Теория вероятностей и математическая статистика

Дисперсия. Функция распределения. Распределения Бернулли и биномиальное.

Глеб Карпов

ВШБ Бизнес-информатика

Характеристики С.В.: Дисперсия

- Дисперсия случайной величины - это число, **константа**, которое помогает нам понять степень разброса потенциальных значений случайной величины относительно среднего.

Характеристики С.В.: Дисперсия

- Дисперсия случайной величины - это число, **константа**, которое помогает нам понять степень разброса потенциальных значений случайной величины относительно среднего.
- Для случайной величины X её дисперсия обозначается $Var[X]$ (Variance) и определяется как:

$$Var[X] = E\left[(X - E[X])^2\right]$$

Характеристики С.В.: Дисперсия

- Дисперсия случайной величины - это число, **константа**, которое помогает нам понять степень разброса потенциальных значений случайной величины относительно среднего.
- Для случайной величины X её дисперсия обозначается $Var[X]$ (Variance) и определяется как:

$$Var[X] = E\left[(X - E[X])^2\right]$$

- Формально говоря, дисперсия это математическое ожидание новой случайной величины:

$$Y = g(X) = (X - E[X])^2,$$

описывающей квадрат расстояния между значением случайной величины X и её математическим ожиданием.

Характеристики С.В.: Дисперсия

- Дисперсия случайной величины - это число, **константа**, которое помогает нам понять степень разброса потенциальных значений случайной величины относительно среднего.
- Для случайной величины X её дисперсия обозначается $Var[X]$ (Variance) и определяется как:

$$Var[X] = E\left[(X - E[X])^2\right]$$

- Формально говоря, дисперсия это математическое ожидание новой случайной величины:

$$Y = g(X) = (X - E[X])^2,$$

описывающей квадрат расстояния между значением случайной величины X и её математическим ожиданием.

- Также вводится понятие стандартного отклонения (standard deviation, std) величины X, которое является квадратным корнем из дисперсии: $std[X] = \sqrt{Var[X]}$.

Характеристики С.В.: Дисперсия

Вычисление дисперсии

Существует упрощенная формула:

$$E\left[(X-E[X])^2\right] = E\left[X^2 - 2XE[X] + (E[X])^2\right] =$$
$$= E[X^2] - E\left[2E[X]X\right] + E\left[(E[X])^2\right] = E[X^2] - 2E[X]E[X] + (E[X])^2 =$$
$$\boxed{E[X^2] - (E[X])^2}$$

В процессе мы используем свойство линейности математического ожидания, а также знание того, что $E[X]$ - это константа, при этом $E[c] = c$, то есть $E\left[E[X]\right] = E[X]$.

Свойства дисперсии

1. **Неотрицательность**: Для любой случайной величины X:

$$Var[X] \geq 0$$

2. **Дисперсия константы**: Если c - константа, то:

$$Var[c] = 0$$

3. **Масштабирование**: Для любой константы c:

$$Var[cX] = c^2 Var[X]$$

Дисперсия: пример 1

Посчитаем дисперсию на примере нашей задачи. Считаем уже известным, что $E[X] = 1.2$.

x	0	-1	-2	1	2	3
$p_X(x)$	0.1	0.1	0.1	0.2	0.2	0.3

- Нам нужно вычислить $E[Y]$, где $Y = (X - E[X])^2$. Сразу воспользуемся подходом, позволяющим миновать шаг построения таблицы вероятности для Y:

$$E[Y] = \sum_{i=1}^{6} g(x_i) P_X(X = x_i) = \sum_{i=1}^{6} (x_i - 1.2)^2 P_X(X = x_i) =$$

$$= 1.44 \cdot 0.1 + 4.84 \cdot 0.1 + 10.24 \cdot 0.1 + 0.04 \cdot 0.2 + 0.64 \cdot 0.2 + 3.24 \cdot 0.3 = 2.76$$

- Но для дисперсии у нас есть ещё более умный подход, описанный на одном из слайдов выше:

$$Var[X] = E\left[X^2\right] - \left(E\left[X\right]\right)^2 = 4.2 - (1.2)^2 = 2.76$$

$$E\left[X^2\right] = \sum_{i=1}^{6} (x_i)^2 P_X(X = x_i) = 0.1 + 4 \cdot 0.1 + 1 \cdot 0.2 + 4 \cdot 0.2 + 9 \cdot 0.3 = 4.2$$

- Voilà!

