от среднего не больше чем на три стандартных отклонения). Получаем: $\frac{1}{0{,}1}+3\cdot\frac{\sqrt{0{,}9}}{0{,}1}\approx 10+3\cdot 10\cdot 0{,}95=38{,}5$. То есть стрелка придется снабдить 39 патронами, не рассчитывая на то, что число выстрелов будет близко к ожидаемым десяти.

Пример 3. Гипергеометрическое распределение. Из урны, в которой N шаров, из них k черных, а остальные – белые, наудачу вынимают n шаров. Найти математическое ожидание, дисперсию и стандартное отклонение случайной величины X «Число черных шаров среди вынутых».

Указание. Если вы разобрались в примере 2, то задача не покажется вам сложной. Нужно только учесть характер зависимости индикаторов. Скорее, трудности будут с алгебраическими преобразованиями в конце.

Ответ: $\mathrm{E}\,X=n\frac{k}{N}$. $\mathrm{D}\,X=\frac{kn(N-n)(N-k)}{N^2(N-1)}$.