Основные задачи и методы теории статистических выводов

Пусть задана выборка $X_1, \ldots, X_n \sim F$

Необходимо сделать выводы о распределении F

а) Параметрические и непараметрические модели

1

Статистической моделью $\mathfrak{F}$ **называется множество распределений (плотностей, регрессионных зависимостей и т.п.). Параметрическая модель** $\mathfrak{F}$ **может быть параметризована конечным числом параметров**

$$\mathfrak{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad \mu \in \mathbb{R},\ \sigma > 0 \right\}$$

Общий вид параметрической модели:

$$\mathfrak{F} = \left\{ f(x;\theta) : \ \theta \in \Theta \right\}$$

Компоненты θ, не представляющие интереса для исследователя, называют мешающими параметрами

Непараметрическая модель – множество $\mathfrak{F}$, которое нельзя параметризовать конечным числом параметров

Например, $\mathfrak{F}$ - это множество всех кумулятивных функций распределения

<u>Пример</u> (оценка одномерного параметра). $X_1, \ldots, X_n$ - i.i.d., имеющие распределение Бернулли с параметром p

<u>Пример</u> (оценка двумерного параметра). $X_1, \ldots, X_n$ **- i.i.d., имеющие нормальное распределение с параметрами** μ **и** σ

<u>Пример</u> (непараметрическое оценивание). $X_1, \ldots, X_n$ **- i.i.d. с распределением** $F \in \mathfrak{F}_{\mathsf{ALL}}$ **(класс всех функций распределения)**

<u>Пример</u> (непараметрическая оценка плотности распределения).

$X_1, \ldots, X_n$ **- i.i.d. с распределением** F **и плотностью** $f = F'$**. Нельзя оценить** f**, предполагая только, что** $F \in \mathfrak{F}_{\mathsf{ALL}}$**. Необходимы дополнительные предположения о гладкости** f**, например, что**

$$f \in \mathfrak{F} = \mathfrak{F}_{\mathsf{DENS}} \bigcap \mathfrak{F}_{\mathsf{SOB}}$$

$\mathfrak{F}_{\mathsf{DENS}}$ **- множество всех плотностей распределения**

$$\mathfrak{F}_{\mathsf{SOB}} = \left\{ f : \int (f''(x))^2 dx < \infty \right\}$$

- пространство Соболева

<u>Пример</u> (непараметрическое оценивание функционалов). $X_1, \ldots, X_n$ **- i.i.d. с распределением** F**. Необходимо оценить**

$$\mu = \mathbb{E}(X_1) = \int x \, dF(x)$$

предполагая только, что μ **существует. Будем использовать статистический функционал**

$$\mu = T(F) = \int x \, dF(x),$$

заменив F **на его оценку**

<u>Пример</u> (регрессия, предсказание и классификация). По наблюдениям $(X_1, Y_1), \ldots (X_n, Y_n)$ **необходимо восстановить функцию регрессии** $r(x) = \mathbb{E}(Y|X = x)$**. Класс** $r \in \mathfrak{F}$ **может быть как параметрическим, так и непараметрическим. Часто используют следующую запись модели:**

$$Y = r(X) + \epsilon, \mathbb{E}(\epsilon) = 0$$

$\epsilon = Y - r(X)$, $Y = Y + r(X) - r(X) = r(X) + \epsilon$

$$\mathbb{E}(\epsilon) = \mathbb{E}\mathbb{E}(\epsilon|X) = \mathbb{E}(\mathbb{E}(Y - r(X))|X) = \mathbb{E}(\mathbb{E}(Y|X) - r(X)) = \mathbb{E}(r(X) - r(X)) = 0$$

Существует два подхода к статистическим выводам: частотный и байесовский

5

Основные обозначения:

$$\mathfrak{F} = \{f(x;\,\theta):\ \theta \in \Theta\}$$

$$\mathbb{P}_\theta(X \in A) = \int_A f(x;\,\theta)dx$$

$$\mathbb{E}_\theta(r(X)) = \int r(x)f(x;\,\theta)dx$$

$$\mathbb{V}_\theta$$

b) Основные задачи: точечное оценивание, доверительные множества, тестирование гипотез

