**ОП «Политология», 2020-21**
**Введение в ТВиМС**
**Разбор задачи на критерий хи-квадрат (семинар 18)**
*А. А. Макаров, А. А. Тамбовцева*

**Задача 1.** Дана таблица сопряженности двух признаков: *пол респондента* и его ответ на вопрос: «*Любите ли вы горький шоколад?*».

| | Да | Нет |
|---|---|---|
| Женский | 32 | 14 |
| Мужской | 15 | 25 |

(a) Вам необходимо проверить, есть ли связь между этими признаками. Сформулируйте нулевую и альтернативную гипотезу. Какое распределение имеет статистика критерия, используемого для проверки нулевой гипотезы?

**Решение.** Сформулируем гипотезы:

$$H_0 : \text{признаки независимы}$$

$$H_1 : \text{признаки связаны}$$

Статистика критерия имеет распределение хи-квадрат с числом степеней свободы $\text{df} = 1$. Величина хи-квадрат с одной степенью свободы – это просто $Z^2$.

(b) Посчитайте наблюдаемое значение статистики и p-value. Какой вывод относительно нулевой гипотезы можно сделать на 5% уровне значимости?

**Решение.** Подготовим все необходимое для решения (маргинальные частоты и общее количество наблюдений):

| | Да | Нет | |
|---|---|---|---|
| Женский | 32 | 14 | $n_{1.} = 46$ |
| Мужской | 15 | 25 | $n_{2.} = 40$ |
| | $n_{.1} = 47$ | $n_{2.} = 39$ | $N = 86$ |

**Первый способ нахождения наблюдаемого значения (быстрый)**

$$\chi^2_{\text{набл}} = \frac{(n_{11} \cdot n_{22} - n_{12} \cdot n_{21})^2 \cdot N}{n_{1.} \cdot n_{2.} \cdot n_{.1} \cdot n_{.2}} = \frac{(32 \cdot 25 - 14 \cdot 15)^2 \cdot 86}{46 \cdot 40 \cdot 47 \cdot 39} \approx 8.88$$

**Второй способ нахождения наблюдаемого значения (содержательный)**

Для каждой ячейки посчитаем значение частоты, которое ожидается в случае, если нулевая гипотеза верна:

$$n_{11}^{\text{ожид}} = \frac{n_{1.} \cdot n_{.1}}{N} = \frac{46 \cdot 47}{86} \approx 25$$

$$n_{12}^{\text{ожид}} = \frac{n_{1.} \cdot n_{.2}}{N} = \frac{46 \cdot 39}{86} \approx 21$$

$$n_{21}^{\text{ожид}} = \frac{n_{2.} \cdot n_{.1}}{N} = \frac{40 \cdot 47}{86} \approx 22$$

$$n_{22}^{\text{ожид}} = \frac{n_{2.} \cdot n_{.2}}{N} = \frac{40 \cdot 39}{86} \approx 18$$

Теперь сравним наблюдаемые частоты с ожидаемыми. Нам не важно, в какую сторону наблюдаемые отличаются от ожидаемых, поэтому будем рассматривать квадраты разностей и оценивать их относительно ожидаемых частот:

$$\chi^2_{\text{набл}} = \frac{(n_{11}^{\text{набл}} - n_{11}^{\text{ожид}})^2}{n_{11}^{\text{ожид}}} + \frac{(n_{12}^{\text{набл}} - n_{12}^{\text{ожид}})^2}{n_{12}^{\text{ожид}}} + \frac{(n_{21}^{\text{набл}} - n_{21}^{\text{ожид}})^2}{n_{21}^{\text{ожид}}} + \frac{(n_{22}^{\text{набл}} - n_{22}^{\text{ожид}})^2}{n_{22}^{\text{ожид}}} =$$

$$= \frac{(32-25)^2}{25} + \frac{(14-21)^2}{21} + \frac{(15-22)^2}{22} + \frac{(25-18)^2}{18} \approx 9.24$$

Результаты, полученные первым и вторым способом, немного отличаются, но это просто эффект округления (во втором способе мы довольно грубо округлили все до целых). Для расчета pvalue выберем один из результатов, допустим, первый. Получаем:

$$\text{pvalue} = \text{P}(\chi^2 > \chi^2_{\text{набл}}) = \text{P}(\chi^2 > 8.88) = \text{P}(Z^2 > 8.88) = \text{P}(|Z| > \sqrt{8.88}) =$$

$$= 2 \cdot \text{P}(Z > \sqrt{8.88}) = 2 \cdot (1 - \Phi(\sqrt{8.88})) = 2 \cdot (1 - \Phi(2.98)) = 2 \cdot (1 - 0.9986) = 0.0028.$$

Сравниваем pvalue с уровнем значимости $\alpha = 5\%$:

$$0.0028 < 0.05 \Rightarrow H_0 \text{ отвергается}$$

Статистический вывод: на 5%-ном уровне значимости у нас есть основания отвергнуть нулевую гипотезу.

Содержательный вывод: есть связь между полом человека и предпочтениями относительно выбора шоколада.