

# Теория вероятностей и математическая статистика

## Точечные оценки. Метод моментов. Интервальные оценки I.

Глеб Карпов

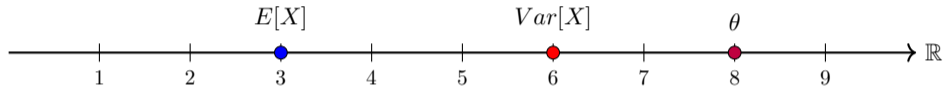
ВШБ Бизнес-информатика

# Повторение

Собираем вместе всё, чего успели коснуться

## Точечная оценка: визуализация

Идеальная ситуация: мы знаем характеристики / параметры



Реальная ситуация: параметры неизвестны

Значения будто скрыты от нас туманом



## Цель точечной оценки

- На основе реализации случайной выборки  $x_1, x_2, \dots, x_n$  получить **предположения**  $\hat{\theta}$  о значениях скрытых в тумане реальности параметров.
- Идея состоит в том, чтобы посчитать значение оценки на реальных имеющихся данных, и чтобы полученное число было бы как можно ближе к истинному значению параметра.
- Следуя аналогии, мы хотим найти затерянные в тумане точки, путём их угадывания специальным способом, с помощью функции **оценки**.

Значения точечных оценок  $\hat{\mu}$ ,  $\hat{\sigma}^2$  и  $\hat{\theta}$   
”попадают” близко к истинным значениям



## Уже известные точечные оценки

Пусть у нас есть случайная выборка  $\mathcal{X} = (X_1, X_2, \dots, X_n)$  (Независимые, одинаково распределенные) с  $\mu \equiv E[X_i]$ ,  $\sigma^2 \equiv Var[X_i]$ .

### 1. Выборочное среднее $\bar{X}$

- Определение:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Характеристики:  $E[\bar{X}] = \mu$  (несмещенная оценка),  $Var(\bar{X}) = \frac{\sigma^2}{n}$
- Распределение:
  - Если  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ , то  $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
  - По ЦПТ: при больших  $n$  выполняется  $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

### 2. Выборочная дисперсия $S^2$

- Определение:  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- Характеристики:  $E[S^2] = \sigma^2$  (несмещенная оценка),  $Var(S^2) = \frac{2\sigma^4}{n-1}$
- Распределение: если  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ , то:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

## Моменты случайной величины

- Момент  $k$ -го порядка случайной величины  $X$  — это математическое ожидание  $k$ -й степени  $X$ :

$$\mu_k = E[X^k]$$

- Первый момент ( $k = 1$ ):  $\mu_1 = E[X]$  — математическое ожидание.
- Второй момент ( $k = 2$ ):  $\mu_2 = E[X^2]$ .
- Центральный момент  $k$ -го порядка:

$$\mu'_k = E[(X - E[X])^k]$$

- Второй центральный момент:  $\mu'_2 = E[(X - E[X])^2] = \text{Var}(X)$  — дисперсия.

# Выборочные моменты

- Для случайной выборки  $X_1, \dots, X_n$  определяем **выборочные моменты**:
- **Выборочный момент  $k$ -го порядка**:

$$m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

- **Выборочный центральный момент  $k$ -го порядка**:

$$m'_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

- Выборочные моменты являются оценками соответствующих теоретических моментов.

# Метод моментов

- **Идея метода моментов:** приравнять выборочные моменты к теоретическим моментам распределения.
- Если распределение зависит от  $p$  параметров, используем первые  $p$  моментов.

**Метод моментов** для оценки параметров:

1. Выразить теоретические моменты через неизвестные параметры.
2. Приравнять выборочные моменты к теоретическим:

$$m_k = \mu_k(\theta_1, \dots, \theta_p), \quad k = 1, 2, \dots, p$$

3. Решить систему уравнений относительно параметров  $\theta_1, \dots, \theta_p$ .

## Пример: Метод моментов для экспоненциального распределения

- Пусть  $X_1, \dots, X_n \sim \text{Exp}(\lambda)$  — время до события (например, время до поломки устройства).
- Нужно оценить параметр  $\lambda$  (интенсивность).
- **Шаг 1:** Теоретический момент первого порядка:

$$\mu_1 = E[X] = \frac{1}{\lambda}$$

- **Шаг 2:** Выборочный момент первого порядка:

$$m_1 = \bar{X}$$

- **Шаг 3:** Приравниваем и выражаем параметр как оценку:

$$\bar{X} = \frac{1}{\lambda} \quad \Rightarrow \quad \hat{\lambda}_{MM} = \frac{1}{\bar{X}}$$

## Пример: Метод моментов для равномерного распределения

- Пусть  $X_1, \dots, X_n \sim \text{Uniform}(0, \theta)$  — равномерное распределение на интервале  $[0, \theta]$ .
- Нужно оценить параметр  $\theta$  (верхнюю границу интервала).
- **Шаг 1:** Теоретический момент первого порядка:

$$\mu_1 = E[X] = \frac{0 + \theta}{2} = \frac{\theta}{2}$$

- **Шаг 2:** Выборочный момент первого порядка:

$$m_1 = \bar{X}$$

- **Шаг 3:** Приравниваем и выражаем параметр как оценку:

$$\bar{X} = \frac{\theta}{2} \quad \Rightarrow \quad \hat{\theta}_{MM} = 2\bar{X}$$

- **Интуиция:** Если среднее значение равно  $\frac{\theta}{2}$ , то верхняя граница  $\theta$  в два раза больше среднего.

