Общие понятия статистики

Семинар был посвящен введению в теорию проверки гипотез.

Определение 1. *Пусть $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ — выборка размера n из некоторого распределения F. Гипотезой H называется любое предположение о природе распределения F*

Рассмотрим случай, когда гипотезы всего две — основная

$$H_0 = \{F \in \mathcal{F}_0\}$$

и альтернативная:

$$H_a = \{F \in \mathcal{F}_a\},$$

где $\mathcal{F}_0$ и $\mathcal{F}_a$ — некоторые непересекающиеся семейства распределений.

Определение 2. *Статистическим критерием, различающим основную гипотезу H_0 и альтернативную гипотезу H_a, называется функция*

$$\delta : \mathcal{X}^n \to \{H_0,\, H_a\},$$

которая задается следующим образом:

$$\delta(\vec{X}) = \begin{cases} H_0, & \text{если } \boldsymbol{X} \in \mathbb{R}^n / K; \\ H_a, & \text{если } \boldsymbol{X} \in K. \end{cases}$$

Здесь $\mathcal{X}^n$ — выборочное пространство, а множество K называется критической областью критерия $\delta(\boldsymbol{X})$.

Определение 3. *Ошибкой i-го рода называется вероятность при верной i-ой гипотезе принять другую.*

Ошибка первого рода, в случае если гипотез две, называется *размером или уровнем значимости* критерия. Единица минус ошибка второго рода — *мощностью* критерия.

Определение 4. *Фактически достигаемым уровнем значимости (p-значением) семейства критериев $\{\delta_\varepsilon(\boldsymbol{X})\}, \varepsilon \in (0,1)$ с критическими областями $K(\varepsilon)$ и размерами ε, на выборке $\boldsymbol{X}$ называется случайная величина*

$$\varepsilon(\boldsymbol{X}) = \inf\{\varepsilon : \boldsymbol{X} \in K(\varepsilon)\}.$$

Тесты на случайность. Тесты, основанные на общем количестве серий.

Как понять, что последовательность символов двух типов является случайной?

Серия — это подпоследовательность символов одного типа, за которой следует и которой предшествует другой тип символа или нет символа вообще. Пусть, например, имеется последовательность символов $aabbbbaaaabbbabbabbbbaa$. Здесь 5 серий символов типа a и 4 серии символов типа b.

Ключ к отсутствию случайности дает любая тенденция символов демонстрировать определенный шаблон в последовательности (чередование, кластеризация и т.д.)

Мы сроим критерий, различающий основную гипотезу

$$H_0 = \{\text{последовательность случайна}\},$$

и альтернативу

$$H_a = \{\text{последовательность не случайна}\}.$$

Пусть у нас имеется упорядоченная последовательность n элементов двух типов, причём

n_1 — количество элементов типа 1, R_1— количество серий элементов типа 1,
n_2 — количество элементов типа 2, R_2— количество серий элементов типа 2,

где, с необходимостью, $n = n_1 + n_2$.

Пусть $R = R_1 + R_2$ — общее число серий в последовательности. Мы построим тест, основанный на этой величине.

Распределение R

Распределение R ищется в предположении, что верна H_0. Сначала мы нашли совместное распределение R_1, R_2, доказав следующую

Лемма 1. *Число различных размещений n объектов в r различных ячеек, при которых не остается пустых ячеек, равно $\binom{n-1}{r-1}$, $n \geq r,\ r \geq 1$.*

Основываясь на лемме, можно получить совместное распределение вероятностей R_1, R_2.

Теорема 1. *Совместное распределение R_1 и R_2 задаётся правилом:*

$$f_{R_1,R_2}(r_1, r_2) \equiv \mathbf{P}\left(R_1 = r_1, R_2 = r_2\right) = \frac{c\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n_1}}, \quad r_i = 1, \ldots, n_i, \quad i = 1, 2,$$

где $c = 2$, если $r_1 = r_2$, и $c = 1$, если $r_1 = r_2 \pm 1$ (r_1 и r_2 могут принимать только такие значения).

Просуммировав по всем возможным значениям R_2, получим распределение R_1

$$f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1}}{\binom{n_1+n_2}{n_1}}, \quad r_1 = 1, \ldots, n_1.$$

Аналогично получается распределение для R_2.

Из вышесказанного можно получить распределение для R

Теорема 2.

$$f_R(r) = \begin{cases} \frac{2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n_1+n_2}{n_1}} =: f_1(r), & \text{если } r \text{ четное}, \\ \frac{\binom{n_1-1}{(r-1)/2-1}\binom{n_2-1}{(r-3)/2-1}+\binom{n_1-1}{(r-3)/2-1}\binom{n_2-1}{(r-1)/2-1}}{\binom{n_1+n_2}{n_1}} =: f_2(r), & \text{если } r \text{ нечетное}. \end{cases}$$