# Окончательное решение среднеквадратического вопроса

*Достаточные статистики*

## Чертоги разума

**Определение**. Статистика $T(X_1, \dots, X_n)$ называется *достаточной*, если существует вариант условного распределения $\mu(B, t) = \mathsf{P}(\mathbf{X} \in B | T(\mathbf{X}) = t)$, которое не зависит от параметра $\boldsymbol{\theta}$.

**Теорема (критерий факторизации).** Пусть $\mathbf{X} = (X_1, \dots, X_n)$ — выборка из доминируемого семейства $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}}\}$, где $\mathsf{P}_{\boldsymbol{\theta}}$ имеет плотность $\rho_{\boldsymbol{\theta}}$ по мере $\mu$. Тогда статистика $T(\mathbf{X})$ достаточная тогда и только тогда, когда найдутся борелевские $h$ и $g$ такие, что

$$\rho_{\boldsymbol{\theta}}(\mathbf{x}) = h(\mathbf{x}) \cdot g(T(\mathbf{x}), \boldsymbol{\theta}).$$

**Основной пример.** Если модель принадлежит экспоненциальному семейству, то есть плотность в ней имеет вид

$$\rho_{\boldsymbol{\theta}}(x) = h(x) \cdot \exp\left(g(\boldsymbol{\theta}) + \sum_{i=1}^{k} a_i(\boldsymbol{\theta}) T_i(x)\right), \tag{1}$$

то вектор $\overline{\mathbf{T}(\mathbf{X})} = \left(\overline{T_1(\mathbf{X})}, \dots, \overline{T_k(\mathbf{X})}\right)$ является достаточной статистикой.

**Теорема (Колмогоров, Блэкуэлл, Рао).** Пусть $\widetilde{\theta}(\mathbf{X})$ — несмещённая оценка $\tau(\theta)$ с конечной дисперсией, $T(\mathbf{X})$ — достаточная статистика. Тогда случайная величина $\widehat{\theta}(\mathbf{X}) = \mathsf{E}_\theta(\widetilde{\theta}(\mathbf{X}) | T(\mathbf{X}))$ является несмещённой оценкой $\tau(\theta)$, причём она не хуже $\widetilde{\theta}(\mathbf{X})$ в среднеквадратичном подходе:

$$\forall \theta \in \Theta \colon \mathsf{E}_\theta \left(\widehat{\theta}(\mathbf{X}) - \tau(\theta)\right)^2 \leqslant \mathsf{E}_\theta \left(\widetilde{\theta}(\mathbf{X}) - \tau(\theta)\right)^2.$$

**Определение.** Статистика $S(\mathbf{X})$ называется *полной*, если для любой борелевской $f$ выполнена импликация:

$$[\forall \theta \in \Theta \colon \mathsf{E}_\theta f(S(\mathbf{X})) = 0] \Longrightarrow [\forall \theta \in \Theta \colon f(S(\mathbf{X})) = 0 \quad (\mathsf{P}_\theta\text{-п.н.})].$$

**Теорема (Леман, Шеффе).** Если $T(\mathbf{X})$ — полная достаточная статистика для семейства $\{\mathsf{P}_\theta\}$ и $\mathsf{E}_\theta \theta^*(\mathbf{X}) = \tau(\theta)$, то $\widehat{\theta}(\mathbf{X}) = \mathsf{E}_\theta(\theta^*(\mathbf{X}) | T(\mathbf{X}))$ — лучшая оценка в среднеквадратичном подходе среди несмещённых оценок (такие оценки называют *оптимальными*).

Помимо непосредственной проверки по определению, полноту обеспечивает следующее **достаточное условие полноты.** Рассмотрим модель из экспоненциального семейства (1). Если множество $\mathbf{a}(\Theta) \subset \mathbb{R}^k$ содержит внутреннюю точку, то статистика $\overline{\mathbf{T}(\mathbf{X})}$ полна.

**Главное следствие.** Если для полной достаточной статистики $T(\mathbf{X})$ вы придумаете такую $\varphi$, что $\mathsf{E}_\theta \varphi(T(\mathbf{X})) = \tau(\theta)$, то $\varphi(T(\mathbf{X}))$ будет оптимальной оценкой $\tau(\theta)$.

# Aufgaben

**1.** Для семейства распределений
**(а)** $\mathcal{P} = \{\mathrm{Bern}(p) \colon p \in (0,1)\}$;
**(б)** $\mathcal{P} = \{\mathcal{N}(a, \sigma^2) \colon a \in \mathbb{R}, \sigma^2 > 0\}$;
**(в)** $\mathcal{P} = \{\mathrm{Beta}(\alpha, \beta) \colon \alpha, \beta > 0\}$
предъявите достаточную статистику фиксированной размерности.

**2.** Не всякой статистикой $T(\mathbf{X})$ можно улучшать оценки посредством взятия УМО — в этом и заключается особенность достаточных статистик. Для выборки

$$X_1, \ldots, X_n \sim \mathcal{N}(a, 1)$$

приведите пример НЕдостаточной статистики $T(\mathbf{X})$ такой, что $\mathsf{E}_a(\overline{\mathbf{X}} | T(\mathbf{X}))$ зависит от неизвестного $a$ и, значит, не является статистикой.

**3.** Докажите, что статистика $X_{(1)}$ для выборки $X_1, \ldots, X_n$ в модели сдвига

$$\mathcal{P} = \{\mathrm{Exp}(1) + a \colon a \in \mathbb{R}\}$$

является полной и достаточной. Постройте с помощью неё оптимальную оценку параметра $a$.

**4.** Докажите, что в модели $\mathrm{U}(\theta, \theta + 1)$, где $\Theta = \{\theta > 0\}$, статистика $(X_{(1)}, X_{(n)})$ является достаточной, но не полной. Есть ли в ней хоть какая-нибудь полная достаточная статистика?

**5.** Найдите оптимальную оценку функции $e^{-\theta}$ для выборки $X_1, \ldots, X_n$ из распределения $\mathrm{Pois}(\theta)$.
*Указание.* В качестве стартовой несмещённой оценки возьмите оценку из первого листка.

**6.** Пусть $\xi_1, \ldots, \xi_{10}$ — н.о.р.с.в. из распределения $\mathrm{U}[0; 11]$. Обозначим $X = \min(\xi_1, \ldots, \xi_{10})$, $Y = \max(\xi_1, \ldots, \xi_{10})$. С помощью арсенала достаточных статистик найдите $\mathsf{D}(X - 4Y)$ (решения «в лоб» не принимаются).

**7*.** **(а)** Достаточная статистика $S(\mathbf{X})$ называется *минимальной*, если для любой другой достаточной статистики $T(\mathbf{X})$ найдётся функция $\varphi$ такая, что $S = \varphi \circ T$. Пусть про статистику $S(\mathbf{X})$ известно, что

$$\forall \mathbf{x}, \mathbf{y} \colon \left( S(\mathbf{x}) = S(\mathbf{y}) \Longleftrightarrow \frac{\rho_\theta(\mathbf{x})}{\rho_\theta(\mathbf{y})} \equiv \mathrm{const} \right),$$

где $\rho_\theta(\mathbf{x})$ — совместная плотность выборки. Покажите, что она является минимальной достаточной статистикой.
*Замечание.* Можете считать известным факт, что найдётся измеримая $S^{-1}$ такая, что $S(S^{-1}(s)) \equiv s$.
**(б)** Найдите минимальную достаточную статистику в модели масштаба распределения Коши:

$$\rho_\theta(x) = \frac{\theta}{\pi(x^2 + \theta^2)}.$$