7. Информационное неравенство.

Пусть имеется выборка размера 1 из бернуллиевского распределения. Насколько хорошо я могу оценить вероятность успеха? Очевидно, что не могу сказать ничего определенного, одно наблюдение может помочь мне составить только самое общее представление о параметре. При этом одно наблюдение из $R[0,\theta]$, скажем, будет несколько более информативным — я заведомо смогу вычеркнуть из потенциальных значений θ отрезок $[0, X_1]$.

В рамках сегодняшнего материала нам хотелось бы отказаться от второго типа оценок, которые позволяют отдельным наблюдениям сужать область рассматриваемого параметра. Для этого, нам хотелось бы некоторую структуру пространства Θ и плотностей распределения θ. Наложим следующие условия, которые мы на прошлом занятии назвали условиями регулярности:

1) Пусть Θ — открытый интервал прямой (возможно бесконечный).

2) Распределения F_θ имеют плотности $f_\theta(x)$ и носители этих плотностей (распределений) не зависят от θ.

3) Плотность $f_\theta(x)$ имеет конечную производную по θ в каждой точке Θ при каждом x из носителя.

4) Для величины $U_1 = \frac{\partial}{\partial\theta}\ln f(X_1,\theta)$ $\mathbf{E}U_1 = 0$, $0 < I_1(\theta) = \mathbf{D}U_1 < \infty$.

Условие 2) как раз и запрещает нам полностью отвергать значения θ на основе наблюдений. Условия 3) и 4) говорят о том, что зависимость плотности от θ достаточно гладка.

Аналогичные условия наложим в дискретном случае, рассматривая при этом вместо $f_\theta(x)$ функцию $\mathbf{P}_\theta(X = x)$. Нам потребуется количественная характеристика информации о параметре, содержащейся в выборке. Назовем информацией Фишера одного элемента X следующую величину:

$$I(\theta) = I_1(\theta) = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)^2 = \mathbf{D}_\theta\frac{\partial}{\partial\theta}\ln f_\theta(X)$$

Иначе говоря, функция плотности одной случайной величины $f_\theta(x)$ логарифмируется и затем дифференцируется по θ, в полученную функцию вместо аргумента x подставляется случайная величина X и у этой величины подсчитывается дисперсия.

Информацией Фишера выборки $X_1, ..., X_n$ называют функцию

$$I_n(\theta) = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(X_1, ..., X_n)\right)^2.$$

В нашем случае н.о.р. величин $I_n(\theta) = nI(\theta)$.

Отметим, что в непрерывном случае

$$\mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right) = \int_{\mathbb{R}}\frac{\frac{\partial f_\theta(x)}{\partial\theta}}{f_\theta(x)}f_\theta(x)dx = \frac{\partial}{\partial\theta}\left(\int_{\mathbb{R}}f_\theta(x)dx\right) = \frac{\partial}{\partial\theta}1 = 0,$$

откуда формулы для $I(\theta)$ равносильны (дискретный случай проверяется аналогично).

Пример 1. Подсчитаем информацию Фишера для бернуллиевских величин с параметром $\theta \in (0,1)$. Запишем $\ln f_\theta$ и продифференцируем его:

$$\ln f_\theta(x) = \ln\left(\theta^x(1-\theta)^{1-x}\right) = (1-x)\ln(1-\theta) + x\ln\theta,\ \frac{\partial}{\partial\theta}\ln f_\theta(x) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)},$$

откуда

$$I(\theta) = \mathbf{D}_\theta\frac{\partial}{\partial\theta}\ln f_\theta(X) = \frac{\mathbf{D}_\theta X}{(\theta(1-\theta))^2} = \frac{1}{\theta(1-\theta)},\ I_n(\theta) = \frac{n}{\theta(1-\theta)}.$$

Таким образом, информация о выборке минимальна при θ близким к $1/2$ (в этом случае при небольшом изменении параметра доля 0 и 1 в выборке меняется незначительно), а максимальна (бесконечна) при θ, приближающихся к 0 или 1 (в этом случае даже небольшие изменения параметра могут оказаться

значимыми — в выборке значительно изменится соотношение 0 и 1).