<u>Точечное оценивание</u> – оценка неизвестной величины (параметры в параметрической модели, распределение, плотность распределения, функций регрессии, предсказание будущего значений некоторой случайной величины) в некотором наилучшем смысле

Будем обозначать оценка параметра θ **через** $\widehat{\theta}$ **или** $\widehat{\theta}_n$**.** $\widehat{\theta}$ **- зависит от данных, то есть это случайная величина.**

$X_1, \ldots, X_n$ **- i.i.d. наблюдения**

$$\widehat{\theta}_n = g(X_1, \ldots, X_n)$$

$\mathsf{bias}(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta$ **- смещение оценки**

Оценка $\widehat{\theta}_n$ **несмещенная, если** $\mathbb{E}(\widehat{\theta}_n) = \theta$

Оценка $\widehat{\theta}_n$ **состоятельная, если** $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$

$\mathsf{se} = \mathsf{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)}$ **- стандартное отклонение оценки или среднеквадратическая ошибка (зависит от неизвестного распределения)**

$\widehat{\mathsf{se}}$ **- оценка стандартного отклонения**

<u>Пример</u> (распределение Бернулли). $X_1, \ldots, X_n$ **- i.i.d., имеющие распределение Бернулли с параметром** p

$$\widehat{p}_n = n^{-1} \sum_i X_i$$

$$\mathbb{E}(\widehat{p}_n) = n^{-1} \sum_i \mathbb{E}(X_i) = p$$

$$\mathsf{se} = \sqrt{\mathbb{V}(\widehat{p}_n)} = \sqrt{p(1-p)/n}$$

$$\widehat{\mathsf{se}} = \sqrt{\widehat{p}(1-\widehat{p})/n}$$

Среднеквадратическая ошибка оценки (MSE)

$$\mathrm{MSE} = \mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2,$$

где $\mathbb{E}_\theta(\cdot)$ **- математическое ожидание относительно плотности выборки**

$$f(x_1, \ldots, x_n;\, \theta) = \prod_{i=1}^{n} f(x_i;\, \theta)$$

8 **<u>Утверждение</u> (bias-variance decomposition):**

$$\mathrm{MSE} = \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}_\theta(\widehat{\theta}_n)$$

Пусть $\overline{\theta}_n = E_\theta(\widehat{\theta}_n)$**, тогда так как** $\mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n) = \overline{\theta}_n - \overline{\theta}_n = 0$**, то**

$$\begin{aligned}
\mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2 &= \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2 \\
&= \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 + 2(\overline{\theta}_n - \theta)\mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n) + \mathbb{E}_\theta(\overline{\theta}_n - \theta)^2 \\
&= (\overline{\theta}_n - \theta)^2 + \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 \\
&= \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}(\widehat{\theta}_n)
\end{aligned}$$

<u>Утверждение</u>: Если $\text{bias} \to 0$ **и** $\text{se} \to 0$**, то** $\widehat{\theta}_n$ **- состоятельная оценка**

Оценка $\widehat{\theta}_n$ **асимптотически нормальна, если**

$$\frac{\widehat{\theta}_n - \theta}{\text{se}} \rightsquigarrow N(0,1)$$

<u>Доверительные множества</u>

Доверительным интервалом с доверительной вероятностью $1-\alpha$ **для параметра** θ **называется интервал** $C_n = (a, b)$**, где** $a = a(X_1, \ldots, X_n)$ **и** $b = b(X_1, \ldots, X_n)$ **- такие функции выборки, что** $\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha$ **для всех** $\theta \in \Theta$

Замечание. $C_n = (a, b)$ **- случайная величина, тогда как** θ **- неизвестная детерминированная величина**

Замечание. Если θ **- векторный параметр, то** C_n **называется доверительным множеством**

В заданном определении доверительного интервала предполагается, что

10 $\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ **для всех** $\theta \in \Theta$

