Теория вероятностей и математическая статистика

Введение в задачу статистики. Выборочные распределения.

Глеб Карпов

ФКН ВШЭ

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: $Var[aX \pm bY] = a^2 Var[X] + b^2 Var[Y]$.

Пример из прошлого

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:

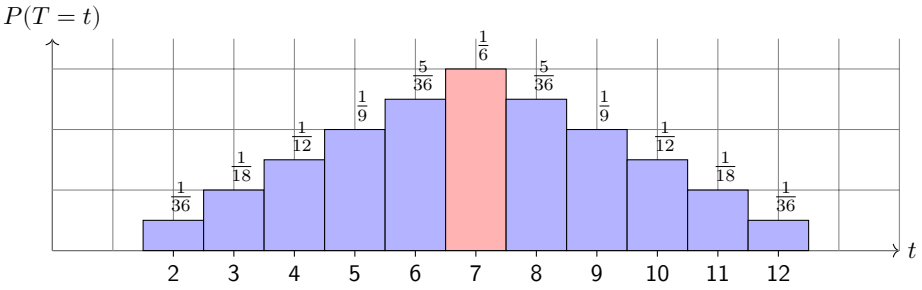
- $T = 2$: только $(1, 1)$
- $T = 3$: $(1, 2), (2, 1)$
- $T = 4$: $(1, 3), (2, 2), (3, 1)$
- $T = 5$: $(1, 4), (2, 3), (3, 2), (4, 1)$
- $T = 6$: $(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)$
- $T = 7$: $(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$
- $T = 8$: $(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)$
-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

Пример из прошлого

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



$E[S_n] = n \cdot \mu$; $Var[S_n] = n \cdot \sigma^2$

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

$X_1 \dots X_n$

$E[X_i] = \mu$

$Var[X_i] = \sigma^2$

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 $n \geq 30$.

Введение в статистику: пример

X – прибыль в день

$F_X(x, \omega)$

$E[X]$

$Var[X]$

$$\frac{x_1 + \ldots + x_{30}}{30}$$

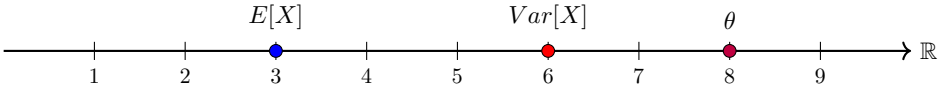
$$\frac{x_1 \quad \ldots \quad x_{365}}{365}$$

Введение в статистику: пример

- Хотим открыть магазин определенного вида товаров. Первичная задача: оценить средний доход в день, чтобы планировать бизнес. В соседнем районе есть похожий магазин, и мы пользуемся старым добрым промышленным шпионажем, чтобы собрать информацию о доходах такого магазина.
- Если формализовать, X - случайная величина, доход магазина в день. Каждый день мы получаем некую её **реализацию** x. На основе собранных $(x_1, ..., x_n)$ реализаций хотим узнать, как минимум, неизвестное $E[X]$, i.e. средний доход магазина в день.
- Случайные процессы вокруг нас — запаянные чёрные ящики — случайные генераторы. Мы хотим "расшифровать" один такой ящик и выяснить его параметры и/или свойства, например, математическое ожидание и дисперсию. Для этого мы наблюдаем случайную величину некоторое количество раз и затем делаем выводы на основе накопленной информации.

Введение в статистику: визуализация

Идеальная ситуация: мы знаем характеристики / параметры



Реальная ситуация: параметры неизвестны

Значения будто скрыты от нас туманом

- Параметры случайной величины — числа, **точки** на числовой оси. Процесс угадывания этих значений обычно называется **точечной оценкой**, то есть мы хотим эти точки на числовой оси как можно ближе угадать.

Что изучает статистика?

- Статистика — это совокупность процедур и принципов для сбора и анализа информации с целью принятия решений в условиях неопределенности.
- В теории вероятностей мы идем от предполагаемой модели к вероятности конкретного исхода, т.е. от общего к частному. В статистике задачи решаются почти полностью в обратном направлении. Статистика исследует относительно небольшой конкретный исход, и цель — узнать что-то о глобальных свойствах случайного эксперимента.
- Таким образом, несмотря на тесную связь между вероятностью и статистикой — между ними прослеживается четкое различие.

Случайная выборка и её реализация

- Собранные данные обычно называются выборкой, но в статистике мы одновременно имеем дело с двумя различными типами выборок.
- Случайная выборка — это вектор (коллекция, набор, совокупность) независимых и одинаково распределенных (i.i.d.) случайных величин:

$$\mathcal{X} = (X_1, X_2, \ldots, X_n), \quad f_{X_i}(x) = f_{X_j}(x), \ \forall i, j \in [1, n], \forall x.$$

- Реализация случайной выборки — это набор наблюдений из случайной выборки $\mathcal{X}$, набор конкретных чисел:

$$x = (x_1, x_2, \ldots, x_n).$$

- Генеральная совокупность (популяция) — полное множество объектов, обладающих интересующим признаком, несущих реализацию интересующей нас случайной величины. Извлекая наблюдения из генеральной совокупности, мы можем сформировать реализацию выборки.

Статистики как случайные величины

- **Статистика** — не только название курса, но и любая функция, зависящая только от переменных случайной выборки, т.е. $g = g(X_1, \dots, X_n)$.
- Каждая статистика будет принимать новое значение для новой реализации $(x_1, \dots, x_n)$ случайной выборки по сравнению с предыдущей реализации выборки.
- Поэтому мы рассматриваем статистики как случайные величины! Как случайные величины, они имеют свои собственные распределения и характеристики. Вероятностное распределение статистики $Y = g(X_1, \dots, X_n)$ называется **выборочным распределением** для Y.

Выборочные распределения

- Пусть $\mathcal{X} = (X_1, \dots, X_n)$ — случайная выборка со средним $\mu = E[X_i]$ и дисперсией $\sigma^2 = Var[X_i] < \infty$.
- Возможная статистика — это, например, рассмотренная ранее сумма всех элементов $S_n = \sum\limits_{i=1}^{n} X_i$. Её характеристики: $E[S_n] = n\mu$, $Var[S_n] = n\sigma^2$
- Одна из самых важных статистик — выборочное среднее:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Характеристики выборочного среднего $\bar{X}$:
  - $E[\bar{X}] = \mu$ $\quad E\left[\frac{X_1}{n} + \dots + \frac{X_n}{n}\right] = \frac{1}{n}E[X_1] + \dots + \frac{1}{n}E[X_n] = \frac{n \cdot \mu}{n} = \mu$
  - $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ $\quad Var\left[\frac{X_1}{n} + \dots + \frac{X_n}{n}\right] = \frac{1}{n^2}\sum Var[X_i] = \dots = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$
- При выполнении условий ЦПТ ($n \geq 30$) можем заявлять, что:

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