Математическая статистика. ДЗ 6.

Прохоров Юрий, 776

Задача 1

Пусть $X_1, \ldots, X_N$ — независимые с.в., $X_i \sim \text{Poiss}(\theta_i)$, где θ_i - неизвестны. Пусть известно, что

$$X_1 + X_2 + \ldots + X_N = n$$

Проверить гипотезу однородности

$$H_0: \qquad \theta_1 = \theta_2 = \ldots = \theta_N$$

Решение:

1. Сначала введем обозначения

$$Z = \sum_{i=1}^{N} X_i, \qquad Y_j = \sum_{i \neq j} X_i = Z - X_j, \qquad \Theta = \sum_{i=1}^{N} \theta_i$$

Сумма независимых пуассоновских с.в. — пуассоновская, причем параметры суммируются:

$$Z \sim \text{Poiss}\left(\sum_{i=1}^{N} \theta_i\right), \qquad Y_j \sim \text{Poiss}\left(\sum_{i \neq j} \theta_i\right)$$

Кроме того нам известен факт, что если $\xi \sim \text{Poiss}(\lambda), \eta \sim \text{Poiss}(\mu)$, то

$$\xi \mid \xi + \eta = n \sim \text{Bin}\left(\frac{\lambda}{\lambda + \mu},\ n\right)$$

Это легко доказать расписав условные вероятности по определению.
Тогда

$$X_j \mid X_j + Y_j = n \sim \text{Bin}\left(\frac{\theta_j}{\Theta},\ n\right)$$

2. Построим критерий проверки гипотезы однородности.

Пусть гипотеза верна. Тогда

$$X_j \mid X_j + Y_j = n \sim \text{Bin}\left(\frac{1}{N},\ n\right), \qquad j = \overline{1, N}$$

Пусть нам даны реализации $X_1, \ldots, X_N$: числа $\nu_1, \ldots, \nu_N$, такие, что

$$\nu_1 + \nu_2 + \ldots + \nu_N = n$$

Тогда можно в качестве критерия взять проверку, что ν_1 является реализацией случайной величины, имеющей распределение $\text{Bin}\left(\frac{1}{N},\ n\right)$.

3. Если гипотеза верна, то

$$X_1 = \sum_{k=1}^{n} \xi_k, \qquad \xi_k \sim \text{Be}\left(\frac{1}{N}\right)$$

Далее можно воспользоваться

(a) критерием χ^2-Пирсона, вычислив статистику

$$T = \frac{\left(\nu_1 - \frac{n}{N}\right)^2}{\frac{n}{N}} + \frac{\left(n - \nu_1 - \frac{n(N-1)}{N}\right)^2}{\frac{n(N-1)}{N}} = \ldots = \frac{(\nu_1 N - n)^2}{n(N-1)} = \left[\frac{\nu_1 - n\,\frac{1}{N}}{\sqrt{n\,\frac{1}{N}\left(1-\frac{1}{N}\right)}}\right]^2$$

При $n \geq 50$ можно считать, что $T \sim \chi^2(1)$. Тогда критическая область:

$$\Omega^{(a)}_{\text{кр.},\ \alpha} = (\lambda_\alpha, +\infty),$$

где λ_α — α-квантиль распределения χ^2.

(b) центральной предельной теоремой:

$$\frac{\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}\xi_k}{\sqrt{\sum_{k=1}^n \mathbb{V}\xi_k}} = \frac{\sum_{k=1}^n \xi_k - n\,\frac{1}{N}}{\sqrt{n\,\frac{1}{N}\ \left(1-\frac{1}{N}\right)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

При достаточно больших n можно считать, что

$$\frac{\nu_1 - n\,\frac{1}{N}}{\sqrt{n\,\frac{1}{N}\ \left(1-\frac{1}{N}\right)}} \sim \mathcal{N}(0,1)$$

Тогда критическая область:

$$\Omega^{(b)}_{\text{кр.},\ \alpha} = \left(-\infty, -t_{\alpha/2}\right) \cup \left(+t_{\alpha/2}, +\infty\right),$$

где $t_{\alpha/2}$ — $\frac{\alpha}{2}$-квантиль распределения $\mathcal{N}(0,1)$.

