Семинар №8
Достаточные статистики

Пусть X – наблюдение с неизвестным распределением $P \in \{P_\theta\}_{\theta\in\Theta}$.
Определение. Статистика $S(X)$ — достаточная для семейства распределений $\{P_\theta\}$, если условное распределение $P_\theta(X \in B|S(X) = x)$ не зависит от θ.

Условия регулярности.

1. Будем считать, что семейство распределений $\{P_\theta\}$ доминируемо (т.е. состоит либо только из дискретных распределений, либо только из абсолютно непрерывных).

2. Примем $p_\theta(x)$ равным $P_\theta(X = x)$ в дискретном случае и равным плотности в абсолютно непрерывном случае.

Задача 1. Пусть $X_1, \ldots, X_n$ – выборка из распределения $Bern(\theta)$. Докажите, что $\sum_{i=1}^n X_i$ – достаточная статистика.
Решение. Распишем условное распределение значений случайных величин из выборки при условии, что $\sum_{i=1}^n X_i = s$:

$$P(X_1 = x_1, \ldots, X_n = x_n|S(X) = s) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, \sum\limits_{i=1}^n X_i = s)}{\sum\limits_{i=1}^n X_i = s} =$$

$$= I\{\sum_{i=1}^n x_i = s\}\frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{C_n^s\theta^s(1-\theta)^{n-s}} = \frac{1}{C_n^s} I\{\sum_{i=1}^n x_i = s\}.$$

Видим, что условное распределение не зависит от θ, значит, статистика $\sum\limits_{i=1}^n X_i$ является достаточной для семейства распределений $Bern(\theta)$. ■

Теорема 1. *(Критерий факторизации Неймана-Фишера.) Пусть $\{P_\theta, \theta \in \Theta\}$ – доминируемое семейство распределений с обобщённой плотностью $p_\theta(x)$, X – наблюдение (выборка) из неизвестного распределения $P \in \{P_\theta, \theta \in \Theta\}$. Тогда*

$$S(X) - \textit{достаточная} \iff p_\theta(X) = \psi(S(X), \theta)h(X),$$

где функция $h(x)$ не зависит от параметра θ.

Задача 2. Пусть $X_1, \ldots, X_n$ – выборка из $N(a, \sigma^2)$, $\theta = (a, \sigma^2)$. Найти достаточную статистику.
Решение. Распишем правдоподобие для выборки из $N(a, \sigma^2)$.

$$p_\theta(X_1, \ldots, X_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(X_i - a)^2\right\} =$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum X_i^2 + \frac{a}{\sigma^2}\sum X_i - \frac{na^2}{2\sigma^2}\right\}.$$

Видим, что $h(X) = 1$, потому что в выражении правдоподобия нет функций от выборки, которые отделимы от параметров. Тогда $S(X) = (\sum X_i, \sum X_i^2)$. ■

Теорема 2. *(Колмогоров-Блэкуэлл-Рао, об улучшении несмещённых оценок.)*
Пусть $\widetilde{\theta}$ – несмещённая оценка $\tau(\theta)$, $S(X)$ – достаточная статистика. Тогда $\theta^* = E_\theta(\widetilde{\theta}|S(X))$ – несмещённая оценка, и для любого $\theta \in \Theta$ $D_\theta\theta^* \leq D_\theta\widetilde{\theta}$, причём если равенство достигается при всех $\theta \in \Theta$, то статистика $\widetilde{\theta}$ является $S(X)$ -измеримой.

Замечание 1. *Если* $\tau(\theta) \in \mathbb{R}^k$, *где* $k > 1$, *то выражение* $D_\theta \theta^* \leq D_\theta \widetilde{\theta}$ *означает, что матрица* $D_\theta \widetilde{\theta} - D_\theta \theta^*$ *неотрицательно определена.*

Определение. θ^* – оптимальная оценка $\tau(\theta)$, если θ^* – несмещённая оценка $\tau(\theta)$ с равномерно наименьшей дисперсией в классе несмещённых оценок (т.е. её дисперсию с помощью теоремы Блэкуэлла-Рао уже нельзя уменьшить).

Определение. $S(X)$ – полная для $\{P_\theta, \theta \in \Theta\}$, если для всех функций $f(x)$ таких, что $\forall \theta \in \Theta$ выполнено $E_\theta f(S(X)) = 0$, следует $f(S(X)) = 0$ P_θ-п.н. для всех $\theta \in \Theta$.

Теорема 3. *(Леман-Шефаре)*
Пусть $S(X)$ *– полная достаточная статистика,* $\widetilde{\theta}$ *– несмещённая оценка* $\tau(\theta)$, *тогда* $\theta^* = E_\theta(\widetilde{\theta}|S(X))$ *– оптимальная оценка* $\tau(\theta)$.

