10 Применение теории больших уклонений

10.1 Теорема Санова

Пример 1. Рассмотрим следующую задачу. Пусть мы рассматриваем эксперимент с k возможными исходами, имеющими вероятности $p_1, ..., p_k$, $\vec{N} = (N_1, ..., N_k)$ — число исходов каждого типа за n испытаний, $\vec{\nu} = \vec{N}/n$ — частоты каждого исхода. Тогда мы можем задать вопросом насколько вероятно то, что $\vec{\nu}$ попадет в то или иное множество.

Для результата о больших уклонениях $\vec{\nu}$ нам понадобятся векторы $\vec{X}_i \in \mathbb{R}^k$, принимающие одно из значений $(1, 0, ..., 0)$, $(0, 1, ..., 0)$,..., $(0, 0, ..., 1)$, $(0, 0, ..., 0)$ с вероятностями $p_1, ..., p_k$. Тогда $\vec{S}_n = (N_1, ..., N_{k-1})$. Функция $R(\vec{h}) = \sum_{i=1}^{k-1} p_i e^{h_i} + p_k$. При этом

$$grad \ln R(\vec{h}) = \left(\frac{p_i e^{h_i}}{R(\vec{h})},\ i \le k-1 \right).$$

Для применения многомерной теоремы о больших уклонениях нам понадобится такое $\vec{h}$, что $grad \ln R(\vec{h}) = (\theta_1, \dots, \theta_{k-1})$, $\sum_{i=1}^{k-1} \theta_i \le 1$. При этом

$$\frac{p_k}{R(\vec{h})} = \theta_k,$$

где $\theta_k = 1 - \theta_1 - \dots - \theta_{k-1}$. Тогда

$$\Lambda(\vec{\theta}) = (\vec{\theta}, \vec{h}) - \ln R(\vec{h}) = \sum_{i=1}^{k} \theta_i h_i - \ln R(\vec{h}).$$

Но $h_i = \ln R(\vec{h}) + \ln(\theta_i/p_i)$, откуда

$$\Lambda(\vec{\theta}) = \sum_{i=1}^{k-1} \theta_i h_i - \ln R(\vec{h}) = \sum_{i=1}^{k-1} \theta_i \ln \frac{\theta_i}{p_i} - \theta_k \ln R(\vec{h}) = \sum_{i=1}^{k} \theta_i \ln \frac{\theta_i}{p_i}.$$

Кроме того,

$$\Sigma(\vec{h}) = \begin{pmatrix} \frac{p_1 e^{h_1}}{R(\vec{h})} - \left(\frac{p_1 e^{h_1}}{R(\vec{h})}\right)^2 & -\frac{p_1 p_2 e^{h_1+h_2}}{R(\vec{h})^2} & \dots & -\frac{p_1 p_{k-1} e^{h_1+h_{k-1}}}{R(\vec{h})^2} \\ -\frac{p_1 p_2 e^{h_1+h_2}}{R(\vec{h})^2} & \frac{p_2 e^{h_2}}{R(\vec{h})} - \left(\frac{p_2 e^{h_2}}{R(\vec{h})}\right)^2 & \dots & -\frac{p_2 p_{k-1} e^{h_2+h_{k-1}}}{R(\vec{h})^2} \\ & & \dots & \\ -\frac{p_1 p_{k-1} e^{h_1+h_{k-1}}}{R(\vec{h})^2} & -\frac{p_2 p_{k-1} e^{h_2+h_{k-1}}}{R(\vec{h})^2} & \dots & \frac{p_{k-1} e^{h_{k-1}}}{R(\vec{h})} - \left(\frac{p_{k-1} e^{h_{k-1}}}{R(\vec{h})}\right)^2 \end{pmatrix}.$$

При нашем h

$$\Sigma(\vec{h}) = \left(\begin{array}{cccc} \theta_1 - \theta_1^2 & -\theta_1\theta_2 & \ldots & -\theta_1\theta_{k-1} \\ -\theta_1\theta_2 & \theta_2 - \theta_2^2 & \ldots & -\theta_2\theta_{k-1} \\ & & \ldots & \\ -\theta_1\theta_{k-1} & \theta_2\theta_{k-1} & \ldots & \theta_{k-1} - \theta_{k-1}^2 \end{array} \right).$$

Вынесем из первой строки и первого столбца $\sqrt{\theta_1}$, из вторых θ_2 и так далее, получим матрицу

$$E - (a_1, \ldots, a_{k-1})^T (a_1, \ldots, a_{k-1}),$$

где $a_j = -\sqrt{\theta_j}$. Указанная матрица при вычитании E становится матрицей ранга 1, а значит имеет собственное значение 1 кратности $n-2$. При этом $\vec{a}$ являтся собственным вектором с собственным значением $1 - ||\vec{a}||^2 = \theta_k$. Отсюда,

$$\det(\Sigma(\vec{h})) = \theta_1 \cdots \theta_k.$$

Отсюда в силу локальной предельной теоремы

$$\mathbf{P}(\vec{S}_n = n\vec{\theta}) = \frac{1}{(\sqrt{2\pi n})^{k-1}\sqrt{\theta_1 \cdots \theta_k}} \exp\left(-n \sum_{i=1}^{k} \theta_i \ln \frac{\theta_i}{p_i}\right),$$

где $\vec{\theta}$ — вектор с положительными компонентами, в сумме дающими менее 1. Этот результат называется теоремой Санова.