Дисперсия: пример 2

Предположим, у нас есть другая случайная величина W и ее функция вероятности задана:

w	-6	0	1	15
$p_W(x)$	0.1	0.5	0.3	0.1

- Посчитаем математическое ожидание:

$$E[W] = -6 \cdot 0.1 + 0 + 1 \cdot 0.3 + 15 \cdot 0.1 = 1.2 = E[X]$$ из предыдущего примера

Дисперсия: пример 2

Предположим, у нас есть другая случайная величина W и ее функция вероятности задана:

w	-6	0	1	15
$p_W(x)$	0.1	0.5	0.3	0.1

- Посчитаем математическое ожидание:

$$E[W] = -6 \cdot 0.1 + 0 + 1 \cdot 0.3 + 15 \cdot 0.1 = 1.2 = E[X]$$ из предыдущего примера

- Посчитаем дисперсию для W:

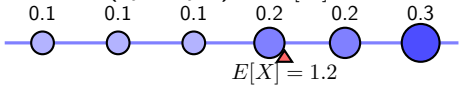
$$Var[W] = E\left[W^2\right] - \left(E\left[W\right]\right)^2 = 26.4 - (1.2)^2 = 24.96$$

$$E\left[W^2\right] = \sum_{i=1}^{4} (w_i)^2 \cdot P_W(W = w_i) = 36 \cdot 0.1 + 0 + 1 \cdot 0.3 + 225 \cdot 0.1 = 26.4$$

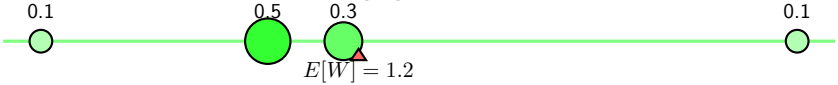
Физическая интерпретация дисперсии

Дисперсия показывает, насколько "разбросаны" значения относительно среднего. Представим случайные величины как массы на стержне:

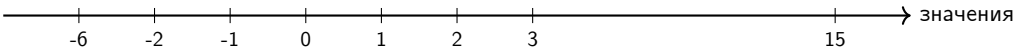
Распределение X (пример 1): $Var[X] = 2.76$



Распределение W (пример 2): $Var[W] = 24.96$



Масштаб не соблюден - для наглядности



Функция распределения (кумулятивная/интегральная)

- **Функция распределения (кумулятивная/интегральная) (CDF)** случайной величины X определяется как:

$$F_X(x) = P(X \leq x), \quad F_X(x) : \mathbb{R} \longrightarrow [0, 1] \subset \mathbb{R}$$

Для дискретной случайной величины:

$$F_X(x) = \sum_{i:\, x_i \leq x} P_X(X = x_i)$$

Функция распределения (кумулятивная/интегральная)

- **Функция распределения (кумулятивная/интегральная) (CDF)** случайной величины X определяется как:

$$F_X(x) = P(X \leq x), \quad F_X(x) : \mathbb{R} \longrightarrow [0, 1] \subset \mathbb{R}$$

Для дискретной случайной величины:

$$F_X(x) = \sum_{i:\, x_i \leq x} P_X(X = x_i)$$

- $F_X(x)$ показывает вероятность того, что случайная величина X примет значение, не превышающее x. Иными словами, сколько вероятности “скопилось”, аккумулировалось, до значения x.

Свойства CDF для дискретных случайных величин

1. **Монотонность**: $F_X(x)$ неубывающая функция
2. **Граничные значения**:
   - $\lim_{x \to -\infty} F_X(x) = 0$
   - $\lim_{x \to +\infty} F_X(x) = 1$
3. **Ступенчатая функция**: CDF дискретной с.в. имеет скачки в точках, где $p_X(x) > 0$

