Математика для Data Science. Теория вероятностей. Шпаргалка

Содержание

Пятая неделя. Статистика

Арифметика случайных величин и нормальное распределение

Следующие формулы выполнены как для дискретных, так и для непрерывных случайных величин:

Математическое ожидание суммы. Математическое ожидание суммы двух случайных величин это сумма их математических ожиданий. То есть

$$E[X+Y] = E[X] + E[Y].$$

Дисперсия суммы. Если две случайных величины независимы, то дисперсия их суммы это сумма дисперсий. То есть

$$Var(X+Y) = Var(X) + Var(Y).$$

Сложение с константой. Для любой случайной величины X и любого числа $c \in \mathbb{R}$ выполнено

$$E[X+c] = E[X] + c$$

и

$$Var(X+c) = Var(X).$$

Умножение на константу. Для любой случайной величины X и любого числа $c \in \mathbb{R}$ выполнено

$$E[cX] = cE[X]$$

и

$$Var(cX) = c^2 Var(X).$$

Нормальное распределение

Нормальное распределение с математическим ожиданием 0 и дисперсией 1 обозначается $N(0,1)$. Оно имеет такую функцию плотности распределения:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}\ x^2\right)}$$

Нормальное распределение с математическим ожиданием μ и дисперсией σ^2 обозначается $N(\mu, \sigma^2)$. У него такая функция плотности распределения:

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

Случайная величина, имеющая распределение $N(\mu, \sigma^2)$ для каких-то μ и σ, называется *нормально распределённой*.

Правило двух и трех сигм. Пусть случайная величина ξ имеет нормальное распределение со средним μ и дисперсией σ^2. Тогда:

- $P(\mu - \sigma < \xi < \mu + \sigma\ ) \ = 0.682\cdots \approx 0.68$
- $P(\mu - 2\sigma < \xi < \mu + 2\sigma\ ) = 0.954\cdots \approx 0.95$. Другими словами, вероятность получить результат, отклоняющийся от μ хотя бы на 2σ, меньше 0.05.
- $P(\mu - 3\sigma < \xi < \mu + 3\sigma\ ) = 0.997\cdots \approx 0.99$. Другими словами, вероятность получить результат, отклоняющийся от μ хотя бы на 3σ, меньше 0.01.

Утверждение. Сумма нескольких совместно независимых нормально распределённых случайных величин – это тоже нормально распределенная случайная величина.

Статистический тест

Реализация случайной величины – это конкретное число, которым стала эта случайная величина после измерения.

Шаблон статистических тестов

1. **Выборка.** Выборка это реализация набора случайных величин $x_1, \ldots, x_n$, то есть это n чисел. Обычно предполагают, что случайные величины $x_1, \ldots, x_n$ совместно независимы и имеют одинаковое распределение.

2. **Гипотезы и предположения.** Выбор гипотез H_0 и H_1, то есть сформулировать свой вопрос на языке теории вероятностей.

3. **Статистика.** Нам нужно как-то объединить величины $x_1, \ldots, x_n$, в одну случайную величину $T(x_1, \ldots, x_n)$. Эту величину называют *статистикой*. При условии что H_0 выполнена, нужно найти распределение случайной величины $T(x_1, \ldots, x_n)$.

4. **Уровень значимости.** *Уровень значимости* это число α отвечающее за вероятность *ошибки первого рода*. То есть за вероятность отвергнуть H_0 в случае, когда H_0 выполнена. Обычно берут $\alpha = 0.05$.

5. **Критическое множество.** Случайная величина $T(x_1, \ldots, x_n)$ принимает значения в $\mathbb{R}$. Нужно выделить подмножество $C_\alpha \subset \mathbb{R}$, по которому мы будем решать, принимать или отвергать H_0. Вероятность попадания T в множество C_α должна быть равна α. Обычно в качестве C_α берут множества вида:

   - $[a, +\infty)$ – если отклонение статистики T вверх свидетельствует в пользу H_1.
   - $(-\infty, b]$ – если отклонение статистики T вниз свидетельствует в пользу H_1.
   - $(-\infty, b] \cup [a, +\infty)$ – если отклонение статистики T от какого-то значения в любую сторону свидетельствует в пользу H_1.

6. **Статистический критерий.** Если реализация T не попала в множество C_α, то мы принимаем H_0. Если T попала в множество C_α, то мы отвергаем H_0 и принимаем H_1.

Ошибка первого рода (принять H_1 при верной H_0). Вероятность допустить ошибку первого рода всегда равна уровню значимости α. Это следует из нашего построения множества C_α.

Ошибка второго рода (принять H_0 при верной H_1). Найдём распределение T при условии, что выполнена H_1. Вероятность того, что так распределённая T не попала в критическое множество, это и есть вероятность ошибки второго рода. Другими словами $\beta = P(T \notin C_\alpha | H_1)$. Действительно, условие $T \notin C_\alpha$ как раз соответствует тому, что мы приняли H_0 и отвергли H_1.

Число $(1 - \beta)$ называют *мощностью* статистического критерия.

ЗБЧ и ЦПТ

Теорема [ЗБЧ]. Пусть $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ — бесконечная последовательность независимых одинаково распределённых случайных величин, имеющих конечное мат.ожидание μ. Обозначим среднее арифметическое первых n случайных величин $\xi_1, \xi_2, \ldots, \xi_n$ так:

$$\bar{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i.$$

Тогда

$$\bar{\xi}_n \xrightarrow[n \to +\infty]{\text{по вероятности}} \mu.$$

То есть $\forall \varepsilon > 0$ выполнено

$$\lim_{n \to +\infty} P(|\bar{\xi}_n - \mu| > \varepsilon) = 0$$

Теорема [ЦПТ]. Пусть $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ — бесконечная последовательность независимых одинаково распределённых случайных величин, имеющих конечное мат.ожидание μ и дисперсию σ^2.

Тогда

$$\frac{\bar{\xi}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to+\infty]{\text{по распределению}} N(0,1)$$

где $N(0,1)$ — нормальное распределение со средним 0 и дисперсией 1.

Обозначение. (неформально) Стрелка

$$\eta_n \xrightarrow[n\to+\infty]{\text{по распределению}} F$$

означает, что при n стремящемся к плюс бесконечности распределение случайной величины η_n близко к распределению F.

Неформально ЦПТ можно сформулировать так:

$$\bar{\xi}_n \xrightarrow[n\to+\infty]{\text{в неформальном смысле}} N\left(\mu, \frac{\sigma^2}{n}\right).$$