Глава 10

Сложные гипотезы и сложные альтернативы

10.1 Базовая часть

10.1.1 Общий подход к проверке гипотез

Ошибки первого и второго рода

В рамках сегодняшнего занятия мы будем рассматривать обобщения рассматриваемого критерия, позволяющего проверять довольно общие параметрические гипотезы. Итак, пусть $\Theta \subseteq \mathbb{R}^k$, $\Theta_0, \Theta_1 \subset \mathbb{R}^k$ $\Theta_0 \cap \Theta_1 = \emptyset$, мы будем рассматривать гипотезу $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$.

Как и прежде критерий мы будем задавать критическим множеством D в $\mathbb{R}^n$, при попадании выборки в которое гипотеза будет отвергнута. При этом у нас будет уже не одна вероятность ошибки I рода, а несколько. Мы будем сравнивать требовать, чтобы максимальная вероятность ошибки I рода была

$$\sup_{\theta \in \Theta_0} \mathbf{P}_\theta((X_1, \ldots, X_n) \in D) = \alpha,$$

где $\alpha \in (0, 1)$ – заданный параметр, называемый уровнем значимости критерием. Функция мощности

$$\beta(\theta) = \mathbf{P}_\theta((X_1, \ldots, X_n) \in D), \quad \theta \in \Theta_1,$$

при этом будет уже не числом, а функцией, поскольку Θ_1 уже не одноточечное множества.

В случае сложных гипотез и сложных альтернатив РНМ критерий зачастую построить невозможно и наша первоочередная цель – научиться строить какие-нибудь критерии с нужным уровнем значимости, желательно несмещенные и состоятельные:

$$\mathbf{P}_\theta((X_1, \ldots, X_n) \in D) \geq \alpha, \ \theta \in \Theta_1, \quad \mathbf{P}_\theta((X_1, \ldots, X_n) \in D) \to 0, \ \theta \in \Theta_1, \ n \to \infty.$$

Статистика критерия

Чаще всего критическое множество задается соотношением $\{\vec{x} : T(\vec{x}) > c\}$ для некоторого c. Статистика T при этом называется статистикой критерия. Параметр c мы определяем, как и прежде, исходя из вероятности ошибки I рода:

$$\sup_{\theta \in \Theta_0} \mathbf{P}_\theta(\vec{X} : T(\vec{X}) > c) = \alpha.$$

Зачастую мы строим критерий из общих соотношений:

- Ищем характеристику выборки $T(\vec{X})$, "склонную" принимать меньшие значения при верной гипотезе и большие при неверной;
- Ищем при каком $\theta_0 \in \Theta_0$ $F_{T(\vec{X});\theta}(x)$ минимально (в идеале такое θ_0 не зависит от x) ;

- Находим из соображения $F_{T(\vec{X});\theta}(c) = 1 - \alpha$.

Пример 1. Пусть $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$, где $X_i \sim \mathcal{N}(\theta, 1)$. Тогда естественно взять в качестве статистики критерия $T(x_1, \ldots, x_n) = x_1 + x_2 + \cdots + x_n$ – это функция достаточной статистики, которая "меньше" при гипотезе и "больше" при альтернативе. При этом

$$\mathbf{P}_\theta(T(X_1, \ldots, X_n) \leq x) = \mathbf{P}_\theta\left(\frac{1}{\sqrt{n}}\left(T(X_1, \ldots, X_n) - \theta n\right) \leq \frac{1}{\sqrt{n}}(x - \theta n)\right) = \Phi\left(\frac{x - \theta n}{\sqrt{n}}\right)$$

монотонно убывает по θ. Значит,

$$\sup_{\theta \leq \theta_0} \mathbf{P}_\theta(T(X_1, \ldots, X_n) > c) = \mathbf{P}_{\theta_0}(T(X_1, \ldots, X_n) > c).$$

При этом c выбирается из соображений

$$\mathbf{P}_{\theta_0}(T(X_1, \ldots, X_n) > c) = \Phi\left(\frac{c - \theta_0 n}{\sqrt{n}}\right) = \alpha,$$

откуда получаем критерий

$$\sum_{i=1}^{n} X_i > \theta_0 n + z_{1-\alpha}\sqrt{n}.$$

Этот критерий имеет вероятность ошибки первого рода α, нетрудно убедиться, что он несмещенный и состоятельный. В действительности, это тот самый РНМ критерий, который мы могли получить с помощью материала факультатива прошлого семинара.

