ОП «Политология», 2019-20 уч. год
Математика и статистика, часть 2
Связь между качественными признаками. Критерий хи-квадрат.
А. А. Макаров, А. А. Тамбовцева, Н. А. Василёнок

Задача. Ниже приведены данные исследования, посвященные маркетинговой эффективности рекламной акции «экономичная упаковка».[1] В ходе исследования были опрошены 100 респондентов. Фиксировался пол респондента и наличие или отсутствие эффекта на него рекламной акции. Заполните пропущенные ячейки таблицы. Проверьте наличие связи между качественными признаками на уровне значимости $\alpha = 0.05$.

Таблица 1: Условие

Пол / Эффект	**Нет**	**Есть**	
Мужчины		15	50
Женщины			50
	40	60	**100**

Решение. Чтобы проверить наличие связи между качественными признаками, нельзя пользоваться коэффициентами корреляции. Нам нужно воспользоваться критерием хи-квадрат для таблиц сопряженности.

Что такое таблицы сопряженности, и как они устроены? Таблица сопряженности описывает совместное частотное распределение качественных признаков. В ячейках таблицы сопряженности находятся частоты *пересечения* признаков. Значения, получающиеся суммированием данных по строчкам и по колонкам называются *маргинальными частотами*. Сумма маргинальных частот по строчкам должна равняться сумме по колонкам и равняться общему числу респондентов.

Зная маргинальные частоты, можно легко восстановить пропущенные ячейки в таблице:

Таблица 2: Наблюдаемые частоты

Пол / Эффект	**Нет**	**Есть**	
Мужчины	*35*	15	50
Женщины	*5*	*45*	50
	40	60	**100**

Давайте договоримся об обозначениях. Частоту в ячейках мы будем обозначать как $n_{ij}^{observed}$, где i – номер строки, а j – номер колонки. Маргинальные частоты, получающиеся суммированием частот по строкам, обозначим как $n_{i.}$. Маргинальные частоты,

[1] Например: 2 литра молока по цене 1.

получающиеся суммированием частот по колонкам, обозначим как $n_{\cdot j}$.

Когда мы рассматривали независимость дискретных случайных величин, то говорили, что условием независимости является следующее равенство:

$$P(A \cap B) = P(A) \times P(B)$$

Похожая логика лежит в анализе таблиц сопряженности. Зная маргинальные частоты, мы можем оценить, как выглядела бы таблица сопряженности, **если бы признаки были независимы**. Для этого нам нужно расчитать *ожидаемую частоту для каждой колонки*:

$$n_{ij}^{expected} = \frac{n_{i.} \times n_{\cdot j}}{N}$$

Запишем таблицу ожидаемых частот:

Таблица 3: Ожидаемые частоты

Пол / Эффект	**Нет**	**Есть**	
Мужчины	20	30	50
Женщины	20	30	50
	40	60	**100**

Теперь можно воспользоваться критерием хи-квадрат.

H_0: признаки независимы
H_a: признаки зависимы

Статистика которого выглядит следующим образом и иллюстрирует отклонения наблюдаемых частот от тех, которые мы могли бы ожидать, если бы признаки были независимы:

$$\chi^2_{df} = \sum_{i,j}^{I,J} \frac{(n_{ij}^{observed} - n_{ij}^{expected})^2}{n_{ij}^{expected}}$$

