# Числовые характеристики распределений. Функция распределения.

Леонид Иосипой

Курс «Вероятностные модели и статистика»
Центр непрерывного образования, ВШЭ

6 апреля 2021

- Повторение
- Задача Лапласа
- Независимость случайных величин
- Числовые характеристики случайных величин

# Повторение

1. Случайной величиной называется функция, которая каждому возможному исходу в эксперименте ставит в соответствие действительное число.

   Можно понимать случайную величину как «кодирование» исходов эксперимента.

   - Дискретные случайные величины имеют конечное или счетное множество значений. Примеры: $\{1, \dots, n\}$, $\mathbb{N}$, $\mathbb{Z}$.
   - Непрерывные случайные величины имеют несчетное множество значений. Примеры: $[0, 1]$, $\mathbb{R}$.

# Повторение

| Дискретные | Непрерывные |
| --- | --- |
| Чтобы задать распределение, необходимо перечислить возможные значения $a_k$ и приписать им вероятности $p_k$, $k \in \mathbb{N}$. | Чтобы задать распределение, необходимо на всех значениях $u \in \mathbb{R}$ ввести функцию плотности $f(u)$. |
| Вероятности $p_k$ — это числа, удовлетворяющие: 1. $p_k \geq 0$ для любого $k \in \mathbb{N}$; 2. $\sum\limits_{k \in \mathbb{N}} p_k = 1$. | Плотность $f(u)$ — это функция, удовлетворяющая: 1. $f(u) \geq 0$ для любого $u \in \mathbb{R}$; 2. $\int_{-\infty}^{+\infty} f(u)du = 1$. |

# Повторение

| Дискретные | Непрерывные |
| --- | --- |
| Как считаем вероятности? $$\mathbb{P}(X \in A) = \sum_{k:\ a_k \in A} p_k.$$ | Как считаем вероятности? $$\mathbb{P}(X \in A) = \int_A f(u)du.$$ |

# Повторение

## 2. Условная вероятность

- Определение:

$$\mathbb{P}(S|H) = \frac{\mathbb{P}(SH)}{\mathbb{P}(H)}, \quad \text{если } \mathbb{P}(H) \neq 0.$$

- Формула полной вероятности:

$$\mathbb{P}(S) = \sum_i \mathbb{P}(S|H_i)\mathbb{P}(H_i).$$

- Формула Байеса:

$$\mathbb{P}(H_k|S) = \frac{\mathbb{P}(S|H_k)\mathbb{P}(H_k)}{\sum_i \mathbb{P}(S|H_i)\mathbb{P}(H_i)}.$$

# Условная вероятность

## Задача

Предположим, что некий тест на какую-нибудь страшную болезнь имеет вероятность успеха 95%; иначе говоря, 5% — это вероятность как ошибки первого рода (ложного срабатывания, false positive), так и ошибки второго рода (пропуска больного человека, false negative). Предположим также, что болезнь очень распространена и имеется у 1% респондентов.

Пусть теперь некий человек получил позитивный результат теста, то есть тест говорит, что страшная болезнь у человека присутствует. С какой вероятностью он действительно болен?

# Условная вероятность

**Решение.** Введем случайные величины:

- $X$ — болен человек или нет;
- $Y$ — тест оказался положительным (зависит от $X$).

Для удобства введем еще следующие события:

- $S = \{\text{тест оказался положительным}\} = \{Y = 1\}$;
- $H_1 = \{\text{человек болен}\} = \{X = 1\}$;
- $H_2 = \{\text{человек не болен}\} = \{X = 0\}$.

# Условная вероятность

По условию задачи:

$$\mathbb{P}(H_1) = 0.01, \quad \mathbb{P}(H_2) = 0.99,$$

$$\mathbb{P}(S|H_1) = 0.95, \quad \mathbb{P}(S|H_2) = 0.05.$$

По формуле Байеса:

$$\mathbb{P}(H_1|S) = \frac{\mathbb{P}(S|H_1)\mathbb{P}(H_1)}{\mathbb{P}(S|H_1)\mathbb{P}(H_1) + \mathbb{P}(S|H_2)\mathbb{P}(H_2)}$$
$$= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.99 \cdot 0.05}$$
$$\approx 0.16.$$

# Задача Лапласа

## Задача (Задача Лапласа)

Пусть солнце всходило $n$ дней подряд. Какова вероятность того, что оно взойдет еще раз?

Предполагается, что вероятность восхода солнца в любой из дней является неизвестной нам постоянной величиной.

Ввиду отсутствия априорных сведений будем считать, что любое из значений внутри отрезка $[0, 1]$ для нее одинаково правдоподобно.

