СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 13. Задачи непараметрической статистики. Часть 1

Введение

Тема лекции – исследование устойчивости и чувствительности основных непараметрических критериев согласия при динамическом изменении моделей распределений данных.

В задачах статистики больших данных наблюдаются **изменения моделей распределения данных** во времени. Например, при анализе спроса или при исследовании периодичности природных процессов.

Изменения характера распределения выборок сильно осложняет их анализ. Например, при построении регрессионных моделей гипотеза о нормальности и постоянстве параметров распределения остатков может приниматься или отклоняться в зависимости от рассматриваемого периода времени.

При анализе данных важно учитывать **наличие выбросов** и их влияние на **устойчивость статистических выводов**.

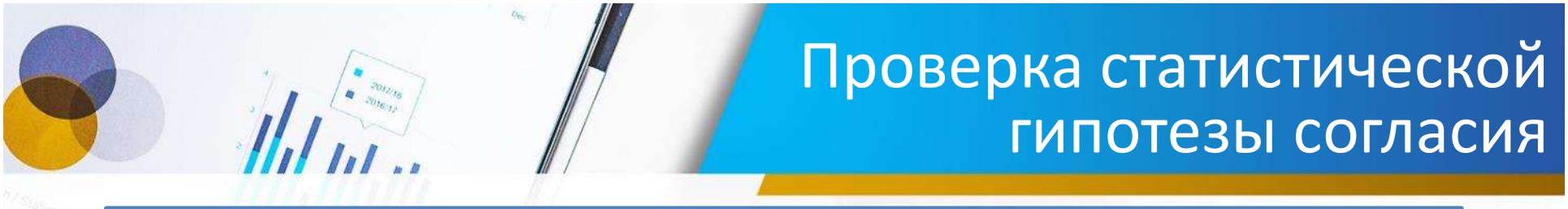
Введение

Далее приведены результаты исследования устойчивости решений ряда задач проверки простой гипотезы согласия. При этом рассматривается несколько моделей изменения распределения данных:

- динамическое изменение смеси распределений;
- изменение параметра формы обобщённого гауссовского распределения;
- изменения числа степеней свободы для распределения Стьюдента.

Исследовались следующие **непараметрические критерии согласия**: хи-квадрат Пирсона, Колмогорова, Крамера-Мизеса и Андерсона-Дарлинга.

Напомним постановку статистической задачи проверки гипотезы согласия и основные критерии согласия.



Проверка статистической гипотезы согласия

Статистическая гипотеза – любое утверждение о виде или свойствах распределения наблюдаемых значений случайных величин.

Пусть у нас имеется некоторая основная гипотеза о распределении вероятностей – будем называть её ***нулевой гипотезой*** и обозначать H_0.

Гипотеза простая, если она однозначно определяет функцию распределение данных. Иначе она называется **сложной.**

Примеры задач проверки статистических гипотез:

- проверка однородности наблюдений,
- проверка независимости,
- проверка наличия выбросов.

Здесь рассмотрена задача проверки гипотезы о том, что реальное распределение выборки согласуются с заданной моделью.

Критерии согласия

Пусть заданы **нулевая гипотеза** H_0: $F_\xi(x) = F_0(x)$

и **альтернативная гипотеза** H_1: $F_\xi(x) \neq F_0(x)$ (сложная гипотеза).

Статистический тест критерия – однозначное правило, согласно которому принимается или отвергается гипотеза.

Ошибка первого рода – вероятность отвергнуть нулевую гипотезу при её справедливости: $P\{H_1|H_0\}$

Задается **уровень значимости** α : $P\{H_1|H_0\} \leq \alpha$. $(\alpha = 0.01,\ 0.05,\ 0.1)$

Ошибка второго рода – вероятность принять нулевую гипотезу при справедливости альтернативы H_1: $P\{H_0|H_1\}$.

Мощность критерия – вероятность принять альтернативу H_1 при ее справедливости: $W = 1 - P\{H_0|H_1\}$

Критерии согласия

Основной идеей построения критериев согласия является оценка расстояния эмпирического распределение $F_n(x)$ от распределения $F_0(x)$ в смысле некоторой метрики $d(F_n, F_0)$.

Если расстояние $d(F_n, F_0)$ меньше заданного порога, то принимается нулевая гипотеза H_0; в противном случае принимается альтернативная гипотеза H_1.

Заданный порог определяется по принятому уровню значимости α.

Все рассматриваемые далее критерии отличаются друг от друга только выбором метрики $d(F_n, F_0)$.

