5 Проверка гипотезы о параметрическом семействе

5.1 О сложной гипотезе и сложной альтернативе

Более частая ситуация заключается в том, что у нас есть гипотеза о принадлежности к параметрическому семейству, например, что выборка нормальная, но с неизвестными параметрами.
Можно, конечно, оценить неизвестные параметры состоятельными оценками, но эта процедура может сместить распределение статистик критерия.

5.1.1 Сложный критерий хи-квадрат и критерий отношения правдоподобий

1. *Критерий хи-квадрат*

   Критерий хи-квадрат в этом случае удается модернизировать, если вместо неизвестных параметров подставить оценки ОМП для них.
   Более конкретно, рассмотрим вероятности $p_i(\theta) = \mathbf{P}_\theta(\Delta_i)$. Подсчитаем количества попаданий ν_i в Δ_i. Найдем ОМП для θ по функции правдоподобия

   $$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} p_i(\theta)^{\nu_i}.$$

   Подставив в $p_i(\theta)$ полученную оценку для θ, можно найти новую статистику хи-квадрат. В этом случае можно показать, что полученная статистика будет иметь распределение χ^2_{k-l-1}, где l — размерность параметра θ, k — число Δ_i.

   Отметим, что важным является то, что оценивание проводится именно на основе группированных данных, то есть правдоподобие строится именно по ним. В противном случае теорема о предельном распределении неверна.

   В случае сложного критерия мы не сможем выбирать интервалы Δ_i так, чтобы вероятности попадания в них были одинаковы, поскольку эти вероятности зависит от неизвестного параметра. Зачастую на практике выбирают интервалы так, чтобы в них попадали близкие количества наблюдений. Более правильный (и более эффективный метод группирования) называется асимптотическое оптимальное группирование. В этом случае группировка осуществляется так, чтобы информация Фишера (информационная матрица Фишера) по сгруппированным данным была как можно ближе к информации исходных данных. В случае скалярного параметра эта задача сводится к максимизации количества информации Фишера о параметре по группированной выборке

   $$\max_{x_0<x_1<\dots<x_k} \sum_{i=1}^{k} \left(\frac{\partial \ln p_i(\theta)}{\partial \theta} \right)^2 p_i(\theta),$$

   где $\Delta_i = [x_{i-1}, x_i)$, а в случае многомерного — максимизации определитель информационной матрицы Фишера. Здесь вместо неизвестных параметров используются их ОМП. Граничные точки разбиения для различных распределений представлены здесь.

   В случае нормального распределения критерий хи-квадрат реализован, например, в пакете nortest в функции pearson.test.

2. *Критерий отношения правдоподобий*

   Это более общий метод, позволяющий работать со сложными параметрическими моделями. Критерий хи-квадрат, в действительности, является апрроксимацией этого критерия.
   Можно представлять этот критерий как некоторое обобщение критерия Неймана-Пирсона. Пусть $\theta = (\theta_1, \dots, \theta_r)$ и основная гипотеза

   $$H_0 : (\theta_{q+1}, \dots \theta_r) = (\theta_{0,q}, \dots, \theta_{0,r}),$$

   то есть часть параметров фиксирована, а остальные произвольны. Тогда найдем статистику отно-

шения правдоподобия

$$T = 2\ln\frac{L(x_1,\dots,x_n,\widehat{\theta})}{L(x_1,\dots,x_n,\widehat{\theta}_0)},$$

где $\widehat{\theta}_0$ — ОМП при H_0, $\widehat{\theta}$ — ОМП в общей параметрической модели. Оказывается, при H_0 в так называемых сильно регулярных моделях справедливо соотношение $T \xrightarrow{d} Y \sim \chi^2_{r-q}$, $n \to \infty$, откуда $T > y_{1-\alpha}$ задает асимптотический критерий уровня α. Этот результат называется теоремой Уилкса. Не будем останавливаться подробно на формулировке условий сильной регулярности, отметим лишь ключевые: правдоподобие должно достаточно гладко зависить от параметра и область изменения параметра является открытым множеством
В частности, критерий работает для выборок из дискретного распределения, где параметричность модели уже не требуется.

Вопрос 1. Как будет выглядеть критерий отношения правдоподобий для дискретных выборок с k возможными значениями?

Критерий Никулина-Рао-Робсона.
Предположим, что исходные наблюдения имели плотность $f(x;\theta)$, причем модель была сильно регулярна, $\widehat{\theta}$ — ОМП в этой модели, а $I(\theta)$ — информационная матрица Фишера. Пусть мы сгруппировали данный и нашли $p_j(\theta) = \mathbf{P}_\theta(\Delta_j)$. Подсчитаем количества попаданий ν_i в Δ_i. Положим $z_j(\theta) = (\nu_j - np_j(\theta))/\sqrt{n}$. Пусть $I(\theta)$ — информационная матрица Фишера модели,

$$C_{i,j}(\theta) = \sum_{l=1}^{k} \frac{1}{p_l(\theta)} \frac{\partial}{\partial\theta_i} p_l(\theta) \frac{\partial}{\partial\theta_j} p_l(\theta),$$

пусть $\widehat{A}$ — диагональная матрица с ν_j на диагонали, $\widehat{I} = I(\widehat{\theta})$, $\widehat{C} = C(\widehat{\theta})$,

$$\widehat{V} = \widehat{A} - \widehat{C}^t\widehat{I}^{-1}\widehat{C}.$$

Тогда критерий Никулина-Рао-Робсона предлагает рассматривать статистику

$$z_j(\widehat{\theta})^t\widehat{V}^{-1}z_j(\widehat{\theta}),$$

которая при гипотезе будет иметь асимптотическое распределение χ^2_{r-1}. Этот критерий учитывает использование ОМП в оценке и достаточно эффективен для соответствующих задач, хотя и громоздок для подсчетов.

