Теория вероятностей и математическая статистика

Устойчивость нормального распределения. ИТМЛ. Хи-квадрат распределение.

Глеб Карпов

ВШБ Бизнес-информатика

Напоминание: Нормальное распределение

- Случайная величина X имеет нормальное распределение с параметрами μ и σ^2, если её плотность:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Обозначение: $X \sim \mathcal{N}(\mu, \sigma^2)$
- Математическое ожидание: $E[X] = \mu$
- Дисперсия: $\mathrm{Var}(X) = \sigma^2$

Многомерное нормальное распределение

- До сих пор мы рассматривали **одномерные** нормальные распределения.
- Как обобщить понятие нормального распределения на случай нескольких случайных величин?

Многомерное нормальное распределение

Пара случайных величин (X, Y) имеет **двумерное нормальное распределение**, если для всех $a, b \in \mathbb{R}$ линейная комбинация $aX + bY$ имеет одномерное нормальное распределение.

Вектор случайных величин $\mathbf{X} = (X_1, X_2, \dots, X_n)$ имеет **многомерное нормальное распределение**, если для всех векторов $\mathbf{a} \in \mathbb{R}^n$ скалярное произведение $\mathbf{a}^\top \mathbf{X}$ имеет одномерное нормальное распределение.

$\mathbf{a}^\top \mathbf{X} = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$

Устойчивость нормального распределения

- **Ключевое свойство**: Нормальное распределение **устойчиво** относительно линейных преобразований.
- Если $X \sim \mathcal{N}(\mu, \sigma^2)$, то для любых констант $a \neq 0$ и b:

$$Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

- **Интерпретация**: Линейное преобразование нормальной случайной величины даёт нормальную случайную величину.

Устойчивость нормального распределения

Линейная комбинация двух независимых нормальных величин

- Частный случай: если $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ и $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ независимы, то:

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

- **Связь с многомерным нормальным**: Независимые нормальные величины (X_1, X_2) формируют двумерное нормальное распределение, поэтому любая линейная комбинация $aX_1 + bX_2$ имеет нормальное распределение.

Устойчивость нормального распределения

Обобщение на n переменных

- Пусть $X_1, \dots, X_n$ — независимые нормальные случайные величины:

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \dots, n$$

- **Важно**: Вектор (коллекция, набор) независимых нормальных случайных величин имеет **многомерное нормальное распределение**.
- По определению многомерного нормального распределения, **любая** линейная комбинация имеет нормальное распределение:

$$Y = \sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Пример 1: Линейная комбинация

- Пусть $X \sim \mathcal{N}(2, 1)$ и $Y \sim \mathcal{N}(3, 4)$ — независимы.
- Найти распределение $W = 2X - 3Y$.
- **Решение**:

$$E[W] = 2 \cdot 2 - 3 \cdot 3 = 4 - 9 = -5$$
$$\mathrm{Var}(W) = 2^2 \cdot 1 + (-3)^2 \cdot 4 = 4 + 36 = 40$$

- Таким образом: $W \sim \mathcal{N}(-5, 40)$

Пример 2: Выборочное среднее

- Пусть $X_1, \dots, X_n$ — независимые наблюдения из $\mathcal{N}(\mu, \sigma^2)$.
- Выборочное среднее: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} S_n$
- Используя свойство устойчивости:

$$\bar{X} = \frac{1}{n} S_n, \quad \text{где } S_n \sim \mathcal{N}(n\mu, n\sigma^2)$$

$$\bar{X} \sim \mathcal{N}\left(\frac{1}{n} \cdot n\mu, \left(\frac{1}{n}\right)^2 \cdot n\sigma^2\right)$$

$$= \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Напоминание: Биномиальное распределение

- Пусть $X \sim \text{Bin}(n, p)$ — биномиальная случайная величина.
- X можно представить как сумму n независимых бернуллиевских случайных величин:

$$X = X_1 + X_2 + ... + X_n, \quad \text{где } X_i \sim \text{Bernoulli}(p)$$

- Характеристики:
  - $E[X] = np$
  - $\text{Var}(X) = np(1-p)$

Интегральная теорема Муавра-Лапласа (ИТМЛ)

Пусть $X \sim \text{Bin}(n, p)$. При больших n биномиальное распределение приближается нормальным:

$$X \sim \mathcal{N}(np, np(1-p))$$

Условия: n достаточно велико (обычно $n \geq 30$), и желательно $np > 5$ и $n(1-p) > 5$.

Это частный случай ЦПТ для суммы бернуллиевских случайных величин!

Нормализованная форма ИТМЛ

- Нормализованная случайная величина:

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

- При больших n:

$$Z \sim \mathcal{N}(0, 1)$$

- Это позволяет использовать стандартную нормальную таблицу для вычисления вероятностей биномиального распределения!

