

Семинар №3

Статистики и оценки.

Пусть есть вероятностно-статистическая модель $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{P})$.

Определение. Измеримое отображение $S : (\mathcal{X}, \mathcal{B}_\mathcal{X}) \rightarrow (\mathcal{Y}, \mathcal{E})$ называется *статистикой*. Если $\mathcal{Y} = \Theta$ (т.е. $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ и мы имеем дело с параметрической моделью), то S называется оценкой параметра θ .

Грубо говоря, если у нас есть выборка $X = (X_1, \dots, X_n)$, то статистикой $S(X)$ называется измеримая функция от выборки, а оценкой – такая статистика, значением которой является параметр распределения или функция от параметра (в случае, когда семейство возможных распределений \mathcal{P} параметризовано, т.е. представимо в виде $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$).

Примеры статистик.

1) Выборочные характеристики.

Пусть $X = (X_1, \dots, X_n)$ – выборка, а $g(x)$ – борелевская функция.

Определение. Величина $\overline{g(X)} = \frac{1}{n} \sum_{i=1}^n g(X_i)$ называется *выборочной характеристикой* $g(x)$.

Например, $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$ называется *выборочным средним*, а $\overline{X^k} = \frac{1}{n} \sum_{i=1}^n X_i^k$ – *выборочным k -тым моментом*.

Эти оценки хороши тем, что они являются несмещёнными и состоятельными оценками математического ожидания и k -тых моментов соответственно (если они есть для рассматриваемого семейства распределений). Несмещённость тривиально следует из линейности математического ожидания, а состоятельность – из закона больших чисел. Кроме того, эти оценки являются асимптотически нормальными (все эти определения читайте дальше в листочке).

2) Функции от выборочных характеристик.

$S(X) = h(g_1(X), \dots, g_k(X))$, где h – борелевская функция в \mathbb{R}^k .

Примером может служить *выборочная дисперсия* $s^2 = \overline{X^2} - (\overline{X})^2$.

3) Порядковые статистики.

Упорядочим значения выборки (X_1, \dots, X_n) по возрастанию: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. Полученная совокупность статистик называется *вариационным рядом*, а его член $X_{(k)}$ – *k -той порядковой статистикой*.

Свойства оценок.

Пусть X – наблюдение из неизвестного распределения $P = \{P_\theta, \theta \in \Theta\}$ (т.е., как правило, либо одна случайная величина, либо выборка).

Определение. Оценка $\theta^*(X)$ называется *несмещённой* оценкой параметра $\tau(\theta)$, если $\forall \theta \in \Theta \quad E_\theta \theta^*(X) = \tau(\theta)$.

Пример. Допустим, у нас есть выборка, состоящая из одной случайной величины X_1 , из неизвестного распределения P_X , причём мы знаем, что P_X может быть одним из трёх следующих распределений: $N(0, 1)$, $N(2, 1)$ и $N(-35, 1)$. Таким образом, в данном случае семейство распределений $\mathcal{P} = \{N(\theta, 1), \theta \in \Theta\}$, где $\Theta = \{-35, 0, 2\}$. Свойство несмещённости статистики $\theta^*(X_1)$ (в дальнейшем будем писать $\theta^*(X)$) означает следующее: если наша выборка сделана из распределения $N(0, 1)$, то $E_\theta \theta^*(X)$ есть $\int_{\mathbb{R}} \theta^*(x) p_{N(0,1)}(x) dx$ и должно

равняться $\tau(0)$, где $p_{N(0,1)}(x)$ – плотность стандартного нормального закона. Если же наша выборка сделана из распределения $N(2, 1)$, то $E_{\theta}\theta^*(X) = \int_{\mathbb{R}} \theta^*(x)p_{N(2,1)}(x)dx = \tau(2)$, то же самое должно выполняться для $\theta = -35$.

В общем случае мы, как правило, знаем семейство распределений \mathcal{P} , например, $\mathcal{P} = \{R[-\theta, \theta], \theta > 0\}$, тогда надо найти распределение статистики $\theta^*(X)$ и посчитать её матожидание как интеграл, зависящий от неизвестного параметра θ , который будет фигурировать в распределении. Для объяснения смысла последующих определений можно проводить аналогичные рассуждения. \square

Пусть $X = (X_1, \dots, X_n)$ – выборка.

Определение. Оценка $\theta^*(X)$ (точнее, последовательность оценок) называется *состоятельной* оценкой параметра $\tau(\theta)$, если $\forall \theta \in \Theta$ выполнено $\theta_n^* \xrightarrow{P_{\theta}} \tau(\theta)$ при $n \rightarrow \infty$.

Определение. Оценка $\theta^*(X)$ называется *сильно состоятельной* оценкой $\tau(\theta)$, если $\forall \theta \in \Theta$ выполнено $\theta_n^* \rightarrow \tau(\theta)$ P_{θ} -п.н.

Определение. Оценка $\theta^*(X)$ называется *асимптотически нормальной* оценкой $\tau(\theta)$, если $\forall \theta \in \Theta$

$$\sqrt{n}(\theta^*(X) - \tau(\theta)) \xrightarrow{d_{\theta}} N(0, \sigma^2(\theta)),$$

где $\sigma^2(\theta)$ называется *асимптотической дисперсией* оценки $\theta^*(X)$.

Основными инструментами доказательства того, что оценка является состоятельной или асимптотически нормальной, являются законы больших чисел и предельные теоремы, а также теорема о наследовании сходимостей.

Лемма 1. (О наследовании асимптотической нормальности)

Пусть $\theta^*(X)$ – асимптотически нормальная оценка θ с асимптотической дисперсией $\sigma^2(\theta)$, а $\tau(\theta)$ – дифференцируемая функция на $\Theta \subset \mathbb{R}$. Тогда $\tau(\theta^*(X))$ – асимптотически нормальная оценка $\tau(\theta)$ с асимптотической дисперсией $\sigma^2(\theta)(\tau'(\theta))^2$.

Задача. Пусть задана выборка $X = (X_1, \dots, X_n)$ из неизвестного распределения, принадлежащего параметрическому семейству распределений $\{P_{\theta}, \theta \in \Theta\}$. Пусть $E_{\theta}|X_1|^k < +\infty$. Доказать, что выборочный k -тый момент $\overline{X^k}$ является несмещённой, состоятельной и асимптотически нормальной оценкой EX_1^k .

Решение.

На самом деле, условие, что распределение, из которого сделана выборка, принадлежит параметрическому семейству распределений, является необязательным для этой задачи. Я просто хотел, чтобы вы привыкали к обозначениям математической статистики.

Итак, докажем несмещённость. $E_{\theta}\overline{X^k} = E_{\theta}\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) = \frac{1}{n} \sum_{i=1}^n E_{\theta}X_i^k = \frac{1}{n} \sum_{i=1}^n E_{\theta}X_1^k = E_{\theta}X_1^k$. Несмещённость доказана.

Теперь состоятельность. По закону больших чисел, $\overline{X^k} = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P_{\theta}} E_{\theta}X_1^k$, что нам и требовалось (т.е. рассуждения как проводятся: предположим, что выборка у нас сделана из распределения P_{θ} , тогда $\overline{X^k}$ по вероятности, взятой с помощью распределения P_{θ} , сходится к $E_{\theta}X_1^k$ по закону больших чисел).

Наконец, асимптотическая нормальность. Из центральной предельной теоремы, $\sqrt{n}(\overline{X^k} - E_{\theta}X_1^k) \xrightarrow{d_{\theta}} N(0, DX_1^k)$, что и требовалось. \blacksquare