Пример 1: CDF для распределения X

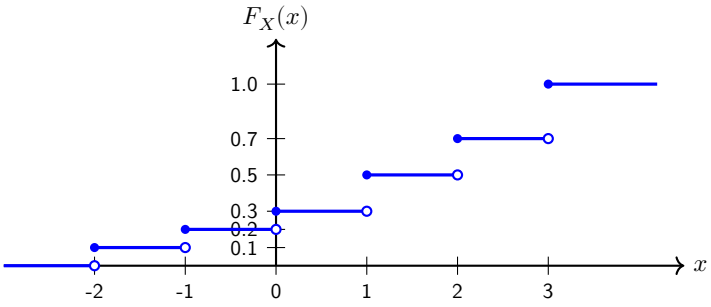
Вспомним наше первое распределение:

x	0	-1	-2	1	2	3
$p_X(x)$	0.1	0.1	0.1	0.2	0.2	0.3

Вычислим CDF:

1. $F_X(x) = 0$ при $x < -2$
2. $F_X(x) = 0.1$ при $-2 \leq x < -1$
3. $F_X(x) = 0.2$ при $-1 \leq x < 0$
4. $F_X(x) = 0.3$ при $0 \leq x < 1$
5. $F_X(x) = 0.5$ при $1 \leq x < 2$
6. $F_X(x) = 0.7$ при $2 \leq x < 3$
7. $F_X(x) = 1.0$ при $x \geq 3$

График CDF для распределения X



Пример 2: CDF для распределения W

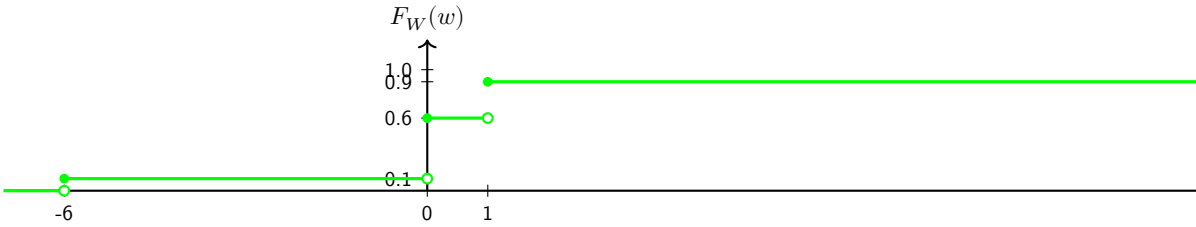
Теперь для второго распределения:

w	-6	0	1	15
$p_W(w)$	0.1	0.5	0.3	0.1

Вычислим CDF:

1. $F_W(w) = 0$ при $w < -6$
2. $F_W(w) = 0.1$ при $-6 \leq w < 0$
3. $F_W(w) = 0.6$ при $0 \leq w < 1$
4. $F_W(w) = 0.9$ при $1 \leq w < 15$
5. $F_W(w) = 1.0$ при $w \geq 15$

График CDF для распределения W



Использование CDF для вычисления вероятностей

CDF позволяет легко вычислять вероятности попадания в интервалы:

Формулы для вычисления вероятностей

1. $P(X \leq a) = F_X(a)$
2. $P(X < a) = F_X(a^-) = \lim_{x \to a^-} F_X(x)$
3. $P(X > a) = 1 - F_X(a)$
4. $P(X \geq a) = 1 - F_X(a^-)$
5. $P(a < X \leq b) = F_X(b) - F_X(a)$
6. $P(a \leq X \leq b) = F_X(b) - F_X(a^-)$

Примеры вычислений с CDF

Для распределения X (на слайдах ранее):

1. $P(X \leq 1) = F_X(1) = 0.5$
2. $P(X > 2) = 1 - F_X(2) = 1 - 0.7 = 0.3$
3. $P(-1 < X \leq 2) = F_X(2) - F_X(-1) = 0.7 - 0.2 = 0.5$
4. $P(0 \leq X < 3) = F_X(3^-) - F_X(0^-) = 0.7 - 0.2 = 0.5$

Распределение Бернулли

Начнем с простейшего случая - **одного случайного эксперимента** с двумя возможными исходами.

Распределение Бернулли - это распределение случайной величины X, которая принимает только два значения:

$$X = \begin{cases} 1 & \text{с вероятностью } p \text{ ("успех")} \\ 0 & \text{с вероятностью } q = 1 - p \text{ ("неудача")} \end{cases}$$

Обозначается как $X \sim Bern(p)$, где $p \in [0, 1]$ - параметр распределения.

