Основные математико-статистические методы и теоремы

1. Теорема Байеса

Если события A и B имеют положительную вероятность, то условная вероятность $P(A \mid B)$ выражается через:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Используется для пересмотра вероятностей с учётом новой информации. Применяется в кредитном скоринге, A/B-тестах, предсказательных моделях.

2. Центральная предельная теорема

Пусть $X_1, \ldots, X_n$ — независимые одинаково распределённые СВ с конечным математическим ожиданием μ и дисперсией σ^2. Тогда при $n \to \infty$:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Обосновывает использование нормального распределения для выборочных средних.

3. Свойства нормального распределения

Распределение $\mathcal{N}(\mu, \sigma^2)$ симметрично и полностью описывается двумя параметрами. Для стандартного нормального $Z \sim \mathcal{N}(0, 1)$:

$$P(|Z| \leq 1.96) \approx 0.95, \quad P(|Z| \leq 2.58) \approx 0.99$$

Широко применяется в доверительных интервалах и тестах гипотез.

4. Биномиальное распределение и нормальная аппроксимация

Если $X \sim \mathrm{Bin}(n, p)$, то при больших n и умеренных p:

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

Используется для приближённого построения интервалов и проверки гипотез по долям.

5. Закон больших чисел

Пусть $X_1, \dots, X_n$ — независимые одинаково распределённые СВ с $\mathbb{E}[X_i] = \mu$. Тогда:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu \quad \text{при } n \to \infty$$

Оправдывает использование выборочного среднего как оценки математического ожидания.

6. Неравенство Чебышёва

Для любой СВ X с конечной дисперсией и любого $\varepsilon > 0$:

$$P(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}$$

Полезно при отсутствии информации о виде распределения.

7. Доверительный интервал для среднего

Если $X_1, \dots, X_n$ — выборка из нормального распределения, то:

$$\bar{X} \pm t_{1-\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$$

где s — выборочное стандартное отклонение.

8. Проверка гипотез (z- и t-тест)

Для проверки $H_0 : \mu = \mu_0$:

$$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{или} \quad t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Решение принимается на основе сравнения статистики с критическими значениями.

9. Доверительный интервал для доли

При достаточной выборке n и пропорции $\hat{p}$:

$$\hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Используется в анализе конверсии, оттока и т. д.

10. Критерий согласия Пирсона (хи-квадрат)

Для категориальных данных:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

При верной H_0, χ^2 имеет распределение χ^2 с $k-1$ степенями свободы.

11. Метод максимального правдоподобия

Для плотности $f(x;\theta)$ и выборки $x_1, \ldots, x_n$:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \ell(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

MLE-оценка: $\hat{\theta}_{\text{MLE}} = \arg\max \ell(\theta)$

12. Логистическая регрессия

Для бинарного Y:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

Логарифм отношения шансов:

$$\log \frac{P(Y = 1)}{1 - P(Y = 1)} = \beta_0 + \cdots + \beta_p x_p$$

13. Bootstrap-оценка

Повторная генерация выборок с возвращением:

- Выбираются B подвыборок из исходной
- Для каждой считается статистика (среднее, медиана и др.)
- Получается эмпирическое распределение оценки

Особенно полезен при малом объёме данных или сложной форме распределения.