ОП «Политология», 2018-19
Математика и статистика, часть 2
Меры связи (16 июня 2019 г.)
А. А. Макаров, А. А. Тамбовцева, Н. А. Василёнок

Коэффициент корреляции К. Пирсона

Используется для выявления *линейной* связи между двумя переменными, измеренными в количественной шкале. Желательно, чтобы в данных при рассмотрении совместного распределения переменных не было нетипичных значений (выбросов), так как их наличие может искажать полученные результаты. Коэффициент корреляции К. Пирсона является неустойчивым к выбросам.

- **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.

Значение R всегда лежит в интервале $[-1; 1]$, если $R > 0$ – связь между показателями прямая (положительная), если $R < 0$ – связь между показателями обратная (отрицательная), а если $R = 0$ – линейной связи между показателями нет. Однако интерпретировать коэффициент имеет смысл в том случае, если он статистически значим (см.ниже).

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

Нулевая и альтернативная гипотезы:

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

$$H_1 : \rho \neq 0 \text{ (связь между показателями есть)}$$

Приведенную выше нулевую гипотезу иногда еще называют *гипотезой о значимости коэффициента корреляции* (хотя интуитивно наоборот – она говорит о незначимости, об отсутствии статистически значимых отличий от нуля).
Наблюдаемое значение t-статистики:

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

где R – значение коэффициента корреляции Пирсона, а n – число наблюдений в выборке.
Соответствующее p-value:

$$\text{p-value} = P(|t| > t_{\text{набл}}) = 2 \cdot P(t > t_{\text{набл}}) = 2 \cdot (1 - P(t < t_{\text{набл}})).$$

Точно посчитать такое p-value без компьютера или специального калькулятора не получится, можно только примерно оценить его значение по таблице распределения Стьюдента. Для этого нужно знать, какое число степеней свободы имеет нужное нам t-распределение:

$$\text{df} = n - 2.$$

Если оценить таким образом p-value не получается, можно пойти другим путем: посчитать критическое значение t-статистики, определить критическую область (область нетипичных при верной нулевой гипотезе значений t-статистики) и понять, попадает ли полученное $t_{\text{набл}}$ в эту область. Получаем следующее:

$$t_{\text{крит}} = t_{(1-\frac{\alpha}{2},\ \text{df}=n-2)}, \text{где } \alpha \text{ – уровень значимости.}$$

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

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

Коэффициент корреляции Ч. Спирмена

Используется для выявления связи между двумя показателями, измеренными в порядковой (ординальной) шкале[1]. Можно использовать и для выявления связи между показателями, измеренными в количественной шкале; более того, данный коэффициент уместно вычислять в случае, когда при рассмотрении совместного распределения данных в выборках обнаруживаются нетипичные значения. Коэффициент корреляции Ч.Спирмена является более устойчивым к выбросам по сравнению с коэффициентом корреляции К.Пирсона.

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

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

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.

[1]Примеры показателей: места в рейтинге, экспертные оценки от 1 до 5, оценки студентов от 1 до 10.

$R_{\text{Спирмена}}$ также лежит в интервале $[-1; 1]$, если $R > 0$ – согласованность рангов прямая, если $R < 0$ – согласованность рангов обратная, $R = 0$ – согласованности между рангами нет.

- **Проверка гипотезы о независимости признаков**
  Нулевая и альтернативная гипотезы:

$$H_0 : \text{признаки независимы (не связаны)}$$

$$H_1 : \text{принаки не являются независимыми (связаны)}$$

  Наблюдаемое значение z-статистики:

$$z_{\text{набл}} = R_{\text{Спирмена}}\sqrt{n-1},$$

  где $R_{\text{Спирмена}}$ – коэффициент корреляции Спирмена и n – число наблюдений в выборке.
  Соответствующее p-value:

$$\text{p-value} = P(|z| > z_{\text{набл}}) = 2 \cdot P(z > z_{\text{набл}}) = 2 \cdot (1 - P(z < z_{\text{набл}})) = 2 \cdot (1 - \Phi(z_{\text{набл}})).$$

  По полученному p-value, как обычно, можем сделать вывод о нулевой гипотезе, учитывая выбранный уровень значимости.

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

Используются 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

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

первой строке (одна строка, все столбцы), а $n_{.1} = 55$ – сумма по первому столбцу (один столбец, все строки). N – сумма всех значений в таблице.

$$n_{11} = 20 \text{ и } n_{12} = 15 \text{ и } n_{21} = 35 \text{ и } n_{22} = 20.$$

- **Проверка гипотезы о независимости признаков**
  Нулевая и альтернативная гипотезы:

$$H_0 : \text{признаки независимы (связи нет)}$$

$$H_1 : \text{признаки не независимы (связь есть)}$$

Статистика используемого критерия имеет распределение хи-квадрат (χ^2) с числом степеней свободы $\text{df} = (r-1)\cdot(c-1)$, где r и c – число строк и столбцов в таблице сопряженности соответственно. В случае таблицы 2×2 все проще: распределение χ^2 с $\text{df} = 1$ – это просто Z^2, квадрат стандартного нормального распределения. Наблюдаемое значение статистики считается следующим образом:

$$z^2_{\text{набл}} = \frac{(n_{11}\cdot n_{22} - n_{21}\cdot n_{12})^2 \cdot N}{n_{1.}\cdot n_{.1}\cdot n_{2.}\cdot n_{.2}}.$$

Посчитаем для нашего случая:

$$z^2_{\text{набл}} = \frac{(20\cdot 20 - 15\cdot 35)^2 \cdot 90}{35\cdot 55\cdot 55\cdot 35} \approx 0.39.$$

Соответствующее p-value:

$$\text{p-value} = P(z^2 > z^2_{\text{набл}}) = P(z^2 > 0.39) = P(|z| > \sqrt{0.39}) = 2\cdot P(z > 0.62) \approx 0.53.$$

На любом разумном уровне значимости нет оснований отвергнуть H_0. Признаки независимы.