10.1.2 Критерий обобщенного отношения правдоподобий

В случае сложной гипотезы нам достаточно, чтобы хотя бы при одном параметре из Θ_0 правдоподобие оказалось достаточно большим. Таким образом, разумно в качестве статистики критерия взять величину

$$\lambda(x_1, ..., x_n) = \frac{\sup_{\theta \in \Theta} L(x_1, ..., x_n; \theta)}{\sup_{\theta \in \Theta_0} L(x_1, ..., x_n; \theta)}.$$

Тогда в качестве критического множества рассмотрим $D_c = \{\lambda(x_1, ..., x_n) > c\}$. Выбирая c (если это удастся) так, что $P_{\theta_0}(D_c) \leq \alpha$ при всех $\theta_0 \in \Theta_0$, мы получим критерий, который называют критерием отношения правдоподобий (к.о.п.). Этот критерий уже не РНМ в общем случае, но, тем не менее, вполне разумен.

Пример 2. Рассмотрим задачу проверки $H_0 : \theta = \theta_0$ с альтернативой $H_1 : \theta \neq \theta_0$, где $X_i \sim \mathcal{N}(\theta, 1)$. Тогда

$$L(x_1, ..., x_n; \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2}.$$

Максимальное значение L при $\theta \in \Theta$ достигается при подстановке ОМП $\hat{\theta}(x_1, ..., x_n) = \sum_{i=1}^n x_i$, при этом правдоподобие принимает вид

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Максимальное значение L при $\theta \in \Theta_0$ достигается при $\theta = \theta_0$ и равно

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \theta_0)^2}.$$

Таким образом,

$$\lambda(x_1,...,x_n)=\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n(x_i-\overline{x})^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n(x_i-\theta_0)^2\right)}=\exp\left(\frac{1}{2}n(\overline{x}-\theta_0)^2\right)>c,$$

где мы воспользовались формулой

$$\frac{1}{n}\sum_{i=1}^n(x_i-\theta_0)^2=\frac{1}{n}\sum_{i=1}^n(x_i-\overline{x})^2+(\theta_0-\overline{x})^2,$$

представляющей формулу $S^2=\overline{Y^2}-\overline{Y}^2$ для $Y_i=X_i-\theta_0$.

Таким образом, мы можем преобразовать критическое множество к виду

$$D_{\tilde{c}}=\{x_1,...,x_n:|\overline{x}-\theta_0|>\tilde{c}\},$$

где

$$P_{\theta_0}(D_{\tilde{c}})=P(|Z|>\tilde{c})=2(1-\Phi(\tilde{c}\sqrt{n}))=\alpha,$$

откуда критерий принимает вид

$$|\overline{x}-\theta_0|>z_{1-\alpha/2}n^{-1/2}.$$

10.1.3 Асимптотический критерий

Зачастую найти в явном виде вероятность $P_\theta(D_c)$ затруднительно, поэтому нам бы хотелось получить некоторый общий результат о предельном распределении статистики отношения правдоподобий. Оказывается, что верна следующая теорема

Теорема 1 (Wilks)**.** *Пусть модель сильно регулярна (то есть выполнены те же условия, что для асимптотической нормальности ОМП), Θ_0 является многообразием размерности l, пространство Θ имеет размерность k. Тогда при всех $\theta\in\Theta_0$ $2\ln\lambda$ сходится по распределению к величине χ^2_{k-l}.*

Эту теорему достаточно сложно найти в аккуратной форме, поэтому приведу ссылку: D. Williams, Weighing the odds: a course in probability and statistics, p.345 (формулировка), p.377 (схема доказательства).

Таким образом, критерий будет иметь вид $2\ln\lambda>x_{1-\alpha}$.