Определение. Семейство $\{P_\theta, \theta \in \Theta\}$ называется экспоненциальным, если плотность распределения P_θ имеет вид

$$p_\theta(x) = h(x) \exp\left(\sum_{i=1}^{k} a_i(\theta) u_i(x) + v(\theta)\right).$$

Теорема 4. *(Об экспоненциальных семействах.)*
Пусть $\theta \in \Theta \subset \mathbb{R}^k$ *и все значения вектора* $(a_1(\theta), \ldots, a_k(\theta))$ *образуют* k*-мерный параллелепипед, тогда* $S(X) = (u_1(x), \ldots, u_k(x))$ *– полная достаточная статистика.*

Замечание 2. *Условие теоремы будет выполнено, если множество* Θ *"телесно" – содержит свои "внутренние" точки (т.е. если некая окрестность без точки лежит в* Θ, *то и сама точка лежит в* Θ*) и функции* $a_1(\theta), \ldots, a_k(\theta)$ *линейно независимы.*

Задача 3. Доказать, что статистика $\sum\limits_{i=1}^{n} X_i$ является полной для семейства распределений $Bern(\theta)$, $\theta \in (0;1)$.

Решение. Итак, нужно доказать, что если для всех $\theta \in (0;1)$ и какой-то функции $f(x)$ выполнено $E_\theta f(\sum X_i) = 0$, то отсюда следует, что $f(x) = 0$ для всех x – значений $\sum X_i$. Так как X_i – бернуллиевские и независимые, то $\sum X_i \sim Bin(n, \theta)$. Распишем матожидание $E_\theta f(\sum X_i)$:

$$E_\theta f\left(\sum X_i\right) = \sum_{k=0}^{n} f(k) C_n^k \theta^k (1-\theta)^{n-k}.$$

Мы получили в правой части многочлен не более чем $n-$ ной степени от θ, он имеет не более чем n корней на отрезке $(0,1)$. Но этот многочлен равен 0 для всех $\theta \in (0,1)$, т.е. все θ из интервала $(0,1)$ являются его корнями — противоречие. Значит, все $f(k) = 0$ для $k = 0, \ldots, n$. Значит, статистика $\sum X_i$ – полная. □

Следствие 1. *(Из теоремы 3.) Пусть* $\varphi(x)$ *— решение уравнения несмещённости* $E_\theta \varphi(S(X)) = \tau(\theta)$ $\forall \theta \in \Theta$, *где* $S(X)$ *— полная достаточная статистика, а* $\tau(\theta)$ *— тот параметр, который мы оцениваем. Тогда* $\varphi(S(X))$ *— оптимальная оценка* $\tau(\theta)$.

Задача 4. Пусть $X_1, \ldots X_n$ — выборка из $R[0, \theta]$. Найти оптимальную оценку для θ.

Решение. Так как для выборки из равномерного закона правдоподобие $p(X, \theta) = \frac{1}{\theta^n} I\{X_{(n)} \leq \theta\}$, то по критерию факторизации (теорема 1) $X_{(n)}$ – достаточная (действительно, $\psi(S(X), \theta) = \frac{1}{\theta^n} I\{X_{(n)} \leq \theta\}$, а $h(X) = 1$). Проверим, что эта статистика – полная.

$$E_\theta f(X_{(n)}) = \int\limits_0^\theta f(x) n \frac{x^{n-1}}{\theta^n} dx = 0 \;\forall \theta > 0 \Longrightarrow \int\limits_0^\theta f(x) x^{n-1} dx = 0 \;\forall \theta > 0.$$

Так как это выполнено для любого $\theta > 0,$ то сама подынтегральная функция равна 0 (кто не верит, возьмите производную по θ от $G(\theta) = \int\limits_0^\theta f(x)x^{n-1}dx,$ $G(\theta)$ дифференцируема по теореме Ньютона-Лейбница и $G'(\theta)$ должна равняться 0, так как $G(\theta) \equiv 0$). Т.е. $f(x)x^n = 0 \ \ \forall x > 0,$ значит, $f(x) \equiv 0$ и $X_{(n)}$ – достаточная.

Итак, мы знаем, что $X_{(n)}$ – полная достаточная статистика, значит, оптимальная оценка есть $\varphi(X_{(n)})$. Решим уравнение несмещённости и найдём φ. Мы уже знаем, что $E_\theta X_{(n)} = \frac{n}{n+1}\theta,$ отсюда, $E_\theta\left(\frac{n+1}{n}X_{(n)}\right) = \theta,$ т.е. $\varphi(x) = \frac{n+1}{n}x$ и $\frac{n+1}{n}X_{(n)}$ – оптимальная. $\square$