# Задача Лапласа

Введем случайную величину $X$, равномерно распределенную на $[0, 1]$ — вероятность восхода солнца.

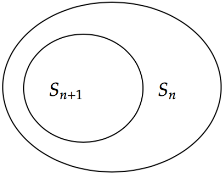
Обозначим $S_n = \{\text{солнце всходило } n \text{ раз подряд}\} = \{Y = n\}$, где $Y \sim \mathbf{B}_X$ (распределение Бернулли с параметром $X$).

В нашей модели: $\mathbb{P}(S_n|X = p) = p^n$.

# Задача Лапласа

Нам необходимо найти $\mathbb{P}(S_{n+1}|S_n)$. По определению:

$$\mathbb{P}(S_{n+1}|S_n) = \frac{\mathbb{P}(S_{n+1} \cdot S_n)}{\mathbb{P}(S_n)} = \frac{\mathbb{P}(S_{n+1})}{\mathbb{P}(S_n)}.$$



# Задача Лапласа

Осталось найти вероятность $\mathbb{P}(S_n)$ для произвольного $n$.

Формула полной вероятности (в дискретном случае) может быть записана так:

$$\mathbb{P}(S_n) = \sum_{0 \le p \le 1} \mathbb{P}(S_n | X = p) \cdot \mathbb{P}(X = p).$$

Для того, чтобы формула стала верной в непрерывном случае, необходимо воспользоваться правилами перехода:

$$\mathbb{P}(S_n) = \int_0^1 \mathbb{P}(S_n | X = p) \cdot f_X(p) dp,$$

где $f_X$ — плотность случайной величины $X$.

# Задача Лапласа

Получим

$$\mathbb{P}(S^n) = \int_0^1 p^n \cdot 1 dp = \left( \frac{p^{n+1}}{n+1} \right) \Big|_0^1 = \frac{1}{n+1}.$$

Подставим ответ в формулу для $\mathbb{P}(S_{n+1}|S_n)$:

$$\mathbb{P}(S_{n+1}|S_n) = \frac{\mathbb{P}(S_{n+1})}{\mathbb{P}(S_n)} = \frac{n+1}{n+2}.$$

# Независимость случайных величин

Перед тем, как мы перейдем к численным характеристикам случайных величин, определим понятие независимости.

Пусть даны две независимые случайные величины $X$ и $Y$,

| $X$ | $a_1$ | $a_2$ | $a_3$ | $\dots$ |
|---|---|---|---|---|
| $\mathbb{P}$ | $p_1$ | $p_2$ | $p_3$ | $\dots$ |

| $Y$ | $b_1$ | $b_2$ | $b_3$ | $\dots$ |
|---|---|---|---|---|
| $\mathbb{P}$ | $q_1$ | $q_2$ | $q_3$ | $\dots$ |

Для произвольных $i$ и $j$ чему должна быть равна вероятность

$$\mathbb{P}(X = a_i | Y = b_j) = \quad \textbf{?}$$

# Независимость случайных величин

Логично было бы определить независимость $X$ и $Y$ так, чтобы для произвольных $i$ и $j$,

$$\mathbb{P}(X = a_i | Y = b_j) = \mathbb{P}(X = a_i).$$

Но пользуясь определением условной вероятности, получим

$$\mathbb{P}(X = a_i | Y = b_j) = \frac{\mathbb{P}(X = a_i, Y = b_j)}{\mathbb{P}(Y = b_j)} \overset{\text{хотим}}{=} \mathbb{P}(X = a_i).$$

То есть достаточно, чтобы

$$\mathbb{P}(X = a_i, Y = b_j) = \mathbb{P}(X = a_i) \cdot \mathbb{P}(Y = b_i).$$

# Независимость случайных величин

Общее определение для дискретного и непрерывного случаев.

Случайные величины $X$ и $Y$ называются независимыми тогда и только тогда, когда для любых (борелевских) подмножеств $A, B \subset \mathbb{R}$ выполняется равенство

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

# Числовые характеристики случайных величин

1. Математическое ожидание.

Рассмотрим сначала дискретный случай.

Для дискретной случайной величины $X$ с таблицей распределения

| $X$ | $a_1$ | $a_2$ | $a_3$ | $\dots$ |
|---|---|---|---|---|
| $\mathbb{P}$ | $p_1$ | $p_2$ | $p_3$ | $\dots$ |

математическим ожиданием (или средним значением) называется число $\mathbb{E}[X]$, которое вычисляется по формуле

$$\mathbb{E}[X] = \sum_i a_i \cdot \mathbb{P}(X = a_i) = \sum_i a_i p_i.$$

# Числовые характеристики случайных величин

1. Математическое ожидание.