Пример 3. Проверим гипотезу $H_0:\theta_1=\mu$ с общей альтернативой, $X_i\sim\mathcal{N}(\theta_1,\theta_2^2)$. Тогда

$$L(x_1,...,x_n;\theta_1,\theta_2)=\left(\frac{1}{\sqrt{2\pi}\theta_2}\right)^n\exp\left(-\frac{1}{2\theta_2^2}\sum_{i=1}^n(x_i-\theta_1)^2\right).$$

Тогда максимум по всем Θ правдоподобия будет при ОМП $\theta_1=\overline{x}$, $\theta_2=S$ и будет

$$\left(\frac{1}{\sqrt{2\pi}S}\right)^n\exp\left(-\frac{1}{2S^2}\sum_{i=1}^n(x_i-\overline{x})^2\right)=\left(\frac{1}{\sqrt{2\pi}S}\right)^n\exp(-n/2).$$

При $\theta_1=\mu$ ОМП будет иметь вид $\theta_2=\sqrt{\frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2}$, откуда максимум правдоподобия будет иметь вид

$$\left(\frac{1}{\sqrt{2\pi\frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2}}\right)^n\exp\left(-\frac{1}{2\frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2}\sum_{i=1}^n(x_i-\mu)^2\right)=\left(\frac{1}{\sqrt{2\pi\sum_{i=1}^n(x_i-\mu)^2}}\right)^n\exp(-n/2).$$

Тогда отношение правдоподобий λ имеет вид

$$\lambda(x_1,...,x_n)=\frac{\left(\frac{1}{\sqrt{2\pi\sum_{i=1}^n(x_i-\mu)^2}}\right)^n\exp(-n/2)}{\left(\frac{1}{\sqrt{2\pi}S}\right)^n\exp(-n/2)}=\left(\frac{\sum_{i=1}^n(x_i-\mu)^2}{\sum_{i=1}^n(x_i-\overline{x})^2}\right)^{n/2}.$$

Можно явно построить соответствующий критерий, чем мы займемся чуть позже, когда будем заниматься нормальными моделями. Сейчас же воспользуемся тем, что модель сильно регулярна и значит асимптотический критерий имеет вид

$$2\ln\lambda(x_1,...,x_n)=\ln\left(\frac{\sum_{i=1}^n(x_i-\mu)^2}{\sum_{i=1}^n(x_i-\overline{x})^2}\right)>x_{1-\alpha},$$

где $x_{1-\alpha}$ — квантиль χ^2_1.

10.2 Факультатив

10.2.1 Проверка независимости к.о.о.п.

Метод отношения правдоподобий достаточно мощен и позволяет решать достаточно общие задачи:

Пример 4. Пусть (X_i,Y_i), $i\le N$, — независимые одинаково распределенные случайные векторы, X_i принимают значения x_i, $i=1...k$, Y_i — значения y_j, $j=1...l$. Рассмотрим гипотезу $H_0: X_i$ и Y_i независимы с общей альтернативой. Тогда в общем случае правдоподобие имеет вид

$$L(x_1,y_1,...,x_n,y_n;p_{i,j},i\le k,j\le l)=\prod_{i=1}^k\prod_{j=1}^l p_{i,j}^{N_{i,j}},$$

где $N_{i,j}=\sum_{m=1}^N I_{X_m=x_i,Y_m=y_j}$ — число пар наблюдений каждого вида, $p_{i,j}=P(X=x_i,Y=y_j)$. Решив задачу 6.1.2, можно увидеть, что ОМП здесь будет $\hat{p}_{i,j}=N_{i,j}/N$. При выполнении гипотезы найдутся такие $p_i=P(X=x_i)$, $i\le k$, $q_j=P(Y=y_j)$, $j\le l$, что $p_{i,j}=p_iq_j$, откуда

$$L(x_1,y_1,...,x_n,y_n;p_i,i\le k,q_j,j\le l)=\prod_{i=1}^k\prod_{j=1}^l p_i^{N_{i,j}}q_j^{N_{i,j}}=\prod_{i=1}^k p_i^{N_{i,\cdot}}\prod_{j=1}^l q_j^{N_{\cdot,j}},$$

где $N_{i,\cdot}=\sum_{j=1}^l N_{i,j}$, $N_{\cdot,j}=\sum_{i=1}^k N_{i,j}$. Тогда поиск ОМП по p_i и q_j призводится независимо и из тех же рассуждений ОМП будет $\hat{p}_i=N_{i,\cdot}/N$, $\hat{q}_j=N_{\cdot,j}/N$.