Заметим, что выполнено $\left(t_{\alpha/2}\right)^2 = \lambda_\alpha$.

Видим, что обоими способами мы приходим фактически к одному и тому же критерию.
Построенный критерий не очень хороший, потому что мы не используем всю данную нам информацию. Этот критерий подошел бы для проверки гипотезы

$$\widetilde{H}_0: \qquad \theta_1 = \frac{\Theta}{N}$$

4. Если $X_i \sim \mathrm{Poiss}(\theta_i)$ — независимые, то несложно вычислить

$$\mathbb{P}\left\{X_1 = k_1, \ldots, X_N = k_N \;\middle|\; \sum_i X_i = n\right\} = \begin{cases} \dfrac{n!}{k_1! \ldots k_N!} \cdot \left(\dfrac{\theta_1}{\Theta}\right)^{k_1} \cdots \left(\dfrac{\theta_N}{\Theta}\right)^{k_N} & , \quad \sum_{i=1}^N k_i = n \\ 0 & , \quad \text{иначе} \end{cases},$$

где $\Theta = \sum_{i=1}^N \theta_i$. То есть условное распределение

$$\left(X_1, \ldots, X_n\right)\big| \sum_i X_i = n \ \sim\ \mathrm{Multi}\left(n, \frac{\theta_1}{\Theta}, \ldots, \frac{\theta_N}{\Theta}\right)$$

является полиномиальным (мультиномиальным) распределением.

На этот случайный вектор можно смотреть как на сгруппированные данные дискретной случайной величины ξ:

$$X_j = \sum_{i=1}^N \mathbb{I}\{\xi = j\}, \qquad \mathbb{P}\{\xi = j\} = \frac{\theta_j}{\Theta} \equiv p_j, \qquad j = \overline{1, N}$$

Таким образом нам нужно проверить простую гипотезу относительно простой сгруппированной выборки:

$$H_0: \qquad p_1 = \ldots = p_N = \frac{1}{N}$$

Это, в свою очередь, легко сделать с помощью критерия χ^2-Пирсона.

Задача 2

Смоделировать выборку $X_1, \ldots, X_n \sim \mathcal{U}[0,1]$ — i.i.d., $n = 100$. С помощью метода инверсий проверить гипотезу случайности для этих чисел.

Решение:

Посчитаем число инверсий для данного набора:

$$T_n = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{I}\{X_i > X_j\}$$

Воспользуемся теоремой:

$$Q_n = \frac{T_n - \mathbb{E}T_n}{\sqrt{\mathbb{V}T_n}} = \frac{T_n - \frac{1}{2}C_n^2}{\sqrt{\frac{n^3}{36} + O(n^2)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

При большом n можно считать, что

$$Q_n = \frac{T_n - \frac{1}{2}C_n^2}{\sqrt{\frac{n^3}{36}}} \sim \mathcal{N}(0,1)$$

Тогда можно построить критическую область:

$$\Omega_{\text{кр.},\alpha} = \left(-\infty, -t_{\alpha/2}\right) \cup \left(+t_{\alpha/2}, +\infty\right),$$

где $t_{\alpha/2}$ — $\frac{\alpha}{2}$-квантиль распределения $\mathcal{N}(0,1)$.

Оказалось, что число инверсий:

$$T_n = 2576, \qquad Q_n = 0.606$$

Соответствующее p-значение:

$$p = \frac{1 - \int\limits_{-Q_n}^{Q_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx}{2} \approx 0.272$$

Получается, что гипотеза будет принята на уровне $\alpha = 0.27$ и будет отвергнута на уровне $\alpha = 0.28$.