   В случае, если ряд

   $$\mathbb{E}[X] = \sum_i a_i \cdot \mathbb{P}(X = a_i) = \sum_i a_i p_i.$$

   не сходится абсолютно, то есть если $\sum |a_i| p_i = \infty$, то говорят, что математическое ожидание не существует.

# Числовые характеристики случайных величин

1. Математическое ожидание.

   Рассмотрим теперь непрерывный случай.

   Для непрерывный случайной величины $X$ с плотностью $f(u)$ математическим ожиданием (или средним значением) называется число $\mathbb{E}[X]$, которое вычисляется по формуле

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} u \cdot f(u) du.$$

# Числовые характеристики случайных величин

1. Математическое ожидание.

   Оба определения согласованы согласно правилам перехода от дискретного случая к непрерывному:

$$\sum_i \longleftrightarrow \int_{-\infty}^{+\infty}, \qquad a_i \longleftrightarrow u \qquad \text{и} \qquad p_i \longleftrightarrow f(u)du.$$

$$\sum_i a_i \cdot p_i \longleftrightarrow \int_{-\infty}^{+\infty} u \cdot f(u)du.$$

# Числовые характеристики случайных величин

1. Математическое ожидание.

**Какой смысл у математического ожидания?**

1) **Физический:** если на невесомом стержне, распределить единичную массу согласно распределению (в дискретном случае поместив в точки $a_i$ массу $p_i$, а в непрерывном — согласно плотности $f$), то точка $\mathbb{E}[X]$ будет координатой «центра тяжести» стержня.

# Числовые характеристики случайных величин

1. Математическое ожидание.

**Какой смысл у математического ожидания?**

2) **Вероятностный:** при многократном повторении эксперимента, ассоциированного с $X$, среднее арифметическое получившихся значений $X_1, X_2, X_3, \ldots$ будет стремиться к $\mathbb{E}[X]$:

$$\frac{X_1 + X_2 + \ldots X_n}{n} \to \mathbb{E}[X] \quad \text{при } n \to \infty.$$

Этот факт называется «Законом больших чисел».

# Числовые характеристики случайных величин

1. Математическое ожидание.

## Задача

Пусть случайная величина $X$ равна числу очков, выпадающих при одном подбрасывании игральной кости. Найти $\mathbb{E}[X]$.

**Решение.** По опеределению

$$\mathbb{E}[X] = \sum_{k=1}^{6} k \cdot \frac{1}{6} = 3.5.$$

В среднем при одном подбрасывании кубика выпадает 3.5 очка. Это несмотря на то, что ни при одном подбрасывании 3.5 очка выпасть не может.

# Числовые характеристики случайных величин

1. Математическое ожидание.

## Задача

Пусть $X$ имеет равномерное распределение на отрезке $[0, 1]$. Найти $\mathbb{E}[X]$.

**Решение.** По опеределению:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} uf(u)du = \int_0^1 udu = \left(\frac{u^2}{2}\right)\Big|_0^1 = \frac{1}{2}.$$

Ничего удивительного: у стержня с массой, равномерно распределенной на $[0, 1]$, центр масс будет находится в $1/2$.

# Числовые характеристики случайных величин

1. Математическое ожидание.

## Задача

Пусть $X \sim \mathcal{N}(0, 1)$. Докажите, что $\mathbb{E}[X] = 0$.

# Числовые характеристики случайных величин

1. Математическое ожидание.

**Решение.** Действительно, для $\mathbb{E}[X]$ верно

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} uf(u)du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} ue^{-\frac{u^2}{2}} du = 0,$$

так как подынтегральная функция является нечетной.

# Числовые характеристики случайных величин

1. Математическое ожидание.

**Свойства математического ожидания:**

(E1) $\mathbb{E}[c] = c$ для любого $c \in \mathbb{R}$;

(E2) $\mathbb{E}[cX] = c\mathbb{E}[X]$ для любого $c \in \mathbb{R}$;

(E3) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ для любых $X$ и $Y$;

(E4) $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$, если $X$ и $Y$ независимы;

(E5) Если $X \geq 0$, то $\mathbb{E}[X] \geq 0$.

# Числовые характеристики случайных величин

1. Математическое ожидание.

## Задача

Пусть $Y \sim \mathbf{B}_{n,p}$ для некоторых $n \in \mathbb{N}$, $p \in [0,1]$. Найти $\mathbb{E}[Y]$.

