Коэффициенты корреляции

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.

Вычисление коэффициента корреляции

Формула расчета:

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

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

Если выборочная ковариация и выборочные дисперсии нам уже известны, вычисления упрощаются:

$$R = \frac{cov(x, y)}{\sqrt{s_x^2}\sqrt{s_y^2}} = \frac{cov(x, y)}{s_x \cdot s_y},$$

где $cov(x, y)$ – выборочная оценка ковариации двух выборок, s_x^2 и s_y^2 – дисперсии первой и второй выборки соответственно.

Значения R лежат в интервале $[-1; 1]$, если $R > 0$ – связь между показателями прямая, если $R < 0$ – связь между показателями обратная, $R = 0$ – линейной связи между показателями нет.

Проверка гипотезы о равенстве теоретического коэффициента корреляции нулю

Нулевая и альтернативная гипотезы:

$$H_0 : r = 0 \text{ (связи нет)}$$

$$H_1 : r \neq 0 \text{ (связь есть)}$$

Наблюдаемое и критическое значение статистики критерия:

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

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

Если проверяем H_0 через построение критической области:

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

Если проверяем H_0 через p-value:

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

Далее сравниваем p-value с уровнем значимости α: если p-value меньше уровня значимости α, H_0 отвергается (связь есть), если больше – не отвергается (связи нет).

Коэффициент корреляции Ч.Спирмена

Используется для выявления монотонной (необязательно линейной) связи между двумя показателями, измеренными в порядковой шкале. Можно использовать и для выявления связи между показателями, измеренными в количественной шкале; более того, данный коэффициент уместно вычислять в случае, когда в совместном распределении 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

Формула расчета:

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

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, $R = 0$ – связи между рангами нет.

Проверка гипотезы о независимости признаков

Нулевая и альтернативная гипотезы:

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

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

Наблюдаемое и критическое значение статистики критерия:

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

$$z_{\text{крит}} = z_{(1-\frac{\alpha}{2})}$$

Если проверяем H_0 через построение критической области:

- $|z_{\text{набл}}| > z_{\text{крит}} \Rightarrow H_0$ отвергается, связь между показателями есть;
- $|z_{\text{набл}}| < z_{\text{крит}} \Rightarrow H_0$ не отвергается, связи между показателями нет.

Если проверяем H_0 через p-value:

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

Далее сравниваем p-value с уровнем значимости α: если p-value меньше уровня значимости α, H_0 отвергается (связь есть), если больше – не отвергается (связи нет).