Таким образом, статистика критерия правдоподобий будет иметь вид

$$\lambda(x_1,...,x_n)=\prod_{i=1}^k\prod_{j=1}^l\frac{\left(\frac{N_{i,j}}{N}\right)^{N_{i,j}}}{\left(\frac{N_{i,\cdot}N_{\cdot,j}}{N^2}\right)^{N_{i,j}}}=\prod_{i=1}^k\prod_{j=1}^l\left(\frac{N_{i,j}N}{N_{i,\cdot}N_{\cdot,j}}\right)^{N_{i,j}},$$

откуда критерий принимает вид

$$\sum_{i=1}^k\sum_{j=1}^l N_{i,j}\ln\left(\frac{N_{i,j}N}{N_{i,\cdot}N_{\cdot,j}}\right)>x_{1-\alpha},$$

где $x_{1-\alpha}$ — квантиль распределения $\chi^2_{kl-1-(k-1+l-1)}=\chi^2_{(k-1)(l-1)}$, поскольку пространство $p_{i,j}$, $\sum_{i=1}^k\sum_{j=1}^l p_{i,j}=1$ имеет размерность $kl-1$, а $p_{i,j}=p_iq_j$, где $p_i:\sum_{i=1}^k p_i=1$, $q_j:\sum_{j=1}^l q_j=1$ — многообразие размерности $k-1+l-1$.

10.2.2 Хи-квадрат аппроксимация к.о.о.п.

Рассмотрим родственное семейство критериев. Начнем с проверки для дискретных X_i с $P(X_i = x_j) = p_j$, $j \le k$, гипотезы $H_0 : p_i = p_i^0$, $i \le k$ с общей альтернативой, где p_i^0 — заданные вероятности. Тогда критерий отношения правдоподобий предписывает рассматривать статистику

$$\lambda = \prod_{i=1}^{k} \frac{\left(\frac{N_i}{N}\right)^{N_i}}{(p_i^0)^{N_i}} = \prod_{i=1}^{k} \left(1 + \frac{N_i - np_i^0}{p_i^0}\right)^{N_i},$$

и пользоваться тем, что

$$-2\ln\lambda = 2\sum_{i=1}^{k} N_i \ln\left(1 + \frac{Np_i^0 - N_i}{N_i}\right)$$

имеет асимптотическое распределение χ^2_{k-1} (условия регулярности здесь выполнены автоматически). При больших N $(N_i - Np_i^0)$ имеет порядок $\sqrt{N}$, поскольку по центральной предельной теореме

$$\frac{N_i - Np_i^0}{\sqrt{N}} \xrightarrow{d} Z \sim \mathcal{N}(0, p_i^0(1-p_i^0)).$$

Следовательно,

$$2\sum_{i=1}^{k} N_i \ln\left(1 + \frac{Np_i^0 - N_i}{N_i}\right) = 2\sum_{i=1}^{k} N_i \left(\frac{Np_i^0 - N_i}{N_i}\right) - \frac{1}{2}\sum_{i=1}^{k} N_i \left(\frac{Np_i^0 - N_i}{N_i}\right)^2 + \sum_{i=1}^{k} N_i Z_i,$$

где $Z_i = O\left(\left(\frac{Np_i^0 - N_i}{N_i}\right)^3\right)$. При этом

$$\frac{(Np_i^0 - N_i)^3}{N_i^2} = \left(\frac{Np_i^0 - N_i}{\sqrt{N_i}}\right)^2 \frac{(Np_i^0 - N_i)}{N_i} \xrightarrow{d} 0,$$

поскольку величина под знаком квадрата стремится к квадрату $\mathcal{N}(0, p_i^0(1-p_i^0))$ случайной величины по центральной предельной теореме, а вторая к нулю по закону больших чисел. Следовательно, все $N_i Z_i$ стремятся к нулю по распределению.