**Решение.** По определению $Y = X_1 + \ldots + X_n$, где $X_1, X_2, \ldots, X_n \sim B_p$ (независимые). Найдем математическое ожидание каждого слагаемого:

$$\mathbb{E}[X_i] = 0 \cdot (1-p) + 1 \cdot p = p, \qquad i = 1, \ldots, n.$$

По свойствам математического ожидания

$$\mathbb{E}[Y] = \mathbb{E}[X_1 + \ldots + \ X_n] = E[X_1] + \ldots + E[X_n] = np.$$

# Числовые характеристики случайных величин

1. Математическое ожидание.

   Обобщим формулу для математического ожидания.

   Пусть $g : \mathbb{R} \to \mathbb{R}$ — произвольная функция, и нам необходимо посчитать $\mathbb{E}[g(X)]$.

# Числовые характеристики случайных величин

1. Математическое ожидание.

   Рассмотрим сначала дискретный случай. Пусть $X$ имеет таблицу распределения

| $X$ | $a_1$ | $a_2$ | $a_3$ | $\ldots$ |
|---|---|---|---|---|
| $\mathbb{P}$ | $p_1$ | $p_2$ | $p_3$ | $\ldots$ |

   Какое распределение будет у случайной величины $g(X)$?

# Числовые характеристики случайных величин

1. Математическое ожидание.

Распределение у $g(X)$ будет таким:

| $g(X)$ | $g(a_1)$ | $g(a_2)$ | $g(a_3)$ | $\dots$ |
|---|---|---|---|---|
| $\mathbb{P}$ | $p_1$ | $p_2$ | $p_3$ | $\dots$ |

Следовательно, по определению,

$$\mathbb{E}[g(X)] = \sum_i g(a_i) p_i.$$

# Числовые характеристики случайных величин

1. Математическое ожидание.

Пусть теперь $X$ непрерывная величина с плотностью $f(u)$. Чтобы получить аналогичную формулу в этом случае воспользуемся правилами перехода

$$\sum_i \longleftrightarrow \int_{-\infty}^{+\infty}, \qquad a_i \longleftrightarrow u \qquad \text{и} \qquad p_i \longleftrightarrow f(u)du,$$

и получим

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(u)f(u)du.$$

Замечание. Явно выписать плотность для $g(X)$ можно. Мы сделаем это позже.

# Числовые характеристики случайных величин

2. Дисперсия.

Дисперсией случайной величины $X$ называется число $\mathrm{Var}(X)$, которое вычисляется по формуле

$$\mathrm{Var}(X) = \mathbb{E}\Big[(X - \mathbb{E}X)^2\Big].$$

Заметим, что $\mathrm{Var}(X) = \mathbb{E}[g(X)]$, где $g(u) = (u - \mathbb{E}X)^2$.

Это определение и для дискретных, и для непрерывных распределений. Но обратите внимание, что мат. ожидание считается в этих случаях по-разному.

# Числовые характеристики случайных величин

2. Дисперсия.

**Какой смысл у дисперсии?**

Дисперсия $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]$ есть среднее значение квадрата отклонения значения случайной величины $X$ от её мат. ожидания.

То есть дисперсия характеризует степень разброса значений случайной величины вокруг её мат.ожидания.

# Числовые характеристики случайных величин

## 2. Дисперсия.

### Задача

Пусть случайная величина $X$ принимает значения $\pm 1$ с равными вероятностями, а случайная величина $Y$ — значения $\pm 10$ с равными вероятностями. Чему равны $\mathrm{Var}(X)$ и $\mathrm{Var}(Y)$?

**Решение.** Легко заметить, что

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0.$$

Поэтому

$$\mathrm{Var}(X) = \mathbb{E}[X^2] = 1, \quad \mathrm{Var}(Y) = \mathbb{E}[Y^2] = 100.$$

# Числовые характеристики случайных величин

## 2. Дисперсия.

### Предложение

*Дисперсия может быть вычислена по формуле:*

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2.$$

**Доказательство.** Обозначим для удобства $a = \mathbb{E}[X]$. Тогда

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X-a)^2] \\ &= \mathbb{E}[X^2 - 2aX + a^2] \\ &= \mathbb{E}[X^2] - 2a\mathbb{E}[X] + a^2 \\ &= \mathbb{E}[X^2] - a^2.\end{aligned}$$

# Числовые характеристики случайных величин

## 2. Дисперсия.

### Задача

Пусть $X$ имеет равномерное распределение на отрезке $[0, 1]$. Найти $\mathrm{Var}(X)$.

# Числовые характеристики случайных величин

2. Дисперсия.