Пример 2. Подсчитаем информацию Фишера для нормальных величин $\mathcal{N}(\theta, 1)$.

$$\ln f_\theta(x) = -\ln\sqrt{2\pi} - \frac{(x-\theta)^2}{2},\ \frac{\partial}{\partial\theta}\ln f_\theta(x) = x - \theta.$$

Таким образом,

$$I(\theta) = \mathbf{D}_\theta(X-\theta) = 1,\ I_n(\theta) = n.$$

Итак, информация Фишера для $\mathcal{N}(\theta, 1)$ постоянна и равна 1, информация Фишера выборки равна n. Информацией о параметре, содержащейся в статистике T назовем

$$I_T(\theta) = \mathbf{D}_\theta\frac{\partial}{\partial\theta}(\ln f_{T(X_1,...,X_n),\theta}(T(X_1,...,X_n))).$$

Можно доказать, что $I_{T(X_1,...,X_n)} \le I_{X_1,...,X_n}$, то есть информация, содержащаяся в выборке не меньше, чем информация, содержащаяся в функции от нее, что вполне естественно. Кроме того равенство $I_{T(X_1,...,X_n)}(\theta) = I_{X_1,...,X_n}(\theta)$ выполняется тогда и только тогда, когда T достаточна.

Пример 3. Найдем в модели предыдущего примера информацию, содержащуюся в $\overline{X}$. Распределение этой величины $\mathcal{N}(\theta, 1/n)$, откуда

$$\ln f_\theta(x) = -\ln\sqrt{2\pi/n} - \frac{n(x-\theta)^2}{2},\ \frac{\partial}{\partial\theta}\ln f_\theta(x) = n(x-\theta),$$

и

$$I_T(\theta) = n^2\mathbf{D}\overline{X} = n = I_n(\theta).$$

Перейдем к базовому результату, связанному с понятием информации Фишера: информационному неравенству:

Теорема 1 (Неравенство Рао-Крамера). Пусть выполнены условия регулярности и $\hat\theta$ — оценка с математическим ожиданием $g(\theta)$. Тогда

$$\mathbf{D}_\theta\hat\theta(X_1,...,X_n) \ge \frac{(g'(\theta))^2}{I_n(\theta)}.$$

В частности, для несмещенных оценок имеем $\mathbf{D}_\theta\hat\theta \ge 1/(nI(\theta))$.

Пример 4. Рассмотрим оценку $\overline{X}$ в схеме Бернулли. Ее дисперсия равна $\theta(1-\theta)/n$. При этом в силу неравенства Рао-Крамера несмещенные оценки в этой модели не могут иметь дисперсию лучше чем $\theta(1-\theta)/n$ (см. пример 3). Значит $\overline{X}$ — оптимальная.

Оптимальные оценки, чья дисперсия совпадает с нижней границей из неравенства Рао-Крамера называются эффективными. Эффективные оценки бывают только в определенных моделях и у определенных функций, но об этом чуть позже.

Пример 5. Оценка $\hat\theta = e^{\overline{X}-1/(2n)}$ в модели $\mathcal{N}(\theta, 1)$ имеет математическое ожидание

$$e^{\theta-1/(2n)}\mathbf{E}e^{\overline{X}-\theta} = e^{\theta-1/(2n)}\mathbf{E}e^{Z/\sqrt{n}} = \frac{e^{\theta-1/(2n)}}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{x/\sqrt{n}}e^{-x^2/2}dx = \frac{e^\theta}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(x-1/\sqrt{n})^2/2}dx = e^\theta,$$

где $Z = \sqrt{n}(\overline{X}-\theta) \sim \mathcal{N}(0,1)$. Следовательно, $\hat\theta$ является оптимальной для функции e^θ как функция от полной достаточной статистики. Покажем, что она не является эффективной оценкой той же функции. Аналогично предыдущим оценкам $\mathbf{E}e^{2\overline{X}} = e^{2\theta+2/n}$, откуда

$$\mathbf{D}_\theta\hat\theta = \mathbf{E}e^{2\overline{X}-1/n} - e^{2\theta} = e^{2\theta+1/n} - e^{2\theta} = e^{2\theta}(e^{1/n}-1)$$

$$\frac{\partial}{\partial\theta}\ln f_\theta(x) = n(x-\theta),$$

При этом $I(\theta) = 1$, $g(\theta) = e^\theta$, откуда нижняя граница в неравенстве Рао-Крамера есть $e^{2\theta}/n$. Асимптотически эти границы эквивалентны, но при фиксированном n $e^{1/n} - 1 > 1/n$, откуда нижняя граница неравенства Рао-Крамера оказывается недостижимой.

