Этакое большое ничего и матстат

Белкин Дмитрий, U-1152
Бертыш Вадим, СПБГЭТУ «ЛЭТИ» 4373

15 июня 2016

Основные определения

Определение 1 (Статистический эксперимент)**.** *Тройка* $(\mathfrak{X}, \mathfrak{F}, \mathcal{P})$ *называется статистическим экспериментом*

- $\mathfrak{X}$ - *Множество результатов эксперимента*
- $\mathfrak{F}$ - *Совокупность наблюдаемых событий*
- $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ - *Семейство вероятностных распределений*

Дальше положим $\mathfrak{X} = \mathbb{R}^n$, $\mathfrak{F} = \sigma(\mathfrak{F}_1 \times \cdots \times \mathfrak{F}_n) = \mathfrak{B}_n$

Определение 2 (Статистика)**.** *Измеримая функция* $T : \mathfrak{X} \to E$ *называется статистикой*

Определение 3 (Подчиненная статистика)**.** *Статистика* T *называется подчиненной, если её распределение не зависит от параметра*

$$P_\theta(T \in A) = P_T(A)$$

Определение 4 (Достаточная статистика)**.** *Статистика* T *назвается достаточной, если условное распределение* X *при условии* T *не зависит от параметра*

$$P_\theta(X \in A|T) = P_{X|T}(A), \forall\theta \in \Theta$$

Подчиненная не содержит информации о параметре, достаточная содержит всю информацию о параметре

Определение 5 (Минимальная достаточная статистика)**.** *Достаточная статистика* T *называется минимальной, если,* $\forall T_1$ *достаточной* $\exists g : T = g(T_1)$

Использование МДС максимально редуцирует имеющиеся данные

Основные типы задач статистики

- Точечное оценивание (статистики $\delta : \mathfrak{X} \to \Theta$)
- Доверительное оценивание с уровнем доверия $1 - \alpha$ ($\mathcal{Y}$ - семейство подмножеств Θ)

  $$\Delta : \mathfrak{X} \to \mathcal{Y}$$

  такие, что $P_\theta(\theta \in \Delta(\vec{X})) \geq 1 - \alpha, \forall\theta \in \Theta$

- Проверка гипотез (принятие решений)
  $H : \theta \in \Theta_*, \Theta_* \subset \Theta$ - Гипотеза. Выдвигают $H_0 : \theta \in \Theta_0$ и $H_A : \theta \in \Theta$
  Решающее правило - критерий

  $$\phi : \mathfrak{X} \to [0; 1]$$

  $\phi(\vec{X})$ - вероятность выбрать альтернативу (отвергнуть H_0)

Асимптотический подход Пусть $(\mathfrak{X}^{(n)}, \mathfrak{F}^{(n)}, \mathcal{P}^{(n)})$ последовательность статистических экспериментов $\mathcal{P}^{(n)} = \{p_\theta^{(n)}, \theta \in \Theta\}$

Определение 6 (Состоятельность оценки)**.** *Точечная оценка $\delta^{(n)}(\vec{X})$ называется состоятельной, если*

$$\delta^{(n)}(\vec{X}) \xrightarrow{p_\theta} \theta, \forall \theta \in \Theta$$

Определение 7 (Сильная состоятельность оценки)**.** *Точечная оценка $\delta^{(n)}(\vec{X})$ называется сильно состоятельной, если*

$$\delta^{(n)}(\vec{X}) \xrightarrow[n\to\infty]{p_\theta = 1} \theta, \forall \theta \in \Theta$$

Определение 8 (Асимптотическая нормальность)**.** *Точечная оценка $\delta^{(n)}(\vec{X})$ называется асимптотически нормальной, если*

$$\sqrt{n}(\delta^{(n)}(\vec{X}) - \theta) \underset{P_\theta}{\rightrightarrows} \mathcal{N}(0, \sigma^2(\theta))$$

Методы накопления статистической информации

- Выборочный метод

  Определение 9 (Выборка)**.** *набор НОРСВ $\vec{X} = (x_1, \ldots, x_n)$ наызывается выборкой*

  Совместное распределение задается распрелелением x_1, Θ не меняется с ростом n

- Группировка
  Разбиваем наблюдения на k групп. Наблюдаемые величины: $(x_1, z_1), \ldots, (x_n, z_n) \in \mathfrak{X}$, где x_i - наблюдаемое значение, а $z_i \in \{1, \ldots, k\{$ - принадлежность его к какой-либо группе. Распределения в пределах группы совпадают. Совместное распределение определяется распределением при каждом z_i - F_s, $s = 1, \ldots, k$