Кроме того

$$2\sum_{i=1}^{k} N_i \left(\frac{Np_i^0 - N_i}{N_i}\right) = N\sum_{i=1}^{k} p_i^0 - \sum_{i=1}^{k} N_i = 0.$$

Отсюда величина

$$2\ln L = \sum_{i=1}^{k} N_i \left(\frac{Np_i^0 - N_i}{N_i}\right)^2 + \sum_{i=1}^{k} N_i Z_i = \sum_{i=1}^{k} \frac{(Np_i^0 - N_i)^2}{N_i} + \sum_{i=1}^{k} N_i Z_i.$$

Второе слагаемое правой части с ростом N стремится к нулю, а левая часть стремится к χ^2_{k-1} распределению (поскольку пространство $p_1 + .. + p_k = 1$ имеет размерность 1, а одноточечное многообразие — размерность 0). Следовательно,

$$T = \sum_{i=1}^{k} \frac{(Np_i^0 - N_i)^2}{N_i} \xrightarrow{d} Y \sim \chi^2_{k-1}.$$

Полученный критерий $T > x_{1-\alpha}$ называется критерием хи-квадрат. Мы можем его рассматривать как аппроксимацию критерия обобщенного отношения правдоподобий, хотя у вас в курсе он получен напрямую. Эта аппроксимация удобна тем, что вместо неудобных логарифмов мы используем более простую квадратичную функцию.

10.2.3 Сложный критерий хи-квадрат

Аналогичным образом строится критерий хи-квадрат для более сложной гипотезы: $H_0 : p_i = p_i(\theta)$, где p_i — заданные функции, а θ — неизвестный параметр, с общей альтернативой. В этом случае мы, соответственно, строим ОМП $\hat{\theta}$ для θ на основе

$$L = \prod_{i=1}^{k} p_i(\theta)^{N_i}$$

и говорим, что статистика

$$\sum_{i=1}^{k} \frac{(N_i - Np_i(\hat{\theta}))^2}{Np_i(\hat{\theta})}$$

имеет асимптотическое распределение χ^2_{k-1-l}, где l — размерность многообразия $p_1(\theta), ..., p_k(\theta)$, которая в наиболее простом случае совпадает с размерностью θ.

Таким образом, критерий хи-квадрат позволяет проверять гипотезу о том, что вероятности значений нашей дискретной выборки принадлежат некоторому параметрическому семейству с общей альтернативой.

Пример 5. Кубик брошен 600 раз, 100 раз выпала 1, 94 раза 2, 103 — 3, 89 — 4, 110 — 5, 104 — 6. Проверим гипотезу о том, что 4 и 5 выпадают одинаково часто. Будем использовать критерий хи-квадрат. Тогда вероятности параметризуются в виде $p_i(\theta) = \theta_i$, $i = 1, 2, 3, 4, 6$, $p_5(\theta) = \theta_4$. Следовательно, правдоподобие приобретает вид

$$L(x_1, ..., x_n) = \theta_1^{N_1}\theta_2^{N_2}\theta_3^{N_3}(2\theta_4)^{N_4+N_5}\theta_6^{N_6}2^{-N_4-N_5},$$

где $2\theta_4$ мы рассматриваем, чтобы сумма оснований степени давала 1. Как мы уже выясняли, ОМП при этом $\theta_i = N_i/N$, $i = 1, 2, 3, 6$, $2\theta_4 = (N_4 + N_5)/N$. Следовательно, критерий приобретает вид

$$T(x_1, ..., x_n) = \sum_{i=1}^{k} \frac{(N_i - Np_i)^2}{np_i} = \frac{(89 - 199/2)^2}{199/2} + \frac{(110 - 199/2)^2}{199/2} = \frac{10.5^2}{199} \approx 0.55.$$

При этом квантиль χ^2_4 ($k = 6$, $l = 1$) уровня 0.95 приблизительно равна 9.5. Следовательно, гипотеза с уверенностью принимается.

Критерий хи-квадрат можно использовать и для случая не дискретных распределений путем дискретизации данных. Скажем, для простой гипотезы $H_0 : F = F_0$ с общей альтернативой $H_1 : F \neq F_0$ рассматриваются некоторые непересекающиеся диапазоны Δ_i, дающие в объединении всю прямую, наблюдения X_i заменяются на Y_i — номер Δ_j, в который попал X_i, а гипотеза $F = F_0$ заменяется на $P(Y_i = j) = p_j^0$, где p_j^0 соответствует вероятности величины с ф.р. F_0 в Δ_i. На практике число Δ_i берут порядка $[\log_2 n]$, а сами Δ_i выбирают так, чтобы вероятности попадания в них были примерно одинаковыми.