2 Проверка принадлежности параметрическому семейству

Мы будем рассматривать сложную гипотезу

$$H_0 : F \in \{F_\theta, \theta \in \Theta\}$$

с общей альтернативой, где $\{F_\theta\}$ – некоторое семейство распределений (например, нормальные распределения с неизвестными параметрами).

Сперва предложим общие подходы к такой задаче, а потом приведем частные случаи для двух популярных семейств.

2.1 Сложные критерии Крамера-фон Мизеса, Андерсона-Дарлинга и Колмогорова

Все три рассмотренных в прошлый критерии естественно модифицировать для проверки нашей гипотезы, заменив F_0 (которого мы теперь точно не знаем) на $F_{\widehat{\theta}}$, где $\widehat{\theta}$ – ОМП для θ в нашем семействе:

$$T_{KS} = \sup \left| \widehat{F}_n(x) - F_{\widehat{\theta}}(x) \right|,$$
$$T_{CvM} = \int_{\mathbb{R}} (\widehat{F}_n(x) - F_{\widehat{\theta}}(x))^2 dF_{\widehat{\theta}}(x),$$
$$T_{AD} = \int_{\mathbb{R}} \frac{(\widehat{F}_n(x) - F_{\widehat{\theta}}(x))^2}{F_{\widehat{\theta}}(x)(1 - F_{\widehat{\theta}}(x))} dF_{\widehat{\theta}}(x),$$

Все три статистики с теми же нормировками, что и раньше, в сильно регулярных абсолютно-непрерывных параметрических семействах будут сходиться к некоторому распределению. Однако теперь уже не к тем распределениям, что раньше, да и вообще к разным распределениям для разных семейств F_θ, но что хуже всего – может быть даже к разным распределениям при разных θ в рамках одного параметрического семейства. Это практически лишает нас возможности использовать эти подходы.

Впрочем, в семействах сдвига-масштаба все значительно лучше – точные распределения статистик не зависят от θ (но зависят от самого семейства), существуют предельные распределения для статистик с теми же нормировками (в случае сильно регулярных семейств), но предельные распределения станут другими и вновь будут зависеть от семейства.

Напомним, что семейства сдвига-масштаба имеют вид

$$F_\theta(x) = F_0\left(\frac{x - \theta_1}{\theta_2}\right),$$

где параметры необязательно независимы (а могут быть функциями одного параметра), а F_0 – известная ф.р.

Таким образом, мы можем строить критерии такого типа для отдельных семейств сдвига-масштаба, а для определения фактического уровня значимости использовать метод Монте-Карло, генерируя вспомогательные выборки из любого представителя нашего семейства.

2.2 Сложные критерии отношения правдоподобий и хи-квадрат

При использовании критерия хи-квадрат или к.о.п. для дискретных данных мы можем безболезненно подставлять ОМП в статистику критерия, используя теорему Уилкса или теорему о распределении статистики критерия множителей Лагранжа:

$$\sum_{i=1}^{k} \frac{(\nu_i - np_i(\widehat{\theta}))^2}{np_i(\widehat{\theta})^2} > y_{1-\alpha}, \quad \sum_{i=1}^{n} \nu_i \ln \frac{\nu_i}{np_i(\widehat{\theta})} > y_{1-\alpha},$$

где y – квантиль χ^2_{k-l}, где мы предполагаем, что поверхность $(p_1(\theta), \ldots, p_k(\theta))$ образует гладкое многообразие размерности l.

ОМП при этом определяется максимизацией правдоподобия

$$\prod_{i=1}^{k} p_i(\theta)^{\nu_i}. \tag{1}$$

Чтобы использовать этот критерий в непрерывном случае (впрочем, для дискретных с большим числом значений придется делать то же самое), мы дискретизируем данные и применяем к ним предыдущий критерий. Главная проблема в том, что ОМП $\widehat{\theta}$ необходимо рассчитывать максимизацией дискретного (!) правдоподобия (1). Таким образом, например, для нормального распределения мы вынуждены оценивать неизвестные параметры μ, σ, максимизируя

$$\prod_{i=0}^{k-1} \left(\Phi\left(\frac{a_{i+1}-\mu}{\sigma}\right) - \Phi\left(\frac{a_i-\mu}{\sigma}\right)\right)^{\nu_i},$$

где $a_0 = -\infty$, $a_k = +\infty$, a_i – точки разбиения прямой. Эту максимизацию сложно производить аналитически.

Второй проблемой является выбор деления на отрезки. Раньше мы могли выбирать отрезки так, что теоретическое распределение всех отрезков было одинаковым. Теперь мы вынуждены выбирать их глядя на данные, а ведь этот может нарушить работу критериев. Впрочем, существуют работы, в которых показано, что, например, деление выборки на равные части не изменяет предельного распределения статистики.

2.3 Три критерия проверки нормальности

Для часто возникающей проверки нормальности три, на наш взгляд, наиболее популярных критерия это критерий Андерсона-Дарлинга (с оговоренными выше поправками), критерий Шапиро-Уилка и К-критерий Д'Агостино.