Поточечным асимптотическим доверительным интервалом будем называть такой доверительный интервал, что

$\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ **для всех** $\theta \in \Theta$

Равномерным асимптотическим доверительным интервалом будем называть такой доверительный интервал, что

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$$

Доверительный интервал не является вероятностным утверждением о θ, поскольку θ - не случайная величина, а детерминированный неизвестный параметр. Интерпретация может быть следующей – либо повтор одного и того же эксперимента, либо построение большого количество доверительных интервалов – в первом случае $1 - \alpha$ процентов времени неизвестный параметр попадет в интервал, во втором случае доля $1 - \alpha$ построенных доверительных интервалов накроет неизвестный параметр.

<u>Пример</u>. $X_1, \ldots, X_n$ **- i.i.d., имеющие распределение Бернулли с параметром** p**.** $\widehat{p}_n = n^{-1} \sum_i X_i$

$$C_n = (\widehat{p}_n - \epsilon_n,\ \widehat{p}_n + \epsilon_n)$$

доверительный интервал с $\epsilon_n^2 = \log(2/\alpha)/(2n)$**.**

12 **<u>Неравенство Хеффдинга</u>:** $Y_1, \ldots, Y_n$ **- i.i.d., при этом** $\mathbb{E}(Y_i) = 0$ **и с вероятностью единица** $a_i \le Y_i \le b_i$**. Тогда для любого** $t > 0$ **и** $\epsilon > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \ge \epsilon\right) \le e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2 (b_i - a_i)^2/8}$$

Отсюда получаем, что $\mathbb{P}(p \in C_n) \ge 1 - \alpha$

<u>Утверждение</u>. Допустим, что $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$ **(оценка асимптотически нормальная). Пусть** Φ **- стандартная нормальная функция распределения,** $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$**, то есть** $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ **и**

$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$**, где** $Z \sim N(0,1)$**,**

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}).$$

Тогда

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha.$$

Действительно, положим $Z_n = (\widehat{\theta}_n - \theta)/\widehat{\mathsf{se}}$**. Тогда согласно предположению** $Z_n \rightsquigarrow Z$**, где** $Z \sim N(0,1)$**.**

$$
\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}} < \theta < \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} < z_{\alpha/2}\right) \\
&\to \mathbb{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= 1 - \alpha
\end{aligned}
$$

Доверительный интервал такого вида является поточечным асимптотическим доверительным интервалом

<u>Пример</u>: Пусть $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$ **,** $\widehat{p}_n = n^{-1} \sum_{i=1}^n X_i$

$$\mathbb{V}(\widehat{p}_n) = n^{-2} \sum_{i=1}^n \mathbb{V}(X_i) = n^{-2} \sum_{i=1}^n p(1-p) = n^{-2} np(1-p)$$

$$= p(1-p)/n$$

$$\mathsf{se} = \sqrt{p(1-p)/n}$$

$$\widehat{\mathsf{se}} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$$

Согласно центральной предельной теореме $\widehat{p}_n \approx N(p, \widehat{\mathsf{se}}^2)$

Тогда приблизительный доверительный интервал с доверительной вероятностью $1-\alpha$ **имеет вид**

$$\widehat{p}_n \pm z_{\alpha/2} \widehat{\mathsf{se}} = \widehat{p}_n \pm z_{\alpha/2} \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

<u>Тестирование гипотез</u>

Делается предположение о процессе, генерирующем данные и задача состоит в том, чтобы определить, содержат ли данные достаточно информации, чтобы отвергнуть это предположение. Если информации не достаточно, то говорится, что опытные данные предположению (гипотезе) не противоречат.

<u>Пример</u>. Пусть $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$

$$H_0 : p = 1/2$$ **(основная гипотеза)**

$$H_1 : p \neq 1/2$$ **(альтернативная гипотеза)**

Рассмотрим статистику $T = |\widehat{p}_n - (1/2)|$**. Если она достаточно большая (пороговое значение будет определено позже), то основная гипотеза отклоняется**