Характеристики распределения Бернулли

Основные характеристики

Для случайной величины $X \sim Bern(p)$:

- Математическое ожидание:
$$E[X] = 0 \cdot (1-p) + 1 \cdot p = p$$
- Дисперсия:
$$Var[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$$
- Стандартное отклонение:
$$\sigma = \sqrt{p(1-p)}$$

Примеры распределения Бернулли

1. **Подбрасывание монеты**: "Орел" = 1, "Решка" = 0, $p = 0.5$ (для честной монеты)
2. **Производственный контроль**: "Дефектное изделие" = 1, "Качественное" = 0
3. **Медицинский тест**: "Положительный результат" = 1, "Отрицательный" = 0
4. **Маркетинг**: "Покупка товара" = 1, "Отказ от покупки" = 0

От одного испытания к процессу

- Что если мы хотим провести несколько независимых испытаний Бернулли?
- Например: подбросить монету 10 раз, проверить 100 изделий на дефекты, провести 50 медицинских тестов.
- Это приводит нас к понятию процесса Бернулли - последовательности независимых испытаний Бернулли.

Процесс Бернулли

Процесс Бернулли - это последовательность независимых одинаково распределенных случайных величин $X_1, X_2, \dots, X_n$, где каждая $X_i \sim Bern(p)$:

$$X_i = \begin{cases} 1 & \text{с вероятностью } p \text{ (''успех'')} \\ 0 & \text{с вероятностью } 1-p \text{ (''неудача'')} \end{cases}$$

Ключевое отличие от одного испытания: теперь у нас есть **последовательность** результатов.

Свойства процесса Бернулли

1. **Независимость**: Результат каждого испытания X_i не зависит от результатов других испытаний
2. **Два исхода**: Каждое испытание может закончиться только успехом (1) или неудачей (0)
3. **Постоянная вероятность**: Вероятность успеха p не изменяется от испытания к испытанию

Примеры процесса Бернулли

1. **Серия подбрасываний монеты**: Последовательность $(X_1, X_2, \ldots, X_{10})$ для 10 подбрасываний
2. **Контроль качества**: Проверка партии из n изделий, где каждое может быть дефектным независимо от других
3. **Клинические испытания**: Тестирование лекарства на n пациентах, где каждый может показать положительную реакцию

Биномиальное распределение

- Теперь, когда у нас есть процесс Бернулли, возникает естественный вопрос: **“Сколько успехов мы получим в n испытаниях?”**

Биномиальное распределение описывает количество успехов в n независимых испытаниях Бернулли с вероятностью успеха p. Если $X_1, X_2, ..., X_n$ - процесс Бернулли с параметром p, то случайная величина:

$$Y = X_1 + X_2 + ... + X_n$$

имеет биномиальное распределение и обозначается $Y \sim Bin(n, p)$. Y может принимать значения $0, 1, 2, ..., n$ (от 0 до n успехов).

Функция вероятности биномиального распределения

Вероятность получить ровно k успехов в n испытаниях:

$$P(Y = k) = C_n^k p^k (1-p)^{n-k}$$

Интуиция формулы

- p^k - вероятность k успехов
- $(1-p)^{n-k}$ - вероятность $(n-k)$ неудач
- C_n^k - количество способов выбрать k позиций из n для размещения успехов

Характеристики биномиального распределения

Основные характеристики

Для случайной величины $Y \sim Bin(n, p)$:

Математическое ожидание:

$$E[Y] = np$$

Дисперсия:

$$Var[Y] = np(1-p)$$

Стандартное отклонение:

$$\sigma = \sqrt{np(1-p)}$$

Вывод математического ожидания:

$E[Y] = E[X_1 + X_2 + ... + X_n] = E[X_1] + E[X_2] + ... + E[X_n] = p + p + ... + p = np$

Визуализация биномиального распределения

Форма биномиального распределения зависит от параметров n и p. Рассмотрим, как влияет значение p на распределение при фиксированном $n = 20$:

Сравнение различных значений p

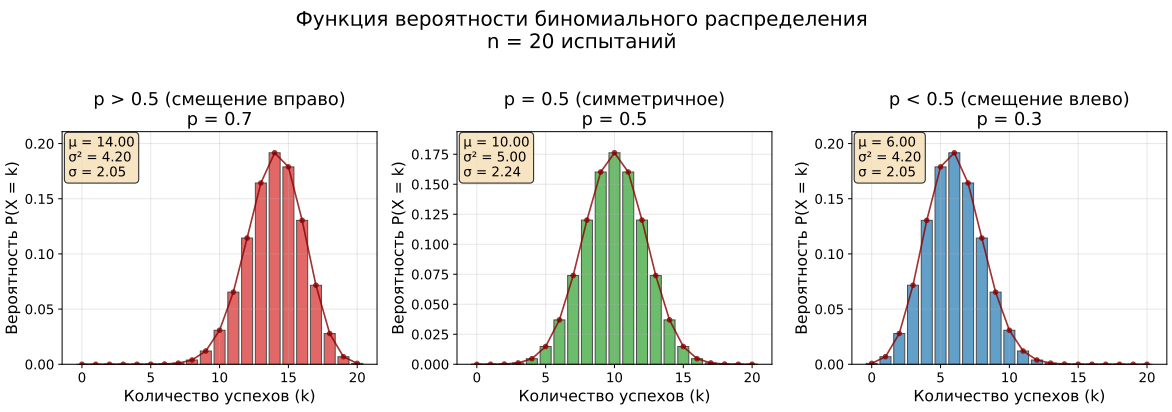


Рис. 1: Сравнение биномиального распределения для разных значений p

Визуализация биномиального распределения

Наблюдения

$p < 0.5$ (например, p = 0.3): Распределение смещено влево, большинство значений сосредоточено в области малых k
$p = 0.5$: Распределение симметрично относительно среднего значения
$p > 0.5$ (например, p = 0.7): Распределение смещено вправо, большинство значений сосредоточено в области больших k

Пример: подбрасывание монеты

Подбросим честную монету 10 раз. Какова вероятность получить ровно 6 орлов?

- $n = 10$, $p = 0.5$, $k = 6$, $Y \sim Bin(10, 0.5)$

$$P(Y = 6) = C_{10}^{6}(0.5)^{6}(0.5)^{4} = \frac{10!}{6! \cdot 4!}(0.5)^{10}$$

$$= \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} \cdot \frac{1}{1024} = 210 \cdot \frac{1}{1024} \approx 0.205$$

Ожидаемое количество орлов: $E[Y] = 10 \cdot 0.5 = 5$

Пример: контроль качества

На заводе производят детали, 3% из которых бракованные. Проверяют партию из 100 деталей.

- $n = 100$, $p = 0.03$ (вероятность брака)

Характеристики:

- $E[Y] = 100 \cdot 0.03 = 3$ (ожидаем 3 бракованные детали)
- $Var[Y] = 100 \cdot 0.03 \cdot 0.97 = 2.91$
- $\sigma = \sqrt{2.91} \approx 1.71$

Вероятность не найти ни одной бракованной детали:

$$P(Y = 0) = C_{100}^{0}(0.03)^{0}(0.97)^{100} = (0.97)^{100} \approx 0.048$$

Пример: контроль качества

На заводе производят детали, 3% из которых бракованные. Проверяют партию из 100 деталей.

- $n = 100$, $p = 0.03$ (вероятность брака)
- $Y \sim Bin(100, 0.03)$ - количество бракованных деталей

Характеристики:

- $E[Y] = 100 \cdot 0.03 = 3$ (ожидаем 3 бракованные детали)
- $Var[Y] = 100 \cdot 0.03 \cdot 0.97 = 2.91$
- $\sigma = \sqrt{2.91} \approx 1.71$

Вероятность не найти ни одной бракованной детали:

$$P(Y = 0) = C_{100}^{0}(0.03)^{0}(0.97)^{100} = (0.97)^{100} \approx 0.048$$

Применения биномиального распределения

1. **Медицина**: Количество пациентов, для которых лекарство окажется эффективным
2. **Маркетинг**: Количество покупателей, совершивших покупку после просмотра рекламы
3. **Социология**: Количество респондентов, давших положительный ответ в опросе
4. **Спорт**: Количество побед команды в серии матчей