Неравенство Рао-Крамера показывает наилучший порядок приближения любой оценкой параметра в регулярной модели. Действительно

$$\mathbf{E}_\theta(\hat\theta - \theta)^2 = \mathbf{D}_\theta\hat\theta + (\mathbf{E}_\theta\hat\theta - \theta)^2.$$

Эта величина по порядку не меньше $1/n$, откуда $\hat\theta - \theta$ имеет порядок не меньше, чем $n^{-1/2}$. Вот почему, рассматривая асимптотически нормальные оценки нам было наиболее интересно смотреть случай $\sigma_n(\theta) = \sigma(\theta)n^{-1/2}$. При этом мы видели, что в нерегулярных моделях такого может и не быть, скажем $(n+1)X_{(n)}/n$ в модели $R[0,\theta]$ имеет дисперсию порядка $1/n^2$.

Из доказательства неравенства Рао-Крамера нетрудно вывести, что равенство там возможно только в так называемых экспоненциальных моделях, в которых

$$f_\theta(x) = \exp(-A(\theta)B(x) + C(\theta))D(x).$$

В них существует эффективная оценка для функции $g(\theta) = C'(\theta)/A'(\theta)$ существует и имеет вид

$$T(x_1, ..., x_n) = n^{-1}\sum_{i=1}^n B(x_i).$$

Нетрудно понять, что эффективные оценки будут также для линейных функций от таких статистик. Других возможных эффективно оцениваемых функций не будет. Таким образом, эффективные оценки существуют лишь для небольшого набора функций лишь в некоторых моделях. В остальных случаях неравенство Крамера не дает наилучших оценок.

Этот результат следует принять во внимание, но пользоваться им как доказанным в рамках занятий нельзя

Пример 6. . Для нормального распределения $\mathcal{N}(\theta, 1)$ плотность имеет вид $f_\theta(x) = (\sqrt{2\pi})^{-1}e^{-(x-\theta)^2/2} = (\sqrt{2\pi})^{-1}e^{-x^2}e^{\theta x - \theta^2/2}$, откуда $A(\theta) = \theta$, $B(x) = x$, $C(\theta) = \theta^2/2$, $D(x) = e^{-x^2}$. Соответственно, оценка $\overline{X}$ является эффективной оценкой θ.

При этом асимптотически эффективные оценки (то есть асимпотически нормальные оценки с асимптотической дисперсией $I(\theta)$) существуют в значительно большем количестве моделей (на прошлом занятии мы сформулировали, что в сильно регулярных моделях такими являются ОМП).

Однако возникает вопрос — действительно ли это наименьшая возможная асимптотическая дисперсия? Вообще говоря, нет, ведь из сходимости распределений не следует сходимость их дисперсий. Примером такого случая служит задача 2.1.3, в которой для модели $X_i \sim \mathcal{N}(\theta, 1)$ строится асимптотически нормальная оценка с асимптотической дисперсией 1 при $\theta \neq 0$ и $b < 1$ при $\theta = 0$, тогда как информация Фишера модели 1.

Однако, при дополнительном условии непрерывности асимптотической дисперсии, оказывается, что асимптотическая дисперсия не может быть меньше $I(\theta)$. Отсюда следует еще одно свойство ОМП в сильно регулярных моделях — их асимптотическая дисперсия наилучшая возможная среди непрерывных.

В многомерном случае (случае многомерного параметра $\theta = (\theta_1, ..., \theta_m)$) вместо $I(\theta)$ рассматривают матрицу из

$$I_{i,j}(\theta) = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta_i}\ln f_\theta(X_1)\frac{\partial}{\partial\theta_j}f_\theta(X_1)\right),$$

называемую информационной матрицей.

Неравенство Рао-Крамера для несмещенных оценок $\hat{\theta}$ при этом принимает такой вид:

$$\Sigma^2(\theta) \geq \frac{1}{n} I^{-1}(\theta),$$

где Σ^2 — ковариационная матрица вектора $\hat{\theta}$, а под $A \geq B$ для матриц мы понимаем следующее: $A - B$ положительно определена. Иначе говоря, для любого вектора $s = (s_1, ..., s_m)$

$$\mathbf{D}_\theta \left(\sum_{i=1}^{m} s_i \hat{\theta}_i \right) \geq \frac{1}{n} s^t I^{-1}(\theta) s.$$

Таким образом, в многомерном случае мы получаем оценку снизу для дисперсий всех линейных комбинаций наших оценок.
Аналогичным образом, информационной матрицей статистики T будем называть

$$I_{i,j}(\theta) = \mathbf{E}_\theta \frac{\partial}{\partial \theta_i} \ln f_{T,\theta}(T) \frac{\partial}{\partial \theta_j} \ln f_{T,\theta}(T),$$

где $f_\theta(t)$ — плотность статистики T. Как и прежде достаточность равносильна тому, что информационная матрица статистики равна $nI(\theta)$.

