Математические и статистические методы в психологии
Проверка статистических гипотез. (6 декабря 2019 г.)
А. А. Макаров, А. А. Тамбовцева, Н. А. Василёнок, Е. П. Шеремет

Таблицы сопряженности и проверка независимости признаков, измеренных в качественной шкале.

Используется для выявления связи между двумя показателями, измеренными в качественной (номинальной) шкале. Примеры таких показателей: пол, уровень образования, согласие/несогласие с утверждением, поддержка/неподдержка кандидата.

Таблица сопряженности

Есть таблица сопряженности 2×2 (пол – любовь к шоколаду) и на 5% уровне значимости мы хотим проверить гипотезу о независимости признаков «пол» и «любовь к шоколаду».

	люблю шоколад	**не люблю шоколад**	
мужчины	20	15	$n_{1.} = 35$
женщины	35	20	$n_{2.} = 55$
	$n_{.1} = 55$	$n_{.2} = 35$	$N = 90$

Нумерация элементов таблицы – как в матрице (первый индекс элемента – номер строки, в которой находится элемент, второй индекс – номер столбца). Точка на месте индекса означает любую строку/столбец. Например, $n_{1.} = 35$ – сумма по первой строке (одна строка, все столбцы), а $n_{.1} = 55$ – сумма по первому столбцу (один столбец, все строки). N – сумма всех значений в таблице.

$$n_{11}^{\text{набл}} = 20$$
$$n_{12}^{\text{набл}} = 15$$
$$n_{21}^{\text{набл}} = 35$$
$$n_{22}^{\text{набл}} = 20$$

Проверка гипотезы о независимости признаков

H_0 : связи между признаками нет, они независимы

H_1 : связь между признаками есть, они не независимы

Для того, чтобы, как всегда, сравнивать наблюдаемое и критическое значение статистики критерия, необходимо определить ожидаемые частоты – значения в ячейках, которые имели бы место, если бы нулевая гипотеза была верна, и признаки были бы независимы. Общая формула расчета выглядит так:

$$n_{ij}^{\text{ожид}} = \frac{n_{i.} \cdot n_{.j}}{N,}$$

где i и j – номер строки и столбца, в которых находится интересующее число n. То есть, мы перемножаем сумму по соответствующей строке и столбцу и делим на общее число N. Рассчитаем ожидаемые значения всех частот в таблице.

$$n_{11}^{\text{ожид}} = \frac{35 \cdot 55}{90} \approx 21.4$$
$$n_{12}^{\text{ожид}} = \frac{35 \cdot 35}{90} \approx 13.6$$
$$n_{21}^{\text{ожид}} = \frac{55 \cdot 55}{90} \approx 33.6$$
$$n_{22}^{\text{ожид}} = \frac{55 \cdot 35}{90} \approx 21.4$$

Интересующие нас наблюдаемые частоты мы берем из таблицы. Получаем такие пары:

$$n_{11}^{\text{набл}} = 20 \text{ и } n_{11}^{\text{ожид}} = \frac{35 \cdot 55}{90} \approx 21.4$$
$$n_{12}^{\text{набл}} = 15 \text{ и } n_{12}^{\text{ожид}} = \frac{35 \cdot 35}{90} \approx 13.6$$
$$n_{21}^{\text{набл}} = 35 \text{ и } n_{21}^{\text{ожид}} = \frac{55 \cdot 55}{90} \approx 33.6$$
$$n_{22}^{\text{набл}} = 20 \text{ и } n_{22}^{\text{ожид}} = \frac{55 \cdot 35}{90} \approx 21.4$$

Статистика используемого критерия имеет распределение хи-квадрат (χ^2). Наблюдаемое значение статистики считается следующим образом:

$$\chi^2_{\text{набл}} = \sum_{i,j=1}^{n} \frac{(n_{ij}^{\text{набл}} - n_{ij}^{\text{ожид}})^2}{n_{ij}^{\text{ожид}}}$$

Посчитаем для нашего случая:

$$\chi^2_{\text{набл}} = \frac{(20-21.4)^2}{21.4} + \frac{(15-13.6)^2}{13.6} + \frac{(35-33.6)^2}{33.6} + \frac{(20-21.4)^2}{21.4} \approx 0.39$$

Считаем p-value (зная, что χ^2 с одной степенью свободы, для таблицы сопряженности 2×2 – это Z^2, где Z – стандартная нормальная величина):

$$\text{p-value} = P(\chi^2 > \chi^2_{\text{набл}}) = P(z^2 > 0.39) = P(|z| > \sqrt{0.39}) =$$
$$= P(|z| > 0.62) = 2P(z > 0.62) = 2 \cdot 0.27 = 0.54$$

Сравниваем полученное значение с $\alpha = 0.05$ ($0.54 > 0.05$) и делаем вывод о том, что на уровне значимости 5% нет оснований отвергнуть нулевую гипотезу о независимости признаков. Любовь к шоколаду никак не связана с полом человека.