- Регрессия

  Определение 10 (Регрессия)**.** *Регрессией величины Y по X называют* $\mathrm{E}[Y|X] = f(x)$

  Модель основана на соотношении

  $$\mathrm{E}_\theta[Y, X(z)] = g_\theta(X(z)^T)$$

  Распределение Y характеризуется F_z

Параметризация

$$\mathcal{P} = \{p_\theta, \theta \in \Theta\}$$

По типу множества Θ можно выделить

- Параметрические $\Theta \subset \mathbb{R}^d$
- Семипараметрическими $\Theta \subset \Theta_1 \times \Theta_2, \Theta_1 \subset \mathbb{R}^d$
- Непараметрические (остальные)

Непараметрическое оценивание (Выборочный подход) Пусть $X_1, \dots X_n$ - выборка из распределения $P_\theta, \theta \in \Theta$, истиное значение P_θ будем называеть *теоретическим распределением.*

Определение 11 (Выборочная функция распределения). *Выборочной функцией распределения называют*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i < x\}} = \frac{\textit{Число наблюдений меньших } x}{\textit{общее число наблюдений}}$$

Теорема 1 (Гливенко-Кантелли).

$$\sup_x \{|F_n(x) - F(x)|\} \xrightarrow[n\to\infty]{P_\theta = 1} 0, \forall F$$

Утверждает сильную состочтельность F_n

Теорема 2 (Колмогорова).

$$\sqrt{n} \sup_x \{F_n(x) - F(x)\} \underset{F}{\rightrightarrows} \mathcal{K}$$

Где $\mathcal{K}$ - распределение Колмогорова, функция распределения $\mathcal{K}(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}$ Даёт возможность строить доверительные области для функции распределения. Пусть α - малое число, x_α - квантиль уровня $1 - \alpha$, тогда

$$1-\alpha \approx P_F(\sqrt(n) \sup_x |F_n(x) - F(x)| < \alpha) = P_F(F_n(x) - \frac{x_\alpha}{\sqrt{n}} \le F(x) \le F_n(x) + \frac{x_\alpha}{\sqrt{n}})$$

Идея доказательства - преобразование Смирнова

Определение 12 (Преобразование Смирнова). *x с непренывной функцией распределения F, $Y = F(x)$, тогда*

$$Y \sim \mathcal{U}(0,1)$$

Определение 13 (Порядковые статистики, вариационный ряд и ранги). *Достаточная статистика*

$$X_{(1)} \le X_{(2)} \le \dots X_{(n)}$$

называется вариационным рядом, $X_(k)$ - k-ая порядковая статистика. Ранг R_k - номер x_k в вариационном ряду

Выборочные характеристики

Определение 14 (Выборочные характеристики). *Пусть $\mathscr{F}$ - подмножество множество распределений $\alpha : \mathscr{F} \to \mathbb{R}$ - числовая характеристика. Тогда*

$\alpha(F)$ - *теоретическая характеристика*

$\alpha(F_n)$ - *выборочная характеристика*

Числовые характеристики можно разделить на две вида:

- Характеристики $H(\mathrm{E}g_1(X), \ldots, \mathrm{E}g_n(X))$
- Непрерывный в равномерной метрике функционал $G(F)$

К первым относятся моментные характеристики, ко вторым квантили.

Теорема 3. *Пусть $X_1, \ldots, X_n$ - Выборка из распределения с функцией распределения F, числовая характеристика $G(F)$ первого или второго типа существует, тогда с вероятностью 1*

$$G(F_n) \underset{n\to\infty}{\to} G(F)$$

Выборочные квантили

Определение 15. *ξ_p - квантиль порядка p, $\xi_p = x_{([np]+1)}$, если $np \in \mathbb{Z}$*

Параметрическое оценивание

Определение 16 (Точечная оценка)**.** *Статистика $\delta(\vec{X})$, $\delta : \mathfrak{X} \to \Theta$ называется точечной оценкой*

Определение 17 (Функция потерь)**.** *пусть θ реально значение параметра, тогда $W(\delta(\vec{X}), \theta)$ функция потерь, если*

- $W(\delta(\vec{X}), \theta) > 0, \forall \vec{X} \in \mathfrak{X}$
- $W(\theta, \theta) = 0$

Используют различные функции потерь (в дальнейшем используем функцию Гаусса)

$$W(\delta, \theta) = |\delta - \theta| \qquad \text{(Лаплас)}$$
$$W(\delta, \theta) = (\delta - \theta)^2 \qquad \text{(Гаусс)}$$

Определение 18 (Риск)**.** *Риском называют $R(\delta, \theta) = \mathrm{E}_\theta[W(\delta(\vec{X}), \theta)]$*