ОП «Политология», 2019-20
Математика и статистика, часть 2
Коэффициенты корреляции. (28.05.2020)
А. А. Макаров, А. А. Тамбовцева, Н. А. Василёнок

1. Коэффициент корреляции К.Пирсона

Используется для выявления линейной связи между двумя показателями, измеренными в количественной шкале.

Расчет коэффициента корреляции

Коэффициент корреляции Пирсона между выборками x и y считается так:

$$R = \frac{cov(x, y)}{s_y \cdot s_y},$$

где $cov(x, y)$ – выборочная ковариация x и y, а s_x и s_y – стандартные отклонения выборок x и y.

Как и любой коэффициент корреляции, R принадлежит интервалу $[-1; 1]$. Если $R > 0$, то связь между показателями прямая, если $R < 0$, то связь между показателями обратная, если $R = 0$, то линейной связи между показателями нет.

Проверка гипотезы о равенстве теоретического коэффициента корреляции нулю

Статистические гипотезы:

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

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

Статистика критерия имеет распределение Стьюдента с $\text{df} = n - 2$, где n – число наблюдений в выборке. Наблюдаемое значение статистики считается так:

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

где R – коэффициент Пирсона, а n – число наблюдений в выборке.

Обычно выбирается двусторонняя альтернативная гипотеза, как H_1 в формулировке выше, поэтому p-value выглядит так:

$$\text{p-value} = \text{P}(t < -t_{\text{набл}}) + \text{P}(t > t_{\text{набл}}).$$

Если p-value больше уровня значимости α, то H_0 не отвергается на этом уровне значимости, поэтому у нас есть основания считать, что связи нет. Если p-value меньше уровня значимости α, то H_0 отвергается на этом уровне значимости, поэтому у нас есть основания считать, что связь есть.

2. Коэффициент корреляции Ч.Спирмена

Используется для выявления связи между двумя показателями, когда хотя бы один из них измерен в порядковой шкале. Можно использовать и для выявления связи между показателями, измеренными в количественной шкале. Коэффициент корреляции Ч.Спирмена уместно вычислять в случае, когда в совместном распределении выборок присутствуют нетипичные значения, так как он является более устойчивым по сравнению с коэффициентом корреляции К.Пирсона.

Расчет коэффициента корреляции

Коэффициент корреляции Спирмена между выборками x и y считается так:

$$R_{\text{Спирмена}} = 1 - \frac{6 \cdot \sum\limits_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot (d_1^2 + d_2^2 + \cdots + d_n^2)}{n \cdot (n^2 - 1)},$$

где d_i – разность между рангом i-того наблюдения в выборке x и рангом i-того наблюдения в выборке y, n – число элементов в выборке.

Как и любой коэффициент корреляции, $R_{\text{Спирмена}}$ принадлежит интервалу $[-1;\ 1]$. Если $R > 0$, то согласованность рангов прямая, если $R < 0$, то согласованность рангов обратная, если $R = 0$, то связи между рангами нет.

Проверка гипотезы о независимости признаков

Статистические гипотезы:

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

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

Статистика критерия имеет стандартное нормальное распределение $N(0, 1)$. Наблюдаемое значение статистики считается так:

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

где n – число элементов в выборке.

Обычно выбирается двусторонняя альтернативная гипотеза, как H_1 в формулировке выше, поэтому p-value выглядит так:

$$\text{p-value} = \text{P}(z < -z_{\text{набл}}) + \text{P}(z > z_{\text{набл}}).$$

Если p-value больше уровня значимости α, то H_0 не отвергается на этом уровне значимости, поэтому у нас есть основания считать, что признаки независимы (связи нет). Если p-value меньше уровня значимости α, то H_0 отвергается на этом уровне значимости, поэтому у нас есть основания считать, что признаки не независимы (связь есть).