Весна 2022 г.

Специальный практикум (V курс)

Применение критерия согласия Пирсона.

Тестирование бинормальной модели для выборочной бимодальной функции распределения

И. И. Никифоров

1. Критерии согласия

Критерий согласия (КС) служит для проверки согласия между выборочным и гипотетическим распределениями или между двумя выборочными распределениями. КС позволяют оценить вероятность того, что полученная выборка не противоречит сделанному предположению о виде закона распределения рассматриваемой случайной величины X. Для этого выбирается некоторая статистика (*критериальная статистика*) $\varkappa$, являющаяся мерой расхождения наблюдаемого и теоретического законов распределения. Функция распределения этой статистики, $\varphi(\varkappa)$, принимается известной, как правило, асимптотически близкой (при $N \to \infty$) к некоторой функции распределения, рассматриваемой в теорвере.

На опыте получают значение меры расхождения $\varkappa_q$, по которому определяют вероятность $\alpha_q \equiv P(\varkappa \geqslant \varkappa_q)$, которая называется *уровнем значимости*. Далее сравнивают α_q с *критическим* уровнем значимости α — малой величиной, значение которой устанавливается в соответствии с характером задачи. Как правило, принимают $\alpha = 0.05$ (уровень значимости $\approx 2\sigma$). Т.е. рассматривают нуль-гипотезу

$$H_0\colon\ X \sim \varphi_{\mathrm{mod}}(x). \tag{1}$$

Гипотеза H_0 принимается или отвергается по следующему правилу:

$$\begin{aligned} &\alpha_q \geqslant \alpha \Longrightarrow \text{расхождение не противоречит } H_0, \\ &\alpha_q < \alpha \Longrightarrow H_0 \text{ отвергается с } P = 1 - \alpha_q \text{ (на уровне значимости } 1 - \alpha_q). \end{aligned} \tag{2}$$

Значения α_q, весьма близкие к 1 (очень хорошее согласие), могут указывать на недоброкачественность выборки (например, из первоначальной выборки без основания выброшены элементы, дающие большие отклонения от среднего).

Наиболее популярные критерии согласия.

1. КС Карла Пирсона (Karl Pearson, 1900): $\varkappa = \chi^2(\varphi_{\mathrm{obs}}(x), \varphi_{\mathrm{mod}}(x))$. Применим в том числе и в случае, если параметры модельного закона были найдены по данной выборке.

2. КС А. Н. Колмогорова: $\varkappa = \max|\Phi_{\text{obs}}(x) - \Phi_{\text{mod}}(x)|$. Применим в случае, когда параметры теоретического закона распределения определяются не по данным исследуемой выборки.

3. КС Н. В. Смирнова – А. Н. Колмогорова: $\varkappa = \max|\Phi_{\text{obs},1}(x) - \Phi_{\text{obs},2}(x)|$. Применяется для проверки гипотезы о принадлежности двух выборок одной генеральной совокупности.

2. Применение критерия Пирсона к тестированию бинормальной модели распределения металличностей шаровых скоплений Галактики

Из наблюдений известны

$$f_i \equiv [\text{Fe/H}]_i = \left[\lg\left(\frac{\text{Fe}}{\text{H}}\right)_* - \lg\left(\frac{\text{Fe}}{\text{H}}\right)_\odot\right]_i, \quad i = 1,\ \ldots, N. \tag{3}$$

Ранее была решена задача разложения наблюдаемого распределения на две нормальные составляющие:

$$\mathcal{N}_1(\bar{F}_1, \sigma_1^2), \quad \mathcal{N}_2(\bar{F}_2, \sigma_2^2). \tag{4}$$

Т.е. были найдены точечные оценки параметров $\bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c$ модельного дифференциального закона распределения

$$\varphi_{\text{mod}}(f) = \frac{c}{\sqrt{2\pi}\sigma_1}\, e^{-\dfrac{(f-\bar{F}_1)^2}{2\sigma_1^2}} + \frac{1-c}{\sqrt{2\pi}\sigma_2}\, e^{-\dfrac{(f-\bar{F}_2)^2}{2\sigma_2^2}}. \tag{5}$$

Задача. Применить критерий Пирсона для модели (5) и для гауссианы

$$\varphi_{\text{mod}}(f) = \mathcal{N}(f; \bar{F}_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0}\, e^{-\dfrac{(f-\bar{F}_0)^2}{2\sigma_0^2}}. \tag{6}$$

Для этих двух модельных распределений найти статистику

$$\chi_q^2 = \sum_{j=1}^{J} \frac{(N_j - Np_j)^2}{Np_j}, \tag{7}$$

Здесь N — объем выборки; J — число ячеек (разрядов); N_j — число объектов, попавших в ячейку j; p_j — теоретическое значение вероятность попадания случайной величины f в j-й интервал, вычисленная для модельного закона распределения по формуле

$$p_j = \int_{\hat{f}_j}^{\hat{f}_{j+1}} \varphi_{\text{mod}}(f)\, df. \tag{8}$$

При $N \to \infty$ закон распределения χ^2_q независимо от вида закона распределения случайной величины стремится к закону χ^2-распределения с $k = J - M - 1$ степенями свободы, где M — число параметров модельного закона распределения, найденных по гистограмме — по числам N_j объектов, попавшим в соответствующие ячейки. В наших случаях можно принять $M = 0$, т.к. параметры искались методом наибольшего правдоподобия, а не по гистограммам. Единица в k вычитается из-за условия нормировки $\sum N_j = N$.