## Пример: Метод моментов для нормального распределения

- Пусть  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$  — нужно оценить  $\mu$  и  $\sigma^2$ .
- **Шаг 1:** Теоретические моменты:

$$\mu_1 = E[X] = \mu$$

$$\mu_2 = E[X^2] = \text{Var}(X) + (E[X])^2 = \sigma^2 + \mu^2$$

- **Шаг 2:** Выборочные моменты:

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

## Пример: Метод моментов для нормального распределения

- **Шаг 3:** Приравниваем:

$$\begin{aligned}\bar{X} &= \mu \\ \frac{1}{n} \sum_{i=1}^n X_i^2 &= \sigma^2 + \mu^2\end{aligned}$$

- **Решение:**

$$\begin{aligned}\hat{\mu}_{MM} &= \bar{X} \\ \hat{\sigma}_{MM}^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\end{aligned}$$

- **Замечание:** Оценка дисперсии методом моментов — это смещенная оценка  $S_n^2$ !

## Доверительные интервалы

- Использование только точечной оценки для оценки параметра — это как ловить рыбу в мутном озере гарпунном, а использование доверительного интервала — как ловить сетью. Мы можем бросить гарпун туда, где увидели рыбу, но скорее всего промахнемся. Если мы закинем сеть в эту область, у нас будет больше шансов, что рыбалка будет успешна.
- Если мы делаем точечную оценку, мы, вероятно, не попадем точно в неизвестный параметр. Если мы используем диапазон правдоподобных значений — доверительный интервал — у нас есть хороший шанс "поймать" параметр.
- Действительно, если наша точечная оценка  $\hat{\theta}$  имеет непрерывное распределение, то  $P_{\theta}\{\hat{\theta} = \theta\} = 0$ .

## Доверительные интервалы

**Доверительный интервал.** Пусть  $X_1, X_2, \dots, X_n$  — случайная выборка случайной величины  $X$ . Пусть задано  $0 < \alpha < 1$ . Пусть  $L = L(X_1, X_2, \dots, X_n)$  и  $U = U(X_1, X_2, \dots, X_n)$  — две статистики. Мы говорим, что интервал  $(L, U)$  является  $(1 - \alpha) \cdot 100\%$  доверительным интервалом для неизвестного параметра  $\theta$ , если

$$1 - \alpha = P_{\theta}\{\theta \in (L, U)\}.$$

Вероятность того, что интервал включает  $\theta$ , равна  $1 - \alpha$ , которая называется **уровнем доверия** интервала.

## Доверительные интервалы

### Иллюстративный пример

Для выборки  $X_1, \dots, X_4$  из  $\mathcal{N}(\mu, 1)$  интервальная оценка  $\mu$  — это, например,  $[\bar{X} - 1, \bar{X} + 1]$ . Найдите вероятность того, что истинный параметр  $\mu$  покрывается этим интервалом.

## Решение

$$P(\bar{X} - 1 < \mu < \bar{X} + 1) = P(-1 < \mu - \bar{X} < 1) = P(-1 < \bar{X} - \mu < 1)$$

Знаем, что  $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \equiv \mathcal{N}\left(\mu, \frac{1}{4}\right)$ , приводим к стандартному нормальному распределению:

$$\begin{aligned} P(-1 < \bar{X} - \mu < 1) &= P\left(\frac{-1}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{1}{\frac{\sigma}{\sqrt{n}}}\right) = \\ P\left(\frac{-1}{\frac{1}{2}} < Z < \frac{1}{\frac{1}{2}}\right) &= P(-2 < Z < 2) \approx 0.9545 \end{aligned}$$

## Действие доверительных интервалов

- Некоторые доверительные интервалы включают в себя  $\theta$ , некоторые нет. Доверительный интервал поймает параметр с вероятностью  $1 - \alpha$ .

● поймал ● не поймал



# Доверительные интервалы для среднего генеральной совокупности

Если дисперсия генеральной совокупности известна

- Нам нужно: случайная выборка размера  $n$ , дисперсия  $\sigma^2 \equiv Var[X_i]$  известна *a priori*.
- Утверждение состоит в том, что мы хотим, чтобы наш доверительный интервал  $(L, U)$  покрывал неизвестное среднее генеральной совокупности  $\mu$  с вероятностью  $1 - \alpha$ :

$$1 - \alpha = P(L(X) < \mu < U(X)) = P(-U < -\mu < -L)$$

- Внедряем в центральную часть неравенства известную случайную величину путём одновременного изменения всех частей неравенства:

$$1 - \alpha = P\left(\frac{\bar{X} - U}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - L}{\frac{\sigma}{\sqrt{n}}}\right)$$

- Обычно интервалы хотят делать симметричными, поэтому делаем симметричную замену переменных:

$$1 - \alpha = P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right)$$

## Доверительные интервалы для среднего генеральной совокупности

Если дисперсия генеральной совокупности известна

- После нахождения критической точки  $z_{\alpha/2}$ , такой, что  $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$ , выполняем обратную замену и восстанавливаем нужные границы:

$$z_{\alpha/2} = \frac{\bar{X} - L}{\frac{\sigma}{\sqrt{n}}} \rightarrow L = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
$$-z_{\alpha/2} = \frac{\bar{X} - U}{\frac{\sigma}{\sqrt{n}}} \rightarrow U = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- В итоге получаем теоретический  $(1 - \alpha)100\%$  доверительный интервал для среднего генеральной совокупности  $\mu$ :

$$\mu \in \left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

# Доверительные интервалы для среднего генеральной совокупности

Если дисперсия генеральной совокупности известна

- Однако на практике  $(1 - \alpha)100\%$  доверительный интервал для среднего генеральной совокупности  $\mu$  записывается как:

$$\mu \in \left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

- Отличие лишь в одной букве:  $\bar{x}$  вместо  $\bar{X}$  — конечно, потому что на практике у нас есть именно **реализация** выборочного среднего, те данные, что нам удалось собрать. Вокруг них и строим интервал.