Меры связи

1 Коэффициент корреляции К.Пирсона.

Используется для выявления *линейной* связи между двумя показателями, измеренными в количественной шкале. Желательно, чтобы в данных не было нетипичных значений (выбросов), так как их наличие может искажать полученные результаты.

- **Расчет коэффициента корреляции** R

$$R = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}},$$

где $\bar{x}$ – среднее арифметическое, посчитанное по первой выборке, где $\bar{y}$ – среднее арифметическое, посчитанное по второй выборке, n – число элементов в выборке.

$R \in [-1; 1]$, если $R > 0$ – связь между показателями прямая, если $R < 0$ – связь между показателями обратная, $R \neq 0$ – линейной связи между показателями нет.

- **Проверка гипотезы о равенстве теоретического коэффициента корреляции ρ нулю**

$H_0 : \rho = 0$ (связи между показателями нет)

$H_A : \rho \neq 0$ (связь между показателями есть)

$$t_{\text{набл}} = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2}$$

$$t_{\text{крит}} = t_{(1-\frac{\alpha}{2},\ df=n-2)}$$

$|t_{\text{набл}}| > t_{\text{крит}} \Rightarrow H_0$ отвергается, связь между показателями есть.

$|t_{\text{набл}}| < t_{\text{крит}} \Rightarrow H_0$ не отвергается, связи между показателями нет.

2 Коэффициент корреляции Ч.Спирмена.

Используется для выявления связи между двумя показателями, измеренными в ранговой (ординальной) шкале. [1] Можно использовать и для выявления связи между показателями, измеренными в метрической шкале; более того, данный коэффициент уместно вычислять в случае, когда в выборках присутствуют нетипичные значения (выбросы), так как коэффициент корреляции Ч.Спирмена является более устойчивым к выбросам по сравнению с коэффициентом корреляции К.Пирсона.

- **Расчет коэффициента корреляции** $R_{\text{Спирмена}}$

$$R_{\text{Спирмена}} = 1 - \frac{6 \cdot \sum\limits_{i=1}^{n} d_i^2}{n(n^2-1)},$$

где d_i – разность между рангом i-того наблюдения в первой выборке и рангом i-того наблюдения во второй выборке, n – число элементов в выборке.

$R_{\text{Спирмена}} \in [-1; 1]$, если $R > 0$ – согласованность рангов прямая, если $R < 0$ – согласованность рангов обратная, $R \neq 0$ – связи между рангами нет.

[1] Примеры показателей: места в рейтинге, экспертные оценки от 1 до 5.

- **Проверка гипотезы о равенстве теоретического коэффициента корреляции ρ нулю**

$H_0 : \rho_{\text{Спирмена}} = 0$ (связи между показателями нет)

$H_A : \rho_{\text{Спирмена}} \neq 0$ (связь между показателями есть)

$$z_{\text{набл}} = R_{\text{Спирмена}}\sqrt{n-1}$$
$$z_{\text{крит}} = z_{(1-\frac{\alpha}{2})}$$

$\left|z_{\text{набл}}\right| > z_{\text{крит}} \Rightarrow H_0$ отвергается, связь между показателями есть.

$\left|z_{\text{набл}}\right| < z_{\text{крит}} \Rightarrow H_0$ не отвергается, связи между показателями нет.

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

Используется для выявления связи между двумя показателями, измеренными в качественной (номинальной) шкале.[2]

- **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_A : 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

[2]Примеры показателей: пол, уровень образования, согласие/несогласие с утверждением, поддержка/неподдержка кандидата.

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:

$n_{11}^{\text{набл}} = 20$ $\quad$ $n_{11}^{\text{ожид}} = \frac{35 \cdot 55}{90} \approx 21.4$

$n_{12}^{\text{набл}} = 15$ $\quad$ $n_{12}^{\text{ожид}} = \frac{35 \cdot 35}{90} \approx 13.6$

$n_{21}^{\text{набл}} = 35$ $\quad$ $n_{21}^{\text{ожид}} = \frac{55 \cdot 55}{90} \approx 33.6$

$n_{22}^{\text{набл}} = 20$ $\quad$ $n_{22}^{\text{ожид}} = \frac{55 \cdot 35}{90} \approx 21.4$

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:

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

Критическое значение статистики критерия определяется следующим образом:

$$\chi^2_{\text{крит}} = \chi^2_{1-\alpha,\ df=(r-1)(c-1)},$$

где r – число строк в таблице сопряженности, c – число столбцов в таблице. В простом случае 2×2, в таком, как наш, число степеней свободы будет всегда равняться 1.

$$\chi^2_{\text{крит}} = \chi^2_{1-0.05,\ df=1} = 3.841$$

Сравниваем наблюдаемое значение и критическое, видим, что наблюдаемое меньше критического, 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.