# 1 Гипотеза согласия

Рассмотрим простую гипотезу $H_0 : F = F_0$, где $F_0$ – заданное распределение. Альтернативу будем предполагать общей $H_1 : F \neq F_0$.

Критерий Колмогорова для простой гипотезы $H_0 : F = F_0$ с общей альтернативой имеет статистику

$$T_K = \sup_x |\widehat{F}_n(x) - F_0(x)|,$$

где $\widehat{F}_n(x)$ – ЭФР. Можно представить статистику в форме

$$T_K = \max\left(\frac{i}{n} - F_0(x_{(i)}), F_0(x_{(i)} - 0) - \frac{i-1}{n}\right).$$

При непрерывном распределении $F_0$ и верной гипотезе $T_K$ имеет некоторое фиксированное распределение, не зависящее от $F_0$. При $n \to \infty$ величина $\sqrt{n}T_K$ сходится к некоторому распределению, называемому распределением Колмогорова. В Python его квантили есть в kolmogi.

Критерий Крамера-фон Мизеса для простой гипотезы $H_0 : F = F_0$ с общей альтернативой имеет статистику

$$T_{CvM} = \int_{\mathbb{R}} (\widehat{F}_n(x) - F_0(x))^2 dF_0(x),$$

где $\widehat{F}_n(x)$ – ЭФР. Это выражение неудобно и используют явную формулу

$$T_{CvM} = \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F_0(X_{(i)})\right)^2 + \frac{1}{12n}.$$

Критерий Андерсона-Дарлинга имеет статистику

$$T_{AD} = \int_{\mathbb{R}} \frac{(\widehat{F}_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x),$$

где $\widehat{F}_n(x)$ – ЭФР. Это выражение неудобно и используют явную формулу

$$T_{AD} = -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\ln F_0(X_{(i)}) + \ln(1 - F_0(X_{(n+1-i)}))).$$

В случае непрерывного распределения $F_0$ обе статистики имеют некоторые распределения, не зависящия от $F_0$, причем при $n \to \infty$ верны соотношения

$$nT_{CvM} \xrightarrow{d} U, \quad nT_{AD} \xrightarrow{d} V,$$

где

$$U = \sum_{j=1}^{\infty} \frac{Z_j^2}{j^2\pi^2}, \quad V = \sum_{j=1}^{\infty} \frac{Z_j^2}{j(j+1)},$$

где $Z_j$ н.о.р. величины со стандартным нормальным распределением.

Еще один знакомый вам критерий – хи-квадрат, использующий статистику

$$T_\chi = \sum_{i=1}^{k} \frac{(\nu_i - nP_0(\Delta_i))^2}{nP_0(\Delta_i)},$$

где $P_0$ – мера, соответствующая $F_0$, $\nu_i$ – число наблюдений, попавших в $\Delta_i$, $\Delta_i$ – разбиение прямой. Здесь ключевой вопрос в выборе $\Delta_i$ и $k$. Обычно предлагают $k = [log_2 n]$ или $k = [n^{1/5}]$. Что касается $\Delta_i$, то их стараются выбрать так, чтобы $P_0(\Delta_i)$ были близки, например, равными.

При верной гипотезе статистика сходится к распределению $\chi^2_{k-1}$, откуда получаем критерий.

Фамилии, начинающиеся с буквы до К включительно решают пункт а), а после К – пункт б).

1. Реализовать критерии а) Крамера-фон Мизеса и б) Андерсона-Дарлинга, определяя $p$-value с помощью метода Монте-Карло. Построить график ЭФР p-value при верной гипотезе и проверить, что критерий работает верно. Учтите, что anderson из scipy.stats – это другое!

2. Построить графики $p$-value критериев Колмогорова, хи-квадрат, а) Крамера-фон Мизеса, б) Андерсона-Дарлинга для проверки гипотезы $H_0 : X_i \sim \mathcal{N}(0,1)$ для а) $X_i \sim p\mathcal{N}(0,1) + (1-p)\mathcal{N}(0,3)$ (под суммой имеется ввиду смесь) б) $X_i \sim X_i \sim 0.5\mathcal{N}(\mu,1) + 0.5\mathcal{N}(-\mu,1)$, $p = 0.9$, $\mu = 0.1$. Подобрать $n$ так, чтобы все критерии были чувствительны к гипотезе (то есть график p-value существенно отличался от биссектрисы, но не становился вертикальным). Какой критерий лучше справляется с задачей?

3. * Для предыдущей задачи построить график мощности всех четырех критериев уровня 95% как функции от параметра $p$ или $\mu$ соответственно.