Применение ИТМЛ: вычисление вероятностей

- Для биномиального распределения $X \sim \text{Bin}(n, p)$:

$$P(a \leq X \leq b) \approx P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - np}{\sqrt{np(1-p)}}\right),$$

  где $Z \sim \mathcal{N}(0, 1)$.

- **Поправка на непрерывность** (без доказательства): поскольку X — дискретная случайная величина, а нормальное распределение — непрерывное, для повышения точности приближения расширяем интервал на 0.5 с каждой стороны:

$$P(a \leq X \leq b) \approx P\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

- Вероятность $P(a \leq X \leq b)$ включает все целые значения от a до b включительно. В непрерывном приближении это соответствует интервалу $[a - 0.5, b + 0.5]$.

Пример: применение ИТМЛ

- Пусть $X \sim \mathrm{Bin}(100, 0.3)$. Найти $P(25 \le X \le 35)$.
- **Точное вычисление** (через биномиальное распределение):

$$P(25 \le X \le 35) = \sum_{k=25}^{35} \binom{100}{k} (0.3)^k (0.7)^{100-k} \approx 0.7698$$

- **Приближение через нормальное распределение**:
  - Параметры: $\mu = np = 30$, $\sigma^2 = np(1-p) = 21$, $\sigma \approx 4.58$
  - С поправкой на непрерывность:

$$\begin{aligned} P(25 \le X \le 35) &\approx P\left(\frac{24.5 - 30}{4.58} \le Z \le \frac{35.5 - 30}{4.58}\right) \\ &= P(-1.20 \le Z \le 1.20) \\ &\approx 0.7699 \end{aligned}$$

- **Сравнение**: Нормальное приближение даёт результат 0.7699, очень близко к точному значению 0.7698.

Распределение X^2 (хи-квадрат)

- Пусть $Z_1, \dots, Z_k$ — независимые стандартные нормальные случайные величины: $Z_i \sim \mathcal{N}(0, 1)$.
- Определим новую случайную величину:

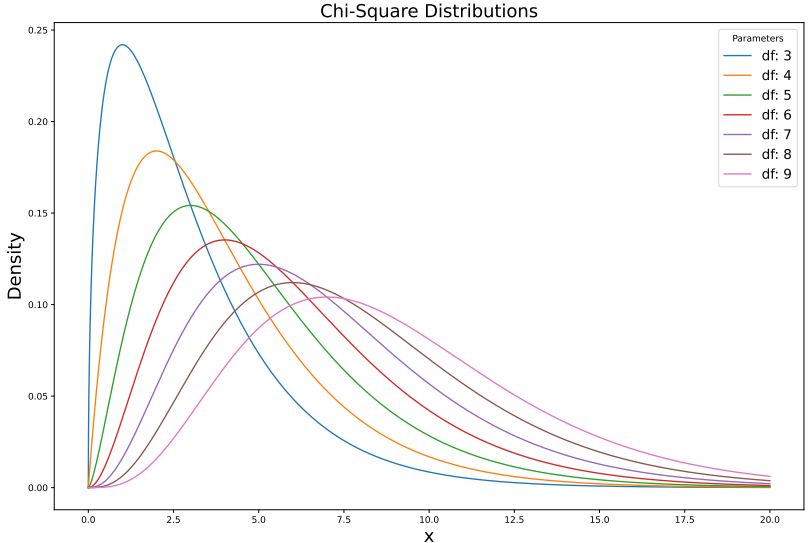
$$\chi^2(k) = \sum_{i=1}^{k} Z_i^2$$

- Распределение такой случайной величины называется **распределением хи-квадрат**.
- Это распределение играет важную роль в статистических процедурах.

Степени свободы

- Функция плотности распределения хи-квадрат **критично зависит** от числа слагаемых k в сумме.
- Число слагаемых k называется **степенями свободы** (degrees of freedom, df).
- Правильное обозначение: $\chi^2(k)$ — читается как "хи-квадрат распределение с k степенями свободы".

Функция плотности распределения хи-квадрат



Математическое ожидание и дисперсия распределения хи-квадрат

- Если $Y \sim \chi^2(k)$, то:

$$E[Y] = k, \quad \mathrm{Var}(Y) = 2k$$

- Доказательство для $E[Y]$:

  Поскольку $Y = \sum_{i=1}^{k} Z_i^2$, где $Z_i \sim \mathcal{N}(0,1)$ независимы:

$$E[Y] = \sum_{i=1}^{k} E[Z_i^2] = \sum_{i=1}^{k} \left(\mathrm{Var}(Z_i) + (E[Z_i])^2\right) = \sum_{i=1}^{k} (1+0) = k$$

Свойство суммы отклонений от среднего

Пусть $X_1, \dots, X_k$ — случайная выборка из некоторой генеральной совокупности. Тогда:

$$\sum_{i=1}^{k}(X_i - \bar{X}) = 0$$

где $\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$ — выборочное среднее.