Вопрос 2. Рассмотрим нормальное распределение $\mathcal{N}(\mu,\sigma^2)$ с неизвестными параметрами. Описать критерии хи-квадрат и отношения правдоподобия в этом случае (не вычисляя в явном виде ОМП по группированным данным)

5.1.2 Критерии Колмогорова и омега-квадрат при сложной гипотезе

1. Критерий Колмогорова-Смирнова также применим к сложной гипотезе при подстановке состоятельных оценок (например, ОМП). Однако в этом случае предельное распределение уже не будет колмогоровским, а будет своим для каждого класса распределений. Это замечание крайне важно и зачастую игнорируется прикладными исследователями.
   Так для нормальных распределений при этом получится так называемое распределение Лиллиефорса (соответствующий критерий называют критерием Лиллиефорса).
   В общем случае можно определить критическое множество с помощью метода Монте-Карло.

2. Критерии Андерсона-Дарлинга и Крамера-Мизеса будут верны и в случае параметрических семейств, но опять-таки предельное распределение станет иным и будет зависеть от распределения выборки.
   Примеры таких критериев можно найти в R, например, EDF_NS.test в пакете EWGoF для экспоненциальности, lillie.test, ad.test, cvm.test для нормальности из nortest.

Задача 1. Смоделировать выборки из $\mathcal{N}(\mu, \sigma^2)$ и проверить их на нормальность с помощью а) критерия Колмогорова-Смирнова с оцененными параметрами б) критерия Лиллефорса. Моделировать 1000 выборок и построить распределение фактического уровня значимости.

5.2 Ответы на вопросы

Ответ 1. Для k возможных значений мы имеем параметрическую гипотезу $H_0 : p_1 = p_1^0, \ldots, p_k = p_k^0$, где $p_1, \ldots, p_{k-1}$ — параметры модели, $p_k = 1 - \cdots - p_{k-1}$, $p_1^0, \ldots, p_k^0$ — известные значения. Тогда при гипотезе H_0 ОМП для параметров это $p_1^0, \ldots, p_{k-1}^0$, поскольку наша модель фактически не имеет неизвестных параметров. В общем случае правдоподобие имеет вид

$$L(x_1, \ldots, x_n) = p_1^{N_1} \cdots p_{k-1}^{N_{k-1}} (1 - p_1 - \cdots - p_{k-1})^{N_k},$$

где N_i — число i-х значений в нашей выборке. Логарифмируя и дифференцируя правдоподобие, мы получим уравнения

$$\frac{N_i}{p_i} - \frac{N_k}{1 - p_1 - \cdots - p_{k-1}} = 0,$$

т.е. $N_i = cp_i$, $i = 1, \ldots, k$ при некотором c. Тогда $N = N_1 + \cdots + N_k = c(p_1 + \cdots + p_k) = c$, то есть оценки будут иметь вид $\widehat{p}_i = N_i/N$. Нетрудно убедиться, что это действительно максимум. Тогда

$$T = 2 \ln \frac{L(x_1, \ldots, x_n, \widehat{p}_1, \ldots, \widehat{p}_k)}{L(x_1, \ldots, x_n, p_1^0, \ldots, p_k^0)} = 2 \sum_{i=1}^{k} \ln \left(\frac{N_i}{N p_i^0} \right)^{N_i} = 2 \sum_{i=1}^{k} N_i \ln \left(\frac{N_i}{N p_i^0} \right).$$

Критерий имеет вид $T > c$, где $c = y_{1-\alpha}$, $y_{1-\alpha}$ — $1-\alpha$-квантиль распределения χ^2_{k-1}.

Ответ 2. Критерии хи-квадрат и отношения правдоподобия будут строиться схожим образом: $T_i > y_{1-\alpha}$, где $y_{1-\alpha}$ есть χ^2_{k-3},

$$T_1 = \sum_{i=1}^{k} \frac{(\nu_i - np_i(\widehat{\theta}_G))^2}{np_i(\widehat{\theta}_G)}, \quad T_2 = 2 \sum_{i=1}^{k} \nu_i \ln \left(\frac{\nu_i}{np_i(\widehat{\theta}_G)} \right),$$

где $\widehat{\theta}_G = (\mu_G, \sigma_G)$ находится из соотношения

$$\prod_{i=1}^{k} p_i(\theta) \to \max, \quad p_i(\theta) = p_i(\mu, \sigma) = \Phi \left(\frac{b_{i+1} - \mu}{\sigma} \right) - \Phi \left(\frac{b_i - \mu}{\sigma} \right).$$