Пример 2. Применим теорему Санова к задаче проверки гипотезы согласия $H_0 : p_1 = p_1^0, ..., p_k = p_k^0$, где p_i^0 заданы, с помощью критериев отношения правдоподобия и хи-квадрат. Критерий отношения правдоподобий основан на статистике

$$T_1(X_1, ..., X_n) = 2\sum_{i=1}^{k} N_i \ln \frac{N_i}{np_i^0},$$

критерий хи-квадрат основан на статистике

$$T_2(X_1, ..., X_n) = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0},$$

обе из которых имеют асимптотическое распределение χ^2_{k-1} при выполнении гипотезы. Тем самым вероятность $\mathbf{P}_0(T_1(X_1, ..., X_n) > c)$ и $\mathbf{P}_0(T_2(X_1, ..., X_n) > c)$ аппроксимируются вероятностью $1 - F_{\chi^2_{k-1}}(c)$.

Можно сказать, что критерий хи-квадрат получается из критерия отношения правдоподобий аппроксимацией логарифма первыми членами разложения, а критерий хи-квадрат вытекает из нормальной аппроксимации для вектора частот $(\nu_1, \dots, \nu_k)$.

Эта аппроксимация довольно точна при c порядка константы, но, скажем, при c порядка n она является неудовлетворительной.

Фактический уровень значимости критериев в этом случае можно оценить с помощью теоремы Санова, используя представления

$$\mathbf{P}(T_1(X_1, ..., X_n) > na) = \mathbf{P}\left(2\sum_{i=1}^{k} \nu_i \ln\left(\frac{\nu_i}{p_i^0}\right) > a\right) = \mathbf{P}((\nu_1, \dots, \nu_{k-1}) \in A),$$

$$\mathbf{P}(T_2(X_1, ..., X_n) > na) = \mathbf{P}\left(\sum_{i=1}^{k} \frac{(\nu_i - p_i^0)^2}{p_i^0} > a\right) = \mathbf{P}((\nu_1, \dots, \nu_{k-1}) \in B),$$

где

$$A = \{\theta_i : 2\sum_{i=1}^{k} \theta_i \ln\left(\frac{\theta_i}{p_i^0}\right) > a\}, \quad B = \{\theta_i : \sum_{i=1}^{k} \theta_i = 1,\ \sum_{i=1}^{k} \frac{(\theta_i - p_i^0)^2}{p_i^0} > a\}.$$

При этом

$$\Lambda(\vec{\theta}) = \sum_{i=1}^{k} \theta_i \ln\left(\frac{\theta_i}{p_i^0}\right)$$

выпуклая функция, достигающая минимума на границе рассматриваемых областей. При этом $\Lambda(\vec{\theta}) = a/2$ при всех $\vec{\theta} \in \partial A$, поэтому наша асимптотика вероятности ошибки первого рода будет порядка $\exp(-na/2)$. Хи-квадрат аппроксимация дала бы вероятность

$$1 - F_{\chi^2_{k-1}}(na) \sim \frac{2}{2^{(k-1)/2}\Gamma((k-1)/2)}(na)^{(k-1)/2-1}e^{-na/2}$$

где последнее соотношение выводится с помощью правила Лопиталя из формулы гамма-плотности. Тем самым, хи-квадрат аппроксимация продолжает работать даже в зоне маленьких ошибок.

А вот B — внешность эллипсоида

$$\sum_{i=1}^{k} \frac{(\theta_i - p_i^0)^2}{p_i^0} > c,$$

пересеченная с плоскостью $\sum_{i=1}^k \theta_i = 1$. Здесь минимум также достигается на границе, то есть в некоторой точке $\theta_1^{(0)}, ..., \theta_k^{(0)}$

$$\sum_{i=1}^k \frac{(\theta_i^{(0)} - p_i^0)^2}{p_i^0} = c.$$

Раскладывая Λ в окрестности точки $\theta^{(0)}$ и, оценивая слагаемые вне этой окрестности, мы получим асимптотику типа (мы смотрим только на экспоненциальную часть)

$$\exp\left(-\sum_{i=1}^k \theta_i^{(0)} \ln \frac{\theta_i}{p_i^0} n\right)$$

при некоторых C, $l < k$. Эта вероятность сложна для аналитической работы, но полученный ответ значительно более точен, чем $1 - F_{\chi^2_{k-1}}(an)$. Это и логично, в критерии хи-квадрат мы фактически подменили исходное распределение частот нормальным, а в зоне больших уклонений нормальная аппроксимация работает плохо.

