Теория вероятностей и математическая статистика

Центральная предельная теорема.

Глеб Карпов

ВШБ Бизнес-информатика

Напоминание: Линейная комбинация случайных величин

Предположим, что у нас есть X и Y — две случайные величины. Следующие свойства работают для любой возможной природы этих переменных.

1. Линейное свойство математического ожидания: $E[aX \pm bY] = aE[X] \pm bE[Y]$.
2. Дисперсия линейной комбинации: $Var[aX \pm bY] = a^2 Var[X] + b^2 Var[Y] \pm 2ab\Big(E[XY] - E[X]E[Y]\Big)$.
3. Если X и Y независимы: $Var[aX \pm bY] = a^2 Var[X] + b^2 Var[Y]$.

Иллюстративный пример

Функция от двух дискретных случайных величин

Предположим, что мы бросаем два 6-гранных кубика, независимых друг от друга. В итоге мы наблюдаем дискретный случайный вектор (X, Y), где X и Y — случайные величины, соответствующие выпавшим числам на каждом из кубиков. Поскольку существует 36 различных пар, совместная функция вероятности задается как: $P(X = x_i, Y = y_j) = \frac{1}{36}$.

Введем новую случайную величину T как функцию от X и Y: $T = f(X, Y) = X + Y$. Построим функцию вероятности для случайной величины T.

Иллюстративный пример

Построение функции вероятности для $T = X + Y$

Для каждого значения t случайной величины T найдем все пары (x, y), которые дают в сумме t:

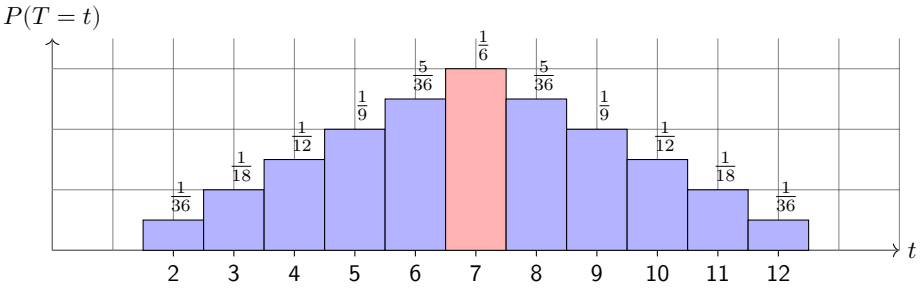
- $T = 2$: только $(1, 1)$
- $T = 3$: $(1, 2), (2, 1)$
- $T = 4$: $(1, 3), (2, 2), (3, 1)$
- $T = 5$: $(1, 4), (2, 3), (3, 2), (4, 1)$
- $T = 6$: $(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)$
- $T = 7$: $(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$
- $T = 8$: $(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)$
- $T = 9$: $(3, 6), (4, 5), (5, 4), (6, 3)$
- $T = 10$: $(4, 6), (5, 5), (6, 4)$
- $T = 11$: $(5, 6), (6, 5)$
- $T = 12$: только $(6, 6)$

Иллюстративный пример

Таблица функции вероятности

t	2	3	4	5	6	7	8	9	10	11	12
$P(T=t)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

График функции вероятности



Промежуточный вывод

Мы на очень простом примере можем пронаблюдать, как функция вероятности суммы одинаково распределенных дискретных величин получает “усиление” в центральных значениях, так как к этим элементарным исходам приводят больше комбинаций изначальных величин.

Центральная предельная теорема

Предпосылки

- Пусть $X_1, X_2, \ldots$ — последовательность независимых случайных величин, взятых из одного и того же распределения, т.е. все X_i имеют одинаковое математическое ожидание $E[X] = \mu$, одинаковую дисперсию $Var[X] = \sigma^2 < \infty$.
- Мы конструируем **новую** случайную величину как частичную сумму первых n случайных величин X_i:

$$S_n = \sum_{i=1}^{n} X_i.$$

- Согласно свойствам математического ожидания и дисперсии, характеристики новой случайной величины:

$$E[S_n] = E\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} E[X_i] = n\mu$$

$$Var[S_n] = Var\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} Var[X_i] = n\sigma^2$$

Центральная предельная теорема

- **Центральная предельная теорема**: распределение такой случайной величины S_n стремится к нормальному распределению при $n \to \infty$:

$$S_n \to Y \sim \mathcal{N}(n\mu, n\sigma^2)$$

- Душно:

$$\forall x \in \mathbb{R} \quad P(S_n < x) \underset{n\to\infty}{\longrightarrow} P(Y < x), \text{ где } Y \sim \mathcal{N}(n\mu, n\sigma^2),$$

что означает, что в каждой точке x функция распределения $F_{S_n}(x)$ стремится к функции распределения $F_Y(x)$, т.е. они буквально "накладываются" друг на друга и теперь характеризуют одну и ту же случайную величину.

- Проще говоря: $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$, когда n достаточно велико.
- Обычно "достаточно велико" это $n \geq 30$.

Центральная предельная теорема

Формулировка через стандартизацию

- Проводим операцию стандартизации введенной ранее случайной величины S_n:

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{Var[S_n]}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- Характеристики:

$$E[Z_n] = E\left[\frac{S_n - n\mu}{\sigma\sqrt{n}}\right] = E\left[\frac{S_n}{\sigma\sqrt{n}}\right] - E\left[\frac{n\mu}{\sigma\sqrt{n}}\right] = \frac{n\mu}{\sigma\sqrt{n}} - \frac{n\mu}{\sigma\sqrt{n}} = 0$$

$$Var[Z_n] = Var\left[\frac{S_n - n\mu}{\sigma\sqrt{n}}\right] = Var\left[\frac{S_n}{\sigma\sqrt{n}}\right] + Var\left[\frac{n\mu}{\sigma\sqrt{n}}\right] = \frac{Var[S_n]}{n\sigma^2} + 0 = 1$$

- Тогда говорим, что:

$$\forall x \in \mathbb{R} \quad P(Z_n < x) \underset{n\to\infty}{\longrightarrow} P(Z < x), \text{ где } Z \sim \mathcal{N}(0,1)$$

Или проще: стандартизованная частичная сумма $Z_n \sim \mathcal{N}(0,1)$, когда n достаточно велико. Обычно требуем $n \geq 30$.