Критерии согласия

Критерий Колмогорова: $d(F_n, F_0) = D_n(F_n, F_0) = \sup_{|x|<\infty} |F_n(x) - F_0(x)|$

Согласно теореме Колмогорова для некоторого порогового значения $t > 0$ выполнено: $\lim_{n\to\infty} P\{\sqrt{n}D_n < t\} = K(t)$, где

$$K(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2t^2}.$$

Если статистика $\sqrt{n}D_n$ превышает порог t_α при заданном уровне значимости α, то нулевая гипотеза H_0 отвергается.

Значение порога t_α определяется из уравнения $K(t_\alpha) = \alpha$

Критерии согласия

Критерии ω^2 – семейство критериев согласия с расстоянием между выборочным и модельным распределением, имеющим вид квадратичной метрики:

$$d(F_n, F_0) = \omega^2 = \int_{-\infty}^{\infty} \big(F_n(x) - F_0(x)\big)^2 \psi(x) dF_0(x)$$

Здесь $\psi(x) > 0$ – весовая функция. Ее выбор определяет конкретный вид критерия.

Критерии согласия

При $\psi(x) = 1$, то получаем статистику **критерия *Крамера-Мизеса-Смирнова:***

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F_0(x_i) - \frac{2i-1}{2n}\right\}^2 .$$

Предельное распределение данной статистики имеет вид $a_1(s)$, т.е.,

$$\lim_{n\to\infty} P\{n\omega_n^2 < s\} = a_1(s),$$

$$a_1(s) = \frac{1}{\sqrt{2s}}\sum_{j=0}^{\infty}\Gamma\left(j+\frac{1}{2}\right)\sqrt{4j+1}\exp\left\{-\frac{(4j+1)^2}{16s}\right\}\left(I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right]\right),$$

Γ – гамма-функция, I_a - модифицированная функция Бесселя.

Критерии согласия

При $\psi(x) = \frac{1}{F_0(x)\,(1-F_0(x))}$ получается статистика **критерия *Андерсона-Дарлинга:***

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln\left(F_0(x_i)\right) + \left(1 - \frac{2i-1}{2n}\right)\left(1 - F_0(x_i)\right)\right\}^2,$$

для которой справедливо $\lim\limits_{n\to\infty} P\{n\Omega_n^2 < s\} = a_2(s)$, где

$$a_2(s) = \frac{\sqrt{2\pi}}{s}\sum_{j=0}^{\infty}\frac{\Gamma\left(j+\frac{1}{2}\right)(4j+1)}{\Gamma\left(\frac{1}{2}\right)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8s}\right\}\int_0^{\infty}\exp\left\{\frac{s}{8(y^2+1)} - \frac{(4j+1)^2\pi^2 y^2}{8s}\right\}dy\,.$$

Критерии согласия

Для последних двух критериев процедура выбора порогового значения s_α осуществляется аналогично выбору порога t_α в критерии Колмогорова. Этот выбор сводится к решению уравнений: $a_1(s_\alpha) = \alpha$ и $a_2(s_\alpha) = \alpha$.

Предельные распределения a_1 и a_2 данных статистик не зависят от вида функции распределения $F_0(x)$. Критерии, обладающие данным свойством, называют *свободными от распределений,* именно такие критерии используются для решения задач проверки гипотез.

Критерии согласия

Критерий χ^2 Пирсона. Данный критерий относится к типу критериев, основанных на группировке данных.

Разобьём область определения случайной величины на k непересекающихся интервалов с границами $a_0 < a_1 < \cdots < a_{k-1} < a_k$. Обозначим:

- n_i – число выборочных наблюдений, попавших в i -ый интервал, $\sum_{i=1}^{k} n_i = n$.
- $p_i = F_0(a_i) - F_0(a_{i-1})$ - вероятность попаданий в i -ый интервал, $\sum_{i=1}^{k} p_i = 1$.

Статистика χ^2 критерия Пирсона имеет вид:

$$\chi_n^2 = n \sum_{i=1}^{k} \frac{\left(\frac{n_i}{n} - p_i\right)^2}{p_i}.$$

Критерии согласия

Предельным распределением данной статистики является распределение хи-квадрат с k-1 степенью свободы:

$$\lim_{n\to\infty} P(\chi_n^2 < x) = \frac{\gamma\left(\frac{k-1}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k-1}{2}\right)},$$

где Γ и γ – полная и неполная гамма-функция, соответственно.

Для проверки статистических гипотез значения статистик сравниваются со значениями квантилей соответствующих распределений.