Приложение. Мотивация информации Фишера

Этот раздел не входит в обязательное к зачету, а приведен прежде всего для мотивации определения информации Фишера. Давайте представим себе, что у нас есть два возможных распределения с плотностями f и g и есть некоторые априорные вероятности p, q того, что распределение имеет плотность f или g соответственно. Тогда после наблюдения реализации выборки $x_1, ..., x_n$ у нас появляются апостериорные вероятности

$$\tilde{p} = \frac{pf(x_1)....f(x_n)}{pf(x_1)....f(x_n) + qg(x_1)....g(x_n)}, \ \tilde{q} = \frac{qg(x_1)....g(x_n)}{pf(x_1)....f(x_n) + qg(x_1)....g(x_n)}.$$

Априорно мы считали, что первая плотность в $d = p/q$ раз вероятнее второй, апостериорно стали считать то же самое с отношением $\tilde{d} = \tilde{p}/\tilde{q}$. Соответственно $\tilde{d}/d$ это некоторый показатель, отражающий количество информации, помогающей нам выбрать из f и g, содержащейся в выборке. Глядя на наши формулы, мы можем увидеть, что это отношение есть $\frac{f(x_1)...f(x_n)}{g(x_1)...g(x_n)}$. Давайте в качестве меры различимости мер f и g возьмем

$$\ln \tilde{d}/d = \sum_{i=1}^{n} \ln \frac{f(x_i)}{g(x_i)}.$$

Назовем *расстоянием Кульбака-Лейблера* между вероятностными распределениями F, G с плотностями f, g следующую величину:

$$I(F : G) = \int_{\mathbb{R}} \ln \left(\frac{f(x)}{g(x)} \right) f(x) dx,$$

в дискретном случае, когда F, G имеют вероятности $f(x)$, $g(x)$ для отдельных значений x,

$$I(F : G) = \sum_{x} \ln \left(\frac{f(x)}{g(x)} \right) f(x).$$

Рассматриваемая величина есть $\mathbf{E}_F \ln(f(X)/g(X))$, где индекс F означает, что X имеет функцию распределения F. Это являет собой среднее значение нашего $\ln \tilde{d}/d$ при условии, что распределение F.

Строго говоря, рассматриваемая величина не является метрикой, поскольку несимметрична, но неотрицательно и обращается в ноль лишь при $F = G$:

$$I(F:G) = -\mathbf{E}_F \ln\left(\frac{g(X)}{f(X)}\right) \geq -\ln\left(\mathbf{E}_F \frac{g(X)}{f(X)}\right) = -\ln\left(\int_{\mathbb{R}} g(x)dx\right) = 0.$$

Введенное расстояние по сути измеряет, насколько вероятны большие различия между F и G, то есть насколько легко различить какое было распределение, глядя на выборку.

Пример 7. Найдем расстояние между $Bernoulli(p)$ и $Bernoulli(1-p)$. Тогда

$$I(p:1-p) = \ln(p/(1-p))p + \ln((1-p)/p)(1-p) = (2p-1)\ln(p/(1-p)).$$

При отдалении p от $1/2$ оно возрастает до бесконечности.
С нашей, сугубо статистической позиции, вполне разумно рассматривать расстояния между распределениями, соответствующими двум различным значениям параметра θ на основе выборки размера n:

$$I(\theta_1:\theta_2,n) = \int_{\mathbb{R}^n} \ln\left(\frac{f_{\theta_1}(x_1,...,x_n)}{f_{\theta_2}(x_1,...,x_n)}\right) f_{\theta_1}(x_1,...,x_n)dx_1...dx_n$$

В привычном для нас случае н.о.р. наблюдений имеем

$$I(\theta_1:\theta_2,n) = \mathbf{E}_{\theta_1}\left(\ln\left(\frac{f_{\theta_1}(X_1)...f_{\theta_1}(X_n)}{f_{\theta_2}(X_1)...f_{\theta_2}(X_n)}\right)\right) = nI(\theta_1:\theta_2,1),$$