10.2 Критерий Неймана-Пирсона

Пример 3. Рассмотрим критерий Неймана-Пирсона проверки простой гипотезы H_0 (X_i имеют плотность $f_0(x)$) с простой альтернативой H_1 (X_i имеют плотность $f_1(x)$). Допустим у нас есть выборка размера n, тогда критерий Неймана-Пирсона предлагает действовать следующим образом: отвергать гипотезу H_0, если выборка попала в множество D

$$D = \left\{ \frac{f_1(x_1)...f_1(x_n)}{f_0(x_1)....f_0(x_n)} > \gamma^n \right\}$$

и принимать ее в противном случае, при некотором $\gamma \in (0, 1)$.

Рассмотрим ошибку первого рода $\alpha_n = \mathbf{P}_0((X_1, ..., X_n) \in D)$, т.е. вероятность того, что гипотеза была верна, а мы ее отвергли. Как ведет себя α_n при $n \to \infty$?

$$\alpha_n = \mathbf{P}_0((X_1, ..., X_n) \in D) = \mathbf{P}_0\left(\sum_{i=1}^n \ln \frac{f_1(X_i)}{f_0(X_i)} \geq n \ln \gamma\right).$$

Для $Y_1 = \ln \frac{f_1(X_1)}{f_0(X_1)}$ при гипотезе H_0

$$R(h) = \mathbf{E}_0 e^{h \ln \frac{f_1(X_i)}{f_0(X_i)}} = \int_{\mathbb{R}} e^{h \ln \frac{f_1(X)}{f_0(x)}} f_0(x) dx = \int_{\mathbb{R}} f_1^h(x) f_0^{1-h}(x) dx.$$

В частности, $R(1) = 1$.

При этом

$$\mathbf{E}_0 Y_1 = \int_{\mathbb{R}} \ln \frac{f_1(x)}{f_0(x)} f_0(x) dx \le \ln \left(\int_{\mathbb{R}} \frac{f_1(x)}{f_0(x)} f_0(x) dx \right) = 0.$$

Аналогичным образом $\mathbf{E}_1 Y_1 \ge 0$. При этом оба неравенства строгие, если только Y_0 не является вырожденной величиной по мере одной из мер. Будем считать, что это не так.

Следовательно, $\mathbf{E}_0 Y_1 = \mu < 0$, $\mathbf{E}_1 Y_1 = \widetilde{\mu} > 0$ (будем считать, что $\mu > -\infty$, $\widetilde{\mu} < \infty$). При $\gamma < \mu$ α_n сходится к 1. При $\gamma > \mu$, $\mu < \ln\gamma < \widetilde{\mu}$ в силу теоремы Петрова

$$\alpha_n \sim \frac{C(\ln\gamma)}{h_{\ln\gamma}\sqrt{n}} \exp(-\Lambda(\ln\gamma)n).$$

Здесь

$$C(\theta) = \frac{1}{\sqrt{2\pi}\sigma(h_\theta)},\ \Lambda(\theta) = \theta h_\theta - \ln R(h_\theta),$$

где

$$h_\theta : \int_{\mathbb{R}} f_1^{h_\theta}(x) f_0^{1-h_\theta}(x)(\ln f_1(x) - \ln f_0(x) - \theta) dx = 0.$$

Аналогично для ошибки второго рода

$$\beta_n = \mathbf{P}_1((X_1, ..., X_n) \notin D) = \mathbf{P}_1 \left(\sum_{i=1}^{n} \ln \frac{f_1(X_i)}{f_0(X_i)} < n \ln\gamma \right).$$

Тогда при $-1 \le h < 0$

$$\widetilde{R}(h) = \mathbf{E}_1 e^{hY_1} = \int_{\mathbb{R}} f_1^{1+h}(x) f_0^{h}(x) dx = R(1+h),$$

$\widetilde{R}(-1) = 1$, $\widetilde{\mu} = \mathbf{E}_1 Y_1 < \infty$. При этом

$$\widetilde{m}(h) = m(1+h),\ \widetilde{\sigma}^2(h) = \sigma^2(1+h),\ \widetilde{h}_\theta : h_\theta - 1$$

при $\theta \in (\mu, \widetilde{\mu})$, откуда

$$\widetilde{\Lambda}(\theta) = \Lambda(\theta) - \theta.$$

Аналогично предыдущему

$$\beta_n \sim \frac{\widetilde{C}(\ln\gamma)}{(1 - h_{\ln\gamma})\sqrt{n}} \exp(-\widetilde{\Lambda}(\ln\gamma)n)$$

при $\mu < \ln\gamma < \widetilde{\mu}$. Таким образом, при всех $\ln\gamma \in (\mu, \widetilde{\mu})$

$$\alpha_n \sim \frac{C(\ln\gamma)}{h_{\ln\gamma}\sqrt{n}} \exp(-\Lambda(\ln\gamma)n), \;\; \beta_n \sim \frac{C(\ln\gamma)}{(1-h_{\ln\gamma})\sqrt{n}} \exp(-\Lambda(\ln\gamma)n + n\ln\gamma).$$

На следующей лекции мы применим полученные результаты для поиска критерия с минимальной асимптотической суммой ошибок.