Для применения критерия Пирсона в общем случае необходимо, чтобы объем выборки N и численности разрядов N_j были достаточно велики (практически считается достаточным, чтобы было $N \geqslant 50 \div 60$, $N_j \geqslant 5 \div 8$).

Выбор границ ячеек:

$$\hat{f}_1 = -\infty, \qquad \hat{f}_{J+1} = +\infty; \tag{9}$$

$$\hat{f}_{j+1} = \hat{f}_j + \Delta\hat{f}, \quad j = 2, \ldots, J-1; \tag{10}$$

$$\Delta\hat{f} = 0.1 \quad \Longrightarrow \quad \hat{f}_2 = -2.3, \;\; \hat{f}_J = -0.1; \tag{11}$$

$$\Delta\hat{f} = 0.2 \quad \Longrightarrow \quad \hat{f}_2 = -2.2, \;\; \hat{f}_J = -0.2. \tag{12}$$

Пусть ячейки включают правые границы (`...--]--]--]--...`). Т.е.

$$(-\infty, \hat{f}_2]; (\hat{f}_2, \hat{f}_3]; \ \ldots; (\hat{f}_{J-1}, \hat{f}_J]; (\hat{f}_J, +\infty).$$

Или для $\Delta\hat{f} = 0.1$

$$(-\infty, -2.30]; (-2.30, -2.20]; \ \ldots; (-0.2, -0.1]; (-0.1, +\infty);$$

для $\Delta\hat{f} = 0.2$

$$(-\infty, -2.20]; (-2.20, -2.00]; \ \ldots; (-0.4, -0.2]; (-0.2, +\infty).$$

Для нормального модельного распределения параметры можно найти по формулам

$$\bar{F}_0 = \frac{1}{N}\sum_{i=1}^{N} f_i, \qquad \sigma_0^2 = \frac{1}{N}\sum_{i=1}^{N}(f_i - \bar{F}_0)^2. \tag{13}$$

Значения вероятностей $\alpha_q(\chi^2_q, k) = P(\chi^2 \geqslant \chi^2_q)$ приводятся в таблицах.

3. Вычисление p_j и $\alpha_q(\chi^2_q, k)$

П/п ndtr.f — вычисляет $P(x \leqslant x_q)$, $x \in \mathcal{N}(0, 1)$.

П/п cdtr.f — вычисляет $P(\chi^2 \leqslant \chi^2_q)$ в зависимости от χ^2_q и k. Нужно взять дополнение до 1, чтобы найти α_q. П/п использует ndtr.f и dlgam.f.

Сверить с таблицами в книжке у Агекяна.

4. Задание и данные

Для заданного набора $f_i, i = 1, \ldots, N$, и для распределений (5) и (6), значений $\Delta\hat{f} = 0.1$ и 0.2 получить/привести результаты.

1. Привести параметры $\bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c$ или $\bar{F}_0, \sigma_0$.
2. Привести N, J, k, χ^2_q, α_q .
3. Графики сравнения наблюдаемого распределения с двумя модельными (для $\Delta\hat{f} = 0.1$ и 0.2).

Данные:

1. all.dat
2. wo_p1.dat
3. wo_p12.dat
4. wo_p123.dat

5. Приложение: визуализация

Нужно сопоставить модельную функцию $\varphi_{\rm mod}(f; \bar{F}_1, \sigma_1, \bar{F}_2, \sigma_2, c)$ с гистограммой наблюдаемого распределения f на одном рисунке.

Гистограмма наблюдаемого распределения. J ячеек длиной $\Delta f = 0.1\,\mathrm{dex}$; в каждой ячейке N_j объектов, $j = 1, 2, \ldots, J$. $\sum N_j = N$ (при нормировке столбцов гистограммы на N). Пусть ячейки гистограммы включают правые границы (. . .--]--]--]--. . .). Например, $(-2.40, -2.30]; (-2.30, -2.20], \ldots$

Модельная функция $\varphi_{\rm mod}(f)$. Нужно согласовать нормировки гистограммы $\varphi_{\rm hist}(f)$ и модельной функции $\varphi_{\rm mod}(f)$:

$$\int_{-\infty}^{+\infty} \varphi_{\rm hist}(f)\, d\varphi = \sum_{j=1}^{J} N_j\, \Delta f = N\, \Delta f, \tag{14}$$

$$\int_{-\infty}^{+\infty} \varphi_{\rm mod}(f)\, d\varphi = 1. \tag{15}$$

Т.е. совместно с гистограммой вместо модельной функции $\varphi_{\rm mod}(f)$ нужно рисовать

$$\boxed{N\, \Delta f\, \varphi_{\rm mod}(f)}. \tag{16}$$

Используя формулу (16), изобразить $\varphi_{\rm mod}$ и две ее нормальные составляющие [см. (5)].