**Решение.** Используем формулу $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$. Чтобы найти $\mathbb{E}[X^2]$, заметим, что это $\mathbb{E}[g(X)]$ для $g(u) = u^2$, поэтому

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} g(u) f(u) du = \int_0^1 u^2 du = \left( \frac{u^3}{3} \right) \Big|_0^1 = \frac{1}{3}.$$

Вспомнив, что $\mathbb{E}[X] = 1/2$, мы получим

$$\mathrm{Var}(X) = \frac{1}{3} - \left( \frac{1}{2} \right)^2 = \frac{1}{12}.$$

# Числовые характеристики случайных величин

2. Дисперсия.

## Задача

Пусть $X \sim \mathcal{N}(0, 1)$. Докажите, что $\mathrm{Var}(X) = 1$.

# Числовые характеристики случайных величин

2. Дисперсия.

**Решение.** Так как $\mathbb{E}[X] = 0$, то $\mathrm{Var}(X) = \mathbb{E}[X^2]$. Используя интегрирование по частям, получим

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{u^2}{2}}\, du \\
&= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u d\left(e^{-\frac{u^2}{2}}\right) \\
&= -\frac{1}{\sqrt{2\pi}} \left( u e^{-\frac{u^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\, du \right) \\
&= 1.
\end{aligned}$$

# Числовые характеристики случайных величин

2. Дисперсия.

**Свойства дисперсии:**

(V1) $\mathrm{Var}(c) = 0$ для любого $c \in \mathbb{R}$.

(V2) $\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$ для любого $c \in \mathbb{R}$.

(V3) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, если $X$ и $Y$ независимы.

(V4) $\mathrm{Var}(X + c) = \mathrm{Var}(X)$ для любого $c \in \mathbb{R}$.

(V5) $\mathrm{Var}(X) \geq 0$.

# Числовые характеристики случайных величин

## 2. Дисперсия.

### Задача

Пусть $Y \sim \mathbf{B}_{n,p}$ для некоторых $n \in \mathbb{N}$, $p \in [0, 1]$. Найти $\mathrm{Var}(Y)$.

**Решение.** По определению $Y = X_1 + \ldots + X_n$, где $X_1, X_2, \ldots, X_n \sim B_p$ (независимые). Найдем дисперсию каждого слагаемого:

$$\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}X_i)^2 = p - p^2 = p(1-p), \quad i = 1, \ldots, n.$$

По свойствам дисперсии

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1 + .. + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n) = np(1-p).$$

# Числовые характеристики случайных величин

## Задача

Пусть $Y \sim \mathcal{N}(a, \sigma^2)$. Докажите, что $\mathbb{E}Y = a$, $\mathrm{Var}(Y) = \sigma^2$.

# Числовые характеристики случайных величин

**Решение.** Вспомним, что $Y = a + \sigma X$, где $X \sim \mathcal{N}(0, 1)$.

Пользуясь свойствами мат. ожидания и дисперсии, получим

$$\mathbb{E}[Y] = \mathbb{E}[a + \sigma X] = a + \sigma \mathbb{E} X = a;$$
$$\operatorname{Var}(Y) = \operatorname{Var}(a + \sigma X) = \sigma^2 \operatorname{Var}(X) = \sigma^2.$$

# Числовые характеристики случайных величин

Аналогично можно получить результат для равномерного распределения $Y$ на произвольном отрезке $[a, b]$, $a < b$.

Представим $Y$ как

$$Y = a + (b - a)X,$$

где $X$ имеет равномерное распределение на $[0, 1]$. Теперь, пользуясь свойствами мат. ожидания и дисперсии:

$$\mathbb{E}[Y] = \mathbb{E}[a + (b - a)X] = a + (b - a)\mathbb{E}X = \frac{a + b}{2};$$

$$\operatorname{Var}(Y) = \operatorname{Var}(a + (b - a)X) = (b - a)^2 \operatorname{Var}(X) = \frac{(b - a)^2}{12}.$$

# Числовые характеристики случайных величин

**Резюме:**

- если $X \sim \mathbf{B}_p$, то $\mathbb{E}[X] = p$, $\mathrm{Var}(X) = p(1-p)$;
- если $X \sim \mathbf{B}_{n,p}$, то $\mathbb{E}[X] = np$, $\mathrm{Var}(X) = np(1-p)$;
- если $X$ равномерно распределен на $[a, b]$, то $\mathbb{E}[X] = (a+b)/2$, $\mathrm{Var}(X) = (b-a)^2/12$;
- если $X \sim \mathcal{N}(a, \sigma^2)$, то $\mathbb{E}[X] = a$, $\mathrm{Var}(X) = \sigma^2$.

Математическое ожидание и дисперсия являются «базовыми» числовыми характеристиками распределений. Они все известны: для любого распределения их можно найти в любом справочнике или в Википедии.

Спасибо за внимание!