т.е. расстояние между параметрами линейно растет с увеличением выборки.
Предположим, что параметр меняется непрерывно и f_θ зависит от него достаточно гладко. Тогда посмотрим насколько хорошо параметр θ отделяется от окрестных значений:

$$I(\theta:\theta+\Delta\theta,1) = -\mathbf{E}_\theta \ln\left(\frac{f_{\theta+\Delta\theta}(X)}{f_\theta(X)}\right) = -\mathbf{E}_\theta \ln\left(1 + \frac{\Delta\theta f'_\theta(X) + \Delta\theta^2 f''_\theta(X)/2 + o(\Delta\theta^2)}{f_\theta(X)}\right) =$$
$$-\Delta\theta \mathbf{E}_\theta \frac{f'_\theta(X)}{f_\theta(X)} - \Delta\theta^2 \mathbf{E}_\theta \frac{f''_\theta(X)}{2f_\theta(X)} + \frac{1}{2}\Delta\theta^2 \mathbf{E}_\theta\left(\frac{f'_\theta(X)}{f_\theta(X)}\right)^2 + o(\Delta\theta^2).$$

Если семейство f_θ устроено достаточно хорошо, чтобы можно было менять порядок интегрирования и дифференцирования, то первые два слагаемых последней суммы нулевые, поскольку

$$\int_{\mathbb{R}} f'_\theta(x)dx = \left(\int_{\mathbb{R}} f_\theta(x)dx\right)' = 0,\ \int_{\mathbb{R}} f''_\theta(x)dx = \left(\int_{\mathbb{R}} f_\theta(x)dx\right)'' = 0,$$

откуда

$$\frac{2I(\theta:\theta+\Delta\theta,1)}{\Delta\theta^2} \to E_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)^2,\ \Delta\theta\to 0.$$

Правая часть характеризует, насколько хорошо параметр θ хорошо отделяется на основе выборки из одного элемента от окрестных значений параметра. Будем называть ее *информацией Фишера* $I(\theta) = I_1(\theta)$. Аналогично

$$I_n(\theta) = \mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(X_1,...,X_n)\right)^2$$

7.1.1 Найти $I(\theta)$ $X \sim Geom(\theta)$. Показать, что $\overline{X}$ эффективна для своего матожидания.
7.2.1 Найти $I(\theta)$ а) $X \sim Poiss(\theta)$ б) $X_i \sim Binom(m,\theta)$. Показать, что $\overline{X}$ эффективна для своего м.о..
7.3.1 Найти $I(\theta)$, $X \sim \exp(\theta)$. Показать, что $\overline{X}$ эффективна для $1/\theta$.
7.1.2 Показать, что $\overline{I_{X\neq 3}}/3$ эффективна для $\mathbf{P}(X=2) = 2\mathbf{P}(X=1) = 2\theta$, $\mathbf{P}(X=3) = 1-3\theta$.
7.2.2 а) $f_X(x) = \theta^{-a}x^{a-1}e^{-x/\theta}I_{x>0}/\Gamma(a)$ б) $\mathbf{P}_\theta(X_i \leq x) = x^{1/\theta}$, $x\in[0,1]$. Найти эффективную оценку θ.
7.3.2 Для $f_X(x) = \theta e^{-|x|\theta}/2$ найти эффективную оценку $1/\theta$.
7.1.3 Пусть X равна a_1, a_2, a_3 с вероятностями $\theta_1, \theta_2, 1-\theta_1-\theta_2$. Найти информационную матрицу для

выборки и для пары статистик $(\overline{I_{X=a_1}}, \overline{I_{X=a_2}})$.

7.2.3 Построить информационную матрицу для $\mathbf{P}(X=0) = 1-\theta_1$, $\mathbf{P}(X=k) = \theta_1\theta_2(1-\theta_2)^{k-1}$, $k>0$.

7.3.3 Построить информационную матрицу для $X_i \sim \mathcal{N}(\theta_1, \theta_2^2)$.

7.0.2* Доказать, через информацию Фишера, что $(\overline{X}, nS^2)$ — достаточная статистика для $X_i \sim \mathcal{N}(\theta_1, \theta_2^2)$, используя то, что $nS^2/\theta_2^2 \sim \Gamma(n/2, 1/2)$, S^2 и $\overline{X}$ независимы.