Свойство суммы отклонений от среднего

$$\sum_{i=1}^{k}\left(X_i - \bar{X}\right) = \sum_{i=1}^{k}\left(X_i\right) - k\bar{X}$$

$$= \sum_{i=1}^{k} X_i - k\frac{\sum_{i=1}^{k} X_i}{k}$$

$$= \sum_{i=1}^{k} X_i - \sum_{i=1}^{k} X_i = 0$$

Теорема о $k\bar{Z}^2$

Пусть $Z_1, \dots, Z_k$ — случайная выборка из стандартного нормального распределения, т.е. $Z_i \sim \mathcal{N}(0, 1)$.

Тогда случайная величина $k\bar{Z}^2$ имеет распределение хи-квадрат с одной степенью свободы:

$$k\bar{Z}^2 \sim \chi^2(1)$$

Теорема о $k\bar{Z}^2$

- Характеристики:

$$E[\bar{Z}] = \frac{1}{k}\sum_{i=1}^{k} E[Z_i] = 0, \quad \mathrm{Var}[\bar{Z}] = \frac{1}{k^2}\sum_{i=1}^{k} \mathrm{Var}[Z_i] = \frac{1}{k^2}k = \frac{1}{k}$$

- По свойству устойчивости, как сумма независимых нормальных случайных величин, $\bar{Z} \sim \mathcal{N}\left(0, \frac{1}{k}\right)$
- Далее рассмотрим случайную величину $\sqrt{k}\bar{Z}$:

$$E[\sqrt{k}\bar{Z}] = 0, \quad \mathrm{Var}(\sqrt{k}\bar{Z}) = k\mathrm{Var}(\bar{Z}) = k\frac{1}{k} = 1$$

- Значит, $\sqrt{k}\bar{Z} \sim \mathcal{N}(0,1)$ — имеет стандартное нормальное распределение. Поэтому одна такая величина в квадрате будет распределена как хи-квадрат:

$$(\sqrt{k}\bar{Z})^2 = k\bar{Z}^2 \sim \chi^2(1)$$

Разложение суммы квадратов

- Рассмотрим преобразование:

$$\sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \left[(Z_i - \bar{Z} + \bar{Z})^2 \right] = \sum_{i=1}^{k} \left[\Big((Z_i - \bar{Z}) + \bar{Z} \Big)^2 \right]$$

$$= \sum_{i=1}^{k} \left[(Z_i - \bar{Z})^2 + 2 (Z_i - \bar{Z}) \bar{Z} + \bar{Z}^2 \right]$$

$$= \sum_{i=1}^{k} (Z_i - \bar{Z})^2 + 2\bar{Z} \sum_{i=1}^{k} (Z_i - \bar{Z}) + k\bar{Z}^2$$

- По свойству суммы отклонений: $\sum\limits_{i=1}^{k} (Z_i - \bar{Z}) = 0$
- Получаем:

$$\sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} (Z_i - \bar{Z})^2 + k\bar{Z}^2$$

Распределение суммы квадратов отклонений

- Из разложения:

$$\underbrace{\sum_{i=1}^{k} Z_i^2}_{k \text{ степеней свободы}} = \underbrace{\sum_{i=1}^{k} (Z_i - \bar{Z})^2}_{(k-1) \text{ степеней свободы}} + \underbrace{k\bar{Z}^2}_{1 \text{ степень свободы}}$$

- Известно:
  - $\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$, и $k\bar{Z}^2 \sim \chi^2(1)$
- В итоге получаем, что оставшееся слагаемое имеет $(k-1)$ степень свободы:

$$\boxed{\sum_{i=1}^{k} (Z_i - \bar{Z})^2 \sim \chi^2(k-1)}$$

- **Ключевой результат**: Сумма квадратов отклонений от выборочного среднего имеет распределение хи-квадрат с $(k-1)$ степенями свободы. Это фундаментальный результат для статистических тестов и доверительных интервалов.

Выборочная дисперсия

- Пусть $X_1, \dots, X_n$ — случайная выборка из нормального распределения: $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
- **Выборочная дисперсия**:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- Выборочная дисперсия используется для оценки неизвестной дисперсии генеральной совокупности σ^2.

Связь выборочной дисперсии с хи-квадрат

- Нормализуем наблюдения: $Z_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$
- Тогда выборочное среднее нормализованных величин: $\bar{Z} = \frac{\bar{X} - \mu}{\sigma}$
- Из предыдущих результатов:

$$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1)$$

Распределение выборочной дисперсии

- Подставляя определение выборочной дисперсии:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

- **Важный результат**:

$$\boxed{\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)}$$

- Это означает, что выборочная дисперсия имеет распределение хи-квадрат с $(n-1)$ степенями свободы.
- **Математическое ожидание и дисперсия**:

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n-1, \quad \mathrm{Var}\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1)$$

- **Следствия для выборочной дисперсии**:

$$E[s^2] = \sigma^2, \quad \mathrm{Var}(s^2) = \frac{2\sigma^4}{n-1}$$