1. Критерий Шапиро-Уилка реализован в scipy.stats.shapiro. Это один из лучших тестов для данной задачи с точки зрения многих исследователей. Он базируется на отношении

   $$T_{SW} = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

   где a_i – достаточно сложные коэффициенты:

   $$\vec{a} = \frac{m^T V^{-1}}{||m^T V^{-2} m||^{1/2}},$$

   где $(m_1, \ldots, m_n)$ – средние порядковых статистик из $\mathcal{N}(0,1)$ выборки, V – их ковариационная матрица. Распределение этой статистики ищется методом Монте-Карло.

2. Критерий Д'Агостино реализован в scipy.stats.normaltest. Идея этого и ряда других критериев (в том числе других версий критерий Д'Агостино) отличия нормального распределения от других ищутся на основе коэффициентов асимметрии

   $$\widehat{\mu}_3 = \frac{\overline{(X - \overline{X})^3}}{S^3}$$

   и эксцесса

   $$\widehat{\mu}_4 = \frac{\overline{(X - \overline{X})^4}}{S^4} - 3,$$

   где S – выборочная дисперсия. Критерий Д'Агостино преобразует эту пару статистик в некоторую

достаточно хитрую функцию вида

$$T_{DA} = C_1((C_2\widehat{\mu}_3))^2 + C_3\left(1 - \frac{1}{(C_4 + C_5\widehat{\mu}_4)^{1/3}}\right)^2,$$

которая имеет предельное χ^2_2 распределение. Здесь C_i – некоторые константы. Эти хитрые преобразования каждой из величин преобразованиями приближают их распределения к нормальному.

2.4 Четыре критерия проверки экспоненциальности

Для проверки другой частой гипотезы – гипотезы экспоненциальности, мы назовем четыре критерия. Первые два из них зачастую называются наиболее эффективными, а два других мы выбрали на основе обзора по ссылке). Итак, мы рассмотрим критерий Андерсона-Дарлинга (со своей поправкой), критерий Шапиро-Уилка для экспоненциальности, критерий Жанга и критерий Фрозини.

1. Критерий Андерсона Дарлинга использует привычную нам статистику, подставляя в качестве эталонного распределение экспоненциальное с оцененным параметром $1/\overline{X}$. При этом распределение статистики будет специфическим именно для экспоненциального распределения и квантили придется рассчитывать отдельно.

2. Критерий Шапиро-Уилка базируется на статистике

$$T_{SW,exp} = \frac{n(\overline{X} - X_{(1)})^2}{(n-1)\sum_{j=1}^n (X_j - \overline{X})^2}.$$

3. А-критерий Жанга имеет статистику

$$-\sum_{j=1}^n \left(\frac{\ln Z_{(j)}}{n - j + 1/2} + \frac{\ln(1 - Z_{(j)})}{j - 1/2}\right),$$

где $Z_j = F_{\widehat{\theta}}(X_j)$.

4. Критерий Фрозини использует статистику

$$T_{Fr} = \frac{1}{\sqrt{n}}\sum_{j=1}^n \left|1 - \exp\left(\frac{X_{(j)}}{\overline{X}}\right) - \frac{j - 0.5}{n}\right|.$$

Увы, ни одного из них в Python не реализовано.

2.5 Задачи

Пункт а, б или в определяется остатком по модулю 3 от длины фамилии (0, 1 или 2 соответственно), а i) или ii) – четностью длины имени (i четно, а ii – нечетно).

1. Реализовать сложные критерии а) Колмогорова, б) Крамера-фон Мизеса, в) Андерсона-Дарлинга для проверки принадлежности i) нормальному ii) экспоненциальному закону, используя метод Монте-Карло для вычисления предельного распределения p-value. Построить ЭФР p-value при гипотезе и при альтернативе, моделируя данные из распределения Стьюдента t_5 в первом случае и χ^2_2 во втором.

2. Реализовать критерий хи-квадрат для проверки i) нормальности, ii) экспоненциальности (ОМП по сгруппированным данным находить численно). Для разбиения на промежутки использовать следующие подходы: 1) делим так, чтобы в ячейки попадало поровну наблюдений, 2) берем квантили уровней i/k теоретического распределения, в качестве неизвестного параметра используем ОМП. Для подсчета p-value использовать предельное распределение.

3. Исследовать сгенерированные модельные данные X на нормальность, где

   - X_i имеют распределение Вейбулла с параметром формы λ;
   - X_i имеют χ^2_n распределение, $n = 5, 10, 25$
   - $X_i \sim Binom(n, 1/2)$.

   i) Использовать критерий из первого пункта, ii) использовать встроенный критерий Андерсона-Дарлинга (scipy.stats.anderson), а также критерии Шапиро-Уилко (scipy.stats.shapiro) и Д'Агостино (scipy.stats.normal Построить ЭФР p-value и график мощности. Подобрать размер выборки так, чтобы сравнение было осмысленным.

4. * Исследовать сгенерированные модельные данные X на экспоциальность, где

   (a) X_i имеют распределение $|\mathcal{N}(\mu, 1)|$;

   (b) X_i имеют χ^2_n распределение;

   (c) $X_i \sim Geom(p)$.

   Построить ЭФР p-value при фиксированных параметрах и график мощности в зависимости от параметра. Подобрать размер выборки и параметр так, чтобы сравнение было осмысленным. Использовать i) встроенный критерий Андерсона-Дарлинга ii) критерий из первого пункта, а также критерии а) Шапиро-Уилка, б) Жанга и в) Фрозини.