Теория вероятностей и математическая статистика

Введение в статистику. Выборочные распределения.

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

Напоминание: функция от двух дискретных случайных величин

Предположим, что мы бросаем два 6-гранных кубика, независимых друг от друга в любом смысле. 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: $P(X = x_i, Y = y_j) = \frac{1}{36}$.

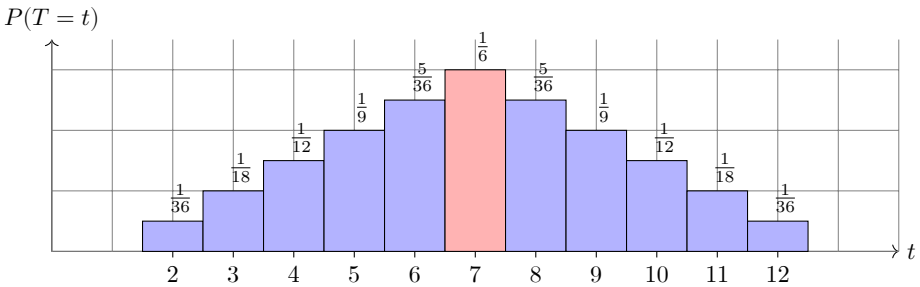
Введем новую случайную величину T как функцию от X и Y: $T = f(X, Y) = X + Y$. Построим функцию вероятности для случайной величины T.

Напоминание: функция от двух дискретных случайных величин

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



Центральная предельная теорема

- Пусть $X_1, \ldots, X_n$ — последовательность независимых случайных величин, взятых из **одного и того же** распределения, т.е. все X_i имеют одинаковое математическое ожидание $\mu = E[X]$ и конечную дисперсию $\sigma^2 = Var[X]$.

- Мы конструируем новую случайную величину: $S_n = \sum_{i=1}^{n} X_i$.

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

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

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

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

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

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

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.
- Итого: перед нами полностью запаянный черный ящик, внутри которого реализуется случайный эксперимент, и нам наружу выкидываются числа. Мы хотим "расшифровать черный ящик" и выяснить свойства этого процесса, например, математическое ожидание и дисперсию - в целом, **параметры** случайной величины. Для этого мы будем наблюдать величину определенное количество раз и делать выводы на основе накопленной информации.

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:

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

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

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

- 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 $\sigma^2 = Var[X_i] < \infty$.

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 $S_n = \sum_{i=1}^{n} X_i$. Её характеристики: $E[S_n] = n\mu$, $Var[S_n] = n\sigma^2$

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:

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

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 $\bar{X}$:

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 $\bar{X}$:
  - $E[\bar{X}] = \mu$

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 $S_n = \sum_{i=1}^{n} X_i$. 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 $\bar{X}$:
  - $E[\bar{X}] = \mu$
  - $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$

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 $\bar{X}$:
  - $E[\bar{X}] = \mu$
  - $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- При выполнении условий ЦПТ ($n \geq 30$) можем заявлять, что:

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