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 $\vec{X} = (x_1, \dots, x_n)$ наызывается выборкой*

Индикатор $\mathbb{1}$

Определение 10 (Точечная оценка)**.** *Статистика* $\delta(\vec{X})$, $\delta : \mathfrak{X} \to \Theta$ *называется точечной оценкой*

Определение 11 (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)

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

Определение 12 (Риск)**.** *Риском называют* $R(\delta, \theta) = \mathrm{E}_\theta[W(\delta(\vec{X}), \theta)]$

Регрессионный анализ

Определение 13 (Регрессия)**.**
Пусть Y - наблюдение, Z - характеристика, определяющая распределение Y, F_Z - распределение Y при фиксированном Z.
Пусть $Y_1, \dots, Y_n$ - независимы. Установим зависимость Y_i от i. Сопоставим $\forall i : i \mapsto Z_i \implies F_i \equiv F_{Z_i}$. Обычно эту зависимость задают параметрически ($\mathrm{ex} : F_i = g_\theta(F_{Z_0}), \theta \in \mathbb{R}^d$).
*Тогда $E_\theta(Y|Z) = g_\theta(Z)$ - **регрессия** Y по Z.*

Определение 14 (Линейная регрессия)**.**
*Регрессия называется **линейной** если*

$$\exists X(Z) = \begin{pmatrix} X_1(Z) \\ \vdots \\ X_n(Z) \end{pmatrix} \text{ - регрессор.}$$

Модель линейной регрессии

$$\mathrm{E}_\theta(Y|Z) = X^T\beta,\ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

В условиях этой модели

$$\begin{aligned} \mathrm{E}_\theta Y_i &= (X(Z_i))^T\beta \\ Y_i &= (X(Z_i))^T\beta + \varepsilon_i \\ Y &= X^T\beta + \varepsilon(\mathrm{E}\varepsilon = 0) \end{aligned}$$

Где $X \in M_m$ - матрица регрессоров, β - $m \times 1$ - столбец параметров, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ - вектор отклонений.

Примеры регрессионных моделей. $Y_1, \dots, Y_n$ - независимые наблюдения.

1. Выборка
$$\mathrm{E}Y_i = \beta_i \text{ Если } \varepsilon_i \text{ - НОРСВ, то } Y_1, \dots, Y_n$$
2. Простая регрессионная модель
$$Y_i = \beta_1 + \beta_2 Z_i + \varepsilon_i,\ X = \begin{pmatrix} 1 & \cdots & 1 \\ Z_1 & \cdots & Z_n \end{pmatrix}$$
3. Полиномиальная модель
$$Y_i = \sum_{j=1}^{s} \beta_j Z_i^{j-1},\ X = \begin{pmatrix} 1 & \cdots & 1 \\ Z_1 & \cdots & Z_n \\ Z_1^2 & \cdots & Z_n^2 \\ \vdots & \ddots & \vdots \\ Z_1^{s-1} & \cdots & Z_n^{s-1} \end{pmatrix}$$

4. Простая группировка (однофакторный дисперсионный анализ)

$$
\begin{aligned}
&Z \in \{1, \ldots, I\} \\
&\beta = (\beta_1, \ldots, \beta_I)^T \\
&\mathrm{E}(Y|Z) = \beta_Z \\
&Y_i = \beta_{Z_i} + \varepsilon_i \\
&\forall i < j : Z_i <= Z_j \\
&X = \begin{pmatrix}
1 & 1 & 0 & 0 & & & & \\
0 & 0 & 1 & 1 & & & \mathbf{0} & \\
0 & 0 & 0 & 0 & & & & \\
& & & & \ddots & & & \\
& & \mathbf{0} & & & \ddots & & \\
& & & & & & 1 & 1
\end{pmatrix}
\end{aligned}
$$