Теория вероятностей и математическая статистика

Точечные оценки.

Глеб Карпов

ВШБ Бизнес-информатика

Напоминание: задачи статистики

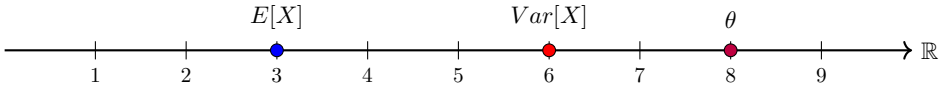
- Случайные процессы вокруг нас — чёрные ящики — случайные генераторы. Мы хотим "расшифровать" один такой ящик и выяснить его параметры и/или свойства, например, математическое ожидание и дисперсию. Для этого мы наблюдаем случайную величину некоторое количество раз и затем делаем выводы на основе накопленной информации.
- 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.

Точечные оценки

- **Оценка** — это специальная статистика, функция от случайной выборки. Её реализация может быть использована в качестве кандидата-значения, **предположения** о значении какого-либо параметра неизвестной случайной величины.
- Пусть θ — неизвестный параметр случайной величины (например, μ или σ^2). Тогда оценка: $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, функция от случайной выборки, тоже является случайной величиной.
- **Реализация оценки** — конкретное числовое значение, полученное из конкретного наблюдения выборки.

Точечная оценка: визуализация

Идеальная ситуация: мы знаем характеристики / параметры



Характеристики / параметры случайной величины X

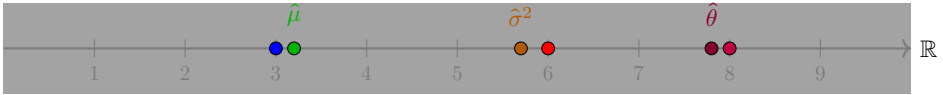
Реальная ситуация: характеристики и параметры неизвестны

Значения будто скрыты от нас туманом

Цель точечной оценки

- На основе реализации случайной выборки $x_1, x_2, \dots, x_n$ получить **предположения** $\hat{\theta}$ для 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 истинным значениям параметров
- 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



Смещенность и несмещенность

Оценка $\hat{\theta}$ называется **несмещенной**, если:

$$E[\hat{\theta}] = \theta$$

Оценка $\hat{\theta}$ называется **смещенной**, если:

$$E[\hat{\theta}] \neq \theta$$

Смещение (bias) определяется как:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Среднеквадратичная ошибка (MSE)

Среднеквадратичная ошибка (Mean Squared Error, MSE) оценки $\hat{\theta}$ определяется как:

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

- MSE измеряет средний квадрат отклонения оценки от истинного значения параметра.
- **Разложение MSE**:
$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$
- MSE учитывает как дисперсию оценки, так и её смещение, и является неким ”мерилом” качества точечной оценки. Чем меньше MSE, тем лучше работает точечная оценка.

Разложение MSE

- Докажем формулу разложения:

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= E[(\hat{\theta}-\theta)^2] \\ &= E[(\hat{\theta}-E[\hat{\theta}]+E[\hat{\theta}]-\theta)^2] \\ &= E[(\hat{\theta}-E[\hat{\theta}])^2]+2E[(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)]+E[(E[\hat{\theta}]-\theta)^2]\end{aligned}$$

- Второе слагаемое равно нулю, так как $E[\hat{\theta}]-\theta$ — константа.
- Получаем:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (E[\hat{\theta}]-\theta)^2 = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2$$

Состоятельность оценки

Оценка $\hat{\theta}$ называется **состоятельной**, если при увеличении размера выборки $n \to \infty$ оценка “сходится” к истинному значению параметра θ.

- **Интуиция**: Чем больше данных мы собираем, тем точнее становится наша оценка, тем ближе ”в тумане” мы находим значения параметров.
- При $n \to \infty$ оценка должна становиться сколь угодно близкой к истинному значению параметра.
- **Важно**: Состоятельность — это асимптотическое свойство (при больших выборках).
- **Связь с несмещённостью**: Несмещённая оценка с дисперсией, стремящейся к нулю при $n \to \infty$, является состоятельной.
- Заметим, что состоятельная оценка может быть смещённой для малых выборок, но смещение должно исчезать при росте n.

Свойства основных оценок

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

Выборочное среднее $\bar{X}$

- Выборочное среднее $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ является оценкой неизвестного математического ожидания μ для исследуемой случайной величины.
- Проверим несмещенность:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot n \cdot \mu = \mu$$

- **Вывод**: $\bar{X}$ — **несмещенная** оценка μ.
- Проверим состоятельность. Дисперсия выборочного среднего:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

- При $n \to \infty$ имеем $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \to 0$.
- **Вывод**: $\bar{X}$ — **состоятельная** оценка μ (несмещённая с дисперсией, стремящейся к нулю).

Свойства основных оценок

Выборочная дисперсия с делением на $n-1$

Выборочная дисперсия $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ является оценкой дисперсии для исследуемой случайной величины.

- Используем факт: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (хи-квадрат с $n-1$ степенями свободы). Для такой хи-квадрат случайной величины:

$$E[\chi^2_{n-1}] = n-1, \quad \mathrm{Var}(\chi^2_{n-1}) = 2(n-1)$$

- Проверим несмещённость:

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1 \Rightarrow E[S^2] = \sigma^2$$

- **Вывод**: S^2 — **несмещенная** оценка для σ^2.
- Проверим состоятельность. Дисперсия S^2:

$$\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1) \Rightarrow \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} \xrightarrow{n\to\infty} 0$$

- **Вывод**: S^2 — **состоятельная** оценка для σ^2 (несмещённая с дисперсией, стремящейся к нулю).

Свойства основных оценок

Выборочная дисперсия с делением на n

Рассмотрим другой вариант оценки дисперсии: $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

- Связь с несмещенной S^2:

$$S_n^2 = \frac{n-1}{n}S^2 = \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S_n^2$$

- Используя $E[S^2] = \sigma^2$, получаем:

$$E[S_n^2] = E\left[\frac{n-1}{n}S^2\right] = \frac{n-1}{n}E[S^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- **Вывод**: S_n^2 — **смещенная** оценка σ^2.
- Вычислим смещение:

$$\text{Bias}(S_n^2) = E[S_n^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{n-1-n}{n}\sigma^2 = -\frac{\sigma^2}{n}$$

Свойства основных оценок

Выборочная дисперсия с делением на n

- Проверим состоятельность. Дисперсия S_n^2:

$$\mathrm{Var}(S_n^2) = \mathrm{Var}\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(S^2) = \left(\frac{n-1}{n}\right)^2 \cdot \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$$

- При $n \to \infty$ смещение стремится к нулю:

$$\lim_{n\to\infty} \mathrm{Bias}(S_n^2) = \lim_{n\to\infty}\left(-\frac{\sigma^2}{n}\right) = 0$$

- И дисперсия стремится к нулю:

$$\lim_{n\to\infty} \mathrm{Var}(S_n^2) = \lim_{n\to\infty} \frac{2(n-1)\sigma^4}{n^2} = 0$$

- **Вывод**: S_n^2 — **состоятельная** оценка σ^2 (смещение стремится к нулю при росте выборки).