Математическая статистика. ДЗ 5.

Прохоров Юрий, 776

Задача 1

В ВУЗ поступают 300 абитуриентов. В таблице приведены их сведения об их оценках по математике в школе и на вступительном экзамене.

	5 в школе	5 на экзамене	5 в школе и на экзамене
# человек	97	48	18

Проверить гипотезу о независимости оценок 5 в школе и на экзамене на уровне значимости $\alpha = 0.1$.

Решение:

Формализуем постановку задачи. Считаем, что у нас есть 2 бернуллиевские случайные величины:

$$X = \mathbb{I}\{5 \text{ в школе}\}, \qquad Y = \mathbb{I}\{5 \text{ на экзамене}\}$$

Нам дана простая выборка из $n = 300$ реализаций случайного вектора $\begin{bmatrix} X \\ Y \end{bmatrix}$. Требуется проверить гипотезу о независимости его компонент.

Составим таблицу из данных в удобном нам формате:

$\Delta_1^Y = \{1\}$	30	18
$\Delta_0^Y = \{0\}$	173	79
	$\Delta_0^X = \{0\}$	$\Delta_1^X = \{1\}$

где Δ_0^X, Δ_1^X — разбиение множества значений X на $r = 2$ ячеек, а Δ_0^Y, Δ_1^Y — разбиение множества значений Y на $l = 2$ ячеек.

Строим модель, считая что X и Y независимы. Тогда

$$p_{ij} = p_i^X p_j^Y, \qquad \text{где} \quad p_{ij} = \mathbb{P}\{X \in \Delta_i^X,\ Y \in \Delta_j^Y\}, \ \ p_i^X = \mathbb{P}\{X \in \Delta_i^X\}, \ \ p_j^Y = \mathbb{P}\{Y \in \Delta_j^Y\}$$

Тогда независимыми параметрами модели являются

$$\theta = \left(p_0^X,\ p_0^Y\right), \qquad \dim\theta = 2$$

Мы свели задачу к проверке сложной гипотезы. Записываем χ^2-статистику:

$$T_n(\theta) = \sum_{j=1}^{l}\sum_{i=1}^{r} \frac{(\nu_{ij} - np_{ij})^2}{np_{ij}}, \qquad \nu_{ij} = \#\ \text{элементов в}\ \Delta_i^X \times \Delta_j^Y$$

Оптимальные параметры:

$$\widehat{\theta}_n = \arg\min_{\theta} T_n(\theta) \qquad \iff \qquad \begin{matrix} p_0^X = \dfrac{\nu_0^X}{n} & p_1^X = \dfrac{\nu_1^X}{n} \\ p_0^Y = \dfrac{\nu_0^Y}{n} & p_1^Y = \dfrac{\nu_1^Y}{n} \end{matrix}$$

где $\nu_0^X = \#$ попаданий X в Δ_0^X, и т.д.

При этом статистика принимает вид

$$T_n(\widehat{\theta}_n) = \sum_{j=1}^{l}\sum_{i=1}^{r} \frac{n\left(\nu_{ij} - \dfrac{\nu_i^X \nu_j^Y}{n}\right)^2}{\nu_i^X \nu_j^Y} \xrightarrow[n\to\infty]{d} \chi^2\big((r-1)(l-1)\big)$$

В нашем случае $T_n(\widehat{\theta}_n) \approx 0.697$, α-квантиль распределения $\chi^2(1)$ равен $\lambda_\alpha = 2.71$.

$$T_n < \lambda_\alpha \qquad \Longrightarrow \qquad \text{данные не противоречат гипотезе на уровне } \alpha = 0.1$$

Можно посчитать, что гипотеза будет принята на уровне $\alpha = 0.4$. То есть p-value $= 0.4$.

Задача 2

Смоделировать последовательность $Y_1, \ldots, Y_{100} \sim \mathcal{U}\{1, \ldots, 5\}$ — i.i.d.

(a) Построить две выборки

$$\mathbf{X}^{(1)} = \big\{Y_{2i-1}\big\}_{i=1}^{50}, \qquad \mathbf{X}^{(2)} = \big\{Y_{2i}\big\}_{i=1}^{50}$$

Проверить гипотезу однородности для этих выборок на уровне $\alpha = 0.05$.

(b) Образовать выборку из двумерных векторов

$$\mathbf{X}:\ X_1, \ldots, X_{50}, \qquad X_i = \begin{bmatrix} Y_{2i-1} \\ Y_{2i} \end{bmatrix}$$

Проверить гипотезу независимости компонент случайного вектора X на уровне $\alpha = 0.05$.

Решение:

(b) Решаем аналогично задаче 1.

Сначала естественным образом разобьем множества значений нечетных и четных Y_i на 5 областей:

$$\Delta_k^{\text{нечет}} = \Delta_k^{\text{чет}} = \{k\}, \qquad k = \overline{1,5}$$

и построим соответствующую таблицу попаданий в $\Delta_i^{\text{нечет}} \times \Delta_j^{\text{чет}}$:

	17	12	7	6	8	
$\Delta_5^{\text{чет}}$	2	3	3	2	0	10
$\Delta_4^{\text{чет}}$	6	0	2	1	3	12
$\Delta_3^{\text{чет}}$	3	3	0	1	2	9
$\Delta_2^{\text{чет}}$	3	3	1	1	1	9
$\Delta_1^{\text{чет}}$	3	3	1	1	2	10
	$\Delta_1^{\text{нечет}}$	$\Delta_2^{\text{нечет}}$	$\Delta_3^{\text{нечет}}$	$\Delta_4^{\text{нечет}}$	$\Delta_5^{\text{нечет}}$	

Видим, что не выполнено условие применимости критерия χ^2 для проверки гипотезы независимости: $\nu_{ij} \geq 5$. Поэтому объединим колонки и строки:

	17	19	14	
$\Delta_{3,4,5}^{\text{чет}}$	11	11	9	31
$\Delta_{1,2}^{\text{чет}}$	6	8	5	19
	$\Delta_1^{\text{нечет}}$	$\Delta_{2,3}^{\text{нечет}}$	$\Delta_{4,5}^{\text{нечет}}$	

Теперь можно пользоваться критерием χ^2. Считаем статистику:

$$T = \sum_{j=1}^{l}\sum_{i=1}^{r} \frac{n\left(\nu_{ij} - \frac{\nu_i^X \nu_j^Y}{n}\right)^2}{\nu_i^X \nu_j^Y} \approx 0.2198$$

Найдем α-квантиль распределения $\chi^2(2)$: $\lambda_\alpha = 5.99$. $T < \lambda_\alpha$, значит, данные не противоречат гипотезе независимости на уровне 0.05.

p-value $= 0.9$

(a) Снова разбиваем носитель на 5 областей: $\Delta_k = \{k\}$.

	27	24	16	15	18	
нечет	17	12	7	6	8	50
чет	10	12	9	9	10	50
	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	

Условия применимости критерия χ^2 для гипотезы однородности выполнены. Считаем статистику:

$$T = \sum_{j=1}^{l} \sum_{i=1}^{r} \frac{n\left(\nu_{ij} - n_j \frac{\nu_i}{n}\right)^2}{n_j \nu_i} = 2.887$$

Найдем α-квантиль распределения $\chi^2(4)$: $\lambda_\alpha = 9.49$. $T < \lambda_\alpha$, значит, данные не противоречат гипотезе однородности на уровне 0.05.

p-value $= 0.58$