ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА ДЛЯ ПРИМЕНЕНИЯ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Лекция 4

Корреляционный анализ позволяет не только установить наличие зависимости между случайными величинами, но и дать качественную характеристику этой связи.

В качестве такой меры служит **коэффициент корреляции**.

Различают следующие виды коэффициентов корреляции:

1. парный линейный выборочный коэффициент корреляции r_{xy};
2. корреляционное отношение η_{xy};
3. множественный коэффициент корреляции $R_{i.jklm}\ldots$ и частный выборочный коэффициент корреляции $r_{ij.klm}\ldots$;
4. ранговые коэффициенты корреляции Спирмена и Кендалла.

Корреляционное отношение

Корреляционное отношение позволяет выявить наличие или отсутствие связи между случайными величинами X и Y.

Определяется корреляционное отношение на основе межгрупповой и общей дисперсий измеряемой величины (принимаем, что на изменчивость случайной величины Y влияют значения случайной величины X).

$$\eta_{yx} = \sqrt{\frac{s^2_{\text{межгруп.по } y}}{s^2_{\text{общ.по } y}}}.$$

Определение общей дисперсии переменной $Y^2_{\text{общ.по } y\ s}$ производится по формуле:

$$s^2_{\text{общ.по } y} = \frac{\sum_{j=1}^{m} \left(y_j - \bar{y}\right)^2 n_j}{N},$$

где y – среднее значение (математическое ожидание) случайной величины Y, оцененное по экспериментальным данным;

n_j – частота встречи значения y_j ;

m – общее число значений y_j ;

N – общее число проведенных экспериментов.

Для оценки межгрупповой дисперсии переменной Y $s^2_{\text{межгруп.по } y}$ необходимо произвести группировку значений переменной Y в зависимости от значений переменной X, т.е. отдельно «собрать» все y_j , которые были отмечены при значении x_1, отдельно «собрать» все y_j , которые были отмечены при значении x_2 и т.д. По каждой полученной группе оценить средние значения величины y, обозначив их y_i .

Межгрупповая дисперсия:

$$s^2_{\text{межгруп.по } y} = \frac{\sum_{i=1}^{k} \left(\bar{y}_i - \bar{y}\right)^2 n_i}{N},$$

где $\bar{y}$ – среднее значение (математическое ожидание) случайной величины Y, оцененное по экспериментальным данным;

$\bar{y}_i$ – групповые средние значения (математические ожидания) случайной величины Y, оцененные по экспериментальным данным, сгруппированным по значением случайной величины X;

n_i – частота встречи значения x_i;

k – общее число значений x_i;

N – общее число проведенных экспериментов.

Свойства корреляционного отношения:

1. $0 \le \eta_{yx} \le 1$.

Причем:

- $\eta_{yx} = 1$ – наличие функциональной зависимости между случайными величинами X и Y;
- $\eta_{yx} = 0$ – отсутствие какой-либо связи между случайными величинами X и Y;
- $0 < \eta_{yx} < 1$ - наличие статистической связи между случайными величинами X и Y.

2. $\eta_{yx} \ne \eta_{xy}$.

Фактически, после определения $r_{xy}=0$ необходимо оценить корреляционное отношение, и только по результатам последнего уже выносить «приговор» зависимости между двумя случайными величинами:

a) если $r_{xy} = 0$, $\eta_{yx} = 1$ – между случайными величинами X и Y наблюдается функциональная зависимость, но она носит нелинейный характер;

b) если $r_{xy} = 0$, $\eta_{yx} = 0$ – между случайными величинами X и Y не наблюдается какой-либо зависимости.

Значимость корреляционного отношения определяется по критерию согласия Фишера – Снедекора.

Наблюдаемое значение критерия определяется по формуле:

$$K_\eta = \frac{\eta_{yx}^2 \cdot (N - m)}{\left(1 - \eta_{yx}^2\right)(m - 1)},$$

где N – общее число опытов;

m – число полученных групп при определении межгрупповой дисперсии (фактически, это число значений случайной величины X).

Корреляционное отношение признается значимым (т.е. основная гипотеза отвергается), если

$$K_\eta \ge F_\alpha(m - 1; N - m)$$

Множественный коэффициент корреляции и частный выборочный коэффициент корреляции

Множественный коэффициент корреляции, равно как и частный выборочный коэффициент корреляции, определяются в случае выявления зависимостей между случайными величинами, чье количество превышает два.

Разница между этими двумя коэффициентами состоит в следующем:

1) множественный коэффициент корреляции оценивает влияние нескольких (больше двух) факторов на параметр оптимизации;
2) частный выборочный коэффициент корреляции оценивает зависимость между двумя параметрами (между двумя факторами, между фактором и параметром оптимизации и т.п.) при исключении влияния остальных параметров взаимодействия.

