Тесты на случайность, основанные на длине самой длинной серии. Тесты на случайность для временных рядов.

На семинаре мы продолжили изучать различные тесты на случайность.

Тесты на основе самой длинной серии

Сначала мы рассмотрели другой способ построения статистического критерия случайности последовательности длины n элементов двух типов, состоящей из n_1 элементов первого типа и $n_2 \equiv n - n_1$ элементов второго типа.

Пусть как и ранее

$$R_1 \text{ — общее число серий из элементов первого типа}$$

и

$$R_2 \text{ — общее число серий из элементов второго типа.}$$

Обозначим через $R_{i,j}$ количество серий длины j, составленных из элементов i-ого типа, где $i = 1, 2; j = 1, 2, \ldots, n_i$. Тогда, с необходимостью, для $i = 1, 2$,

$$\sum_{j=1}^{n_i} R_{i,j} = R_i \qquad \text{и} \qquad \sum_{j=1}^{n_i} j R_{i,j} = n_i.$$

В дальнейшем, для сокращения изложения мы будем использовать следующие обозначения: $\boldsymbol{R_i} = (R_{i,1}, R_{i,2}, \ldots, R_{i,n_i})$, $\boldsymbol{n_i} = (1, 2, \ldots, n_i)$, $(\boldsymbol{u}, \boldsymbol{v})$ используется для обозначения евклидова скалярного произведения векторов $\boldsymbol{u}$ и $\boldsymbol{v}$, а $\|\boldsymbol{u}\|$ — для обозначения l_1-нормы вектора $\boldsymbol{u}$.

Следующая теорема дает вид совместного распределения вектора $(\boldsymbol{R_1}, \boldsymbol{R_2})$.

Теорема 1. *Пусть $\boldsymbol{r_1} \in \mathbb{Z}_+^{n_1}$ и $\boldsymbol{r_2} \in \mathbb{Z}_+^{n_2}$ такие, что $(\boldsymbol{r_i}, \boldsymbol{n_i}) = n_i$ для $i = 1, 2$. Тогда для $r_1 = \|\boldsymbol{r_1}\|$ и $r_2 = \|\boldsymbol{r_2}\|$, при верной основной гипотезе,*

$$f(\boldsymbol{r_1}, \boldsymbol{r_2}) := \mathbf{P}\left(\boldsymbol{R_1} = \boldsymbol{r_1}, \boldsymbol{R_2} = \boldsymbol{r_2}\right) = \frac{C(r_1, r_2) \cdot r_1! \cdot r_2!}{\binom{n_1+n_2}{n_1} \cdot \prod_{i=1}^{2} \prod_{j=1}^{n_i} r_{i,j}!},$$

где

$$C(r_1, r_2) = \begin{cases} 0, & \text{если } |r_1 - r_2| > 1; \\ 1, & \text{если } |r_1 - r_2| = 1; \\ 2, & \text{если } r_1 = r_2. \end{cases}$$

Используя теорему 1 можно получить «маргинальные» распределения элементов $\boldsymbol{R_1}$ и $\boldsymbol{R_2}$.

Теорема 2. *Пусть* $\boldsymbol{r_1} \in \mathbb{Z}_+^{n_1}$ *такой, что* $(\boldsymbol{r_1}, \boldsymbol{n_1}) = n_1$. *Тогда для* $r_1 = \|\boldsymbol{r_1}\|$, *при верной основной гипотезе,*

$$f(\boldsymbol{r_1}) := \mathbf{P}\left(\boldsymbol{R_1} = \boldsymbol{r_1}\right) = \frac{r_1! \cdot \binom{n_2+1}{r_1}}{\prod_{j=1}^{n_1} r_{1,j}! \cdot \binom{n_1+n_2}{n_1}}.$$

Мы построили тест на случайность на основании статистики

$$K = \max\{j : R_{1,j} > 0\}$$

— длины самой длинной серии из элементов первого типа. Используя теорему 2 при верной основной гипотезе можно получить хвост распределения

$$\mathbf{P}\left(K > t\right) = \sum_{k=t}^{n_1} \sum_{\boldsymbol{r_1}} f(\boldsymbol{r_1}),$$

где суммирование во второй сумме ведется по всем таким $\boldsymbol{r_1}$, что

$$\sum_{j=1}^{k} r_{1,j} = r_1, \qquad \sum_{j=1}^{k} j r_{1,j} = n_1 \quad \text{и} \quad r_{1,k} \geq 1,$$

и определить критическую область критерия

$$\delta(\boldsymbol{X}) = \begin{cases} H_0, \text{ если } K \leq q_{1-\alpha}; \\ H_1, \text{ если } K \geq q_{1-\alpha}. \end{cases}$$

Тесты для временных рядов

Пусть мы наблюдаем некоторый временной ряд $\boldsymbol{X_n} = (X_1, X_2, \ldots, X_n)$. Как понять, является ли этот ряд независимыми реализациями какой бы то ни было случайной величины?

Предположим для простоты, что все элементы $\boldsymbol{X_n}$ попарно различны. Через $a_1 < a_2 < \ldots < a_n$ обозначим упорядоченные значения наблюдаемого временного ряда. Основная гипотеза H_0 состоит в том, что наблюдаемые значения случайны, то есть

$H_0 = \{$ряд $\boldsymbol{X_n}$ является случайной перестановкой элементов $a_1, a_2, \ldots, a_n\}$.

Для построения критерия рассмотрим вспомогательную последовательность перемен знаков $\boldsymbol{D_{n-1}} = (D_1, D_2, \ldots, D_{n-1})$, где $D_j = \operatorname{sgn}(X_{j+1} - X_j)$ для всех $j = 1, 2, \ldots, n-1$. Обозначим через R_j число серий длины j в последовательности $\boldsymbol{D_{n-1}}$, $j = 1, 2, \ldots, n-1$, а через $\boldsymbol{R_{n-1}}$ — вектор $(R_{n-1}, R_{n-2}, \ldots, R_1)$.

Сначала мы нашли совместное распределение $f_{n-1}(\boldsymbol{r_{n-1}})$ вектора $\boldsymbol{R_{n-1}}$, используя некоторое естественное рекуррентное соотношение, а затем построили критерий на основе статистики $R_{n-1}(\geq t)$ = количество серий длины $\geq t$:

$$\delta(\boldsymbol{X_n}) = \begin{cases} H_0, \text{ если } R_{n-1}(\geq t) < r; \\ H_1, \text{ если } R_{n-1}(\geq t) \geq r. \end{cases}$$

Критическая область критерия определяется с использованием того факта, что при верной основной гипотезе

$$\mathbf{P}(R_{n-1}(\geq t) \geq r) = \sum_{\boldsymbol{r_{n-1}}} f_{n-1}(\boldsymbol{r_{n-1}}),$$

где суммирование ведётся по всем $\boldsymbol{r_{n-1}}$ таким, что

$$\sum_{k=t}^{n-1} r_k \geq r \qquad \text{и} \qquad \sum_{j=1}^{n-1} j r_j = n - 1.$$