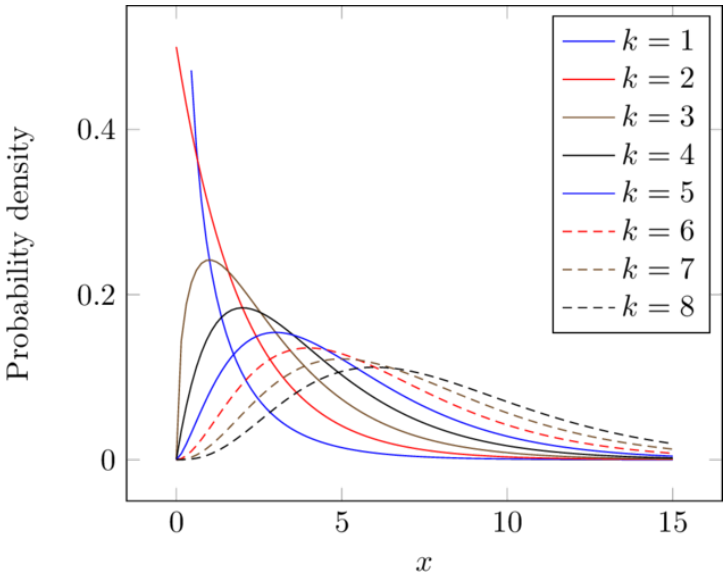
Обратите внимание, что чем ближе статистика к нулю, тем с большей уверенностью мы можем считать, что признаки независимы. Статистика имеет распределение хи-квадрат с числом степеней свободы

$$df = (I-1)(J-1) \ ,$$

где I – количество строчек, а J – количество столбцов в таблице сопряженности. Это число обозначает минимальное количество ячеек в таблице, которые нам нужно знать, чтобы восстановить всю таблицу.

Распределение хи-квадрат $\chi^2_{(df=k)}$ представляет собой сумму k квадратов стандартно нормально распределенных случайных величин: $Z_1^2 + Z_2^2 + ... + Z_k^2$. Следовательно,

случайная величина $\chi^2_{(df=k)}$ принимает значения $[0,\ +\infty)$.



Рассчитаем статистику критерия:

$$\chi^2_{(2-1)(2-1)} = \chi^2_{(1)} = \frac{(35-20)^2}{20} + \frac{(15-30)^2}{30} + \frac{(5-20)^2}{20} + \frac{(45-30)^2}{30} =$$
$$= \frac{450}{20} + \frac{450}{30} = 15 + 22.5 = 37.5$$

Чтобы проверить нулевую гипотезу, нам нужно рассчитать *p-value*.

P-value.

$$Z^2 = \chi^2_{(1)}$$
$$P(\chi^2_{(1)} > x) = P(Z^2 > x) = P(|Z| > \sqrt{x}) = 2 \times P(Z > \sqrt{x})$$
$$P(\chi^2_{(1)} > 37.5) = P(Z^2 > 37.5) = 2 \times P(Z > 6.23) = 0$$

Значение p-value меньше уровня значимости $\alpha = 0.05$, нулевая гипотеза о независимости признаков отвергается.

Упрощенная формула для таблиц 2x2. Формулу выше можно преобразовать в упрощенную формулу, которая для таблиц 2x2 имеет следующий вид:

$$\chi^2_{(2-1)(2-1)} = \frac{N \times (n_{11} \times n_{22} - n_{12} \times n_{21})^2}{n_{1.} \times n_{2.} \times n_{.1} \times n_{.2}}$$

В этой формуле используются только *наблюдаемые* частоты. Обратите внимание, что в числителе находится квадрат определителя матрицы – таблицы сопряженности.

$$\chi^2_{(2-1)(2-1)} = \frac{100 \times (35 \times 45 - 5 \times 15)^2}{40 \times 60 \times 50 \times 50} = 37.5$$

Распространенные значения квантилей распределения хи-квадрат с df=1

1. $\alpha = 0.05$

$$P(-1.96 < Z < 1.96) = 0.95$$

$$Z^2 = 1.96^2 = 3.8416$$

$$P(\chi_1^2 < 3.8416) = 0.95$$

2. $\alpha = 0.1$

$$P(-1.645 < Z < 1.645) \simeq 0.9$$

$$Z^2 = 1.645^2 \simeq 2.71$$

$$P(\chi_1^2 < 2.71) = 0.9$$

3. $\alpha = 0.01$

$$P(-2.576 < Z < 2.576) \simeq 0.99$$

$$Z^2 = 2.576^2 \simeq 6.635$$

$$P(\chi_1^2 < 6.635) = 0.99$$