При взаимодействии нескольких случайных величин обычно строится корреляционная матрица, членами которой являются парные выборочные линейные коэффициенты корреляции между взаимодействующими случайными величинами.

По главной диагонали данной матрицы располагаются единицы, а сама матрица – симметрична относительно главной диагонали.

$$q = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{21} & 1 & r_{23} & \dots & r_{2k} \\ r_{31} & r_{32} & 1 & \dots & r_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & r_{k3} & \dots & 1 \end{bmatrix}.$$

Множественный коэффициент корреляции определяется по формуле

$$R_{j.1,2,\ldots(j-1),(j+1),\ldots k} = \sqrt{1 - \frac{|q|}{q_{jj}}},$$

(Данное обозначение читается следующим образом: «коэффициент корреляции на случайную величину J случайных величин 1, 2, …, K»)

где

IqI - определитель корреляционной матрицы;

q_{jj} – алгебраическое дополнение соответствующего элемента корреляционной матрицы.

Наблюдаемое значение критерия определяется по формуле:

$$K_R = \frac{R^2 \cdot (N - k)}{(1 - R^2)(k - 1)},$$

где N – общее число опытов;

k – число переменных во взаимодействии.

Множественный коэффициент корреляции признается значимым, если

$$K_R \geq F_\alpha(k - 1; N - k).$$

Частный выборочный коэффициент корреляции определяется как

$$r_{ij.1,\ldots,k} = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}},$$

где q_{ij}, q_{ii} q_{jj} – алгебраические дополнения соответствующих элементов корреляционной матрицы.

(Данное обозначение следует читать как «взаимодействие между случайными величинами I и J при исключении влияния остальных случайных величин»)

Значимость частного выборочного коэффициента корреляции определяется по критерию согласия Стьюдента.

$$K_{r.1} = \frac{r_{ij.1,\ldots,k}\sqrt{N-2}}{\sqrt{1-r^2_{ij.1,\ldots,k}}},$$

где N – общее число опытов.

Частный выборочный коэффициент корреляции признается значимым , если

$$\left|K_{r.1}\right| \geq t_\alpha\left(N-k+2\right),$$

где k – число переменных во взаимодействии.

Свойства множественного и частного коэффициентов корреляции совпадают со свойствами корреляционного отношения и парного линейного выборочного коэффициента корреляции соответственно.

Ранговые коэффициенты корреляции

Все перечисленные выше коэффициенты корреляции, несмотря на всю свою необходимость, не позволяют, однако, оценивать зависимости качественных переменных.

В лучшем случае качественные показатели можно подвергнуть процедуре ранжировки, но это не сделает их количественными, а значит – применять описанные выше показатели связи нельзя.

Для оценки ранжированных переменных существуют свои коэффициенты корреляции: коэффициенты Спирмена и Кендалла.

Оба эти коэффициента оценивают совпадение (или не совпадение) рангов двух совокупностей по одному ранжируемому признаку.

Коэффициент ранговой корреляции Спирмена

Для того, чтобы оценить коэффициент ранговой корреляции Спирмена, необходимо, прежде всего, определиться по какому признаку будет производиться ранжирование.

Затем провести оценку рангов по этому признаку для двух совокупностей.

Коэффициент ранговой корреляции Спирмена определяется по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n}$$

где r_i , s_i – ранги i-го объекта по совокупностям X и Y;
n – число пар наблюдений.

Иногда при исследованиях сталкиваются со случаями, когда для разных значений признака ранжирования в одной совокупности существуют одинаковые ранговые значение.

Такие случаи называются случаями со связанными рангами.

Если невозможно решить, какие ранги приписать этим объектам, им всем приписывается одинаковый средний ранг.

В случае связанных рангов коэффициент Спирмена вычисляется

$$\rho = 1 - \frac{\sum_{i=1}^{n}(r_i - s_i)^2}{\frac{1}{6}\left(n^3 - n\right) - \left(T_r + T_s\right)},$$

где $T_r = \frac{1}{12}\sum_{i=1}^{m_r}\left(t_r^3 - t_r\right)$, $T_s = \frac{1}{12}\sum_{i=1}^{m_s}\left(t_s^3 - t_s\right)$ – поправочные коэффициенты;

m_r, m_s – число групп неразличимых рангов у первой и второй совокупности соответственно;

t_r, t_s – число рангов, вошедших в соответствующую группу.

Оценка значимости коэффициента ранговой корреляции Спирмена, независимо от того, по какой из двух формул он вычислялся, производится по критерию согласия Стьюдента.

Наблюдаемое значение критерия определяется по формуле:

$$K_\rho = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}},$$

где n –число пар наблюдений.

Ранговый коэффициент корреляции Спирмена признается значимым (т.е. основная гипотеза отвергается), если

$$\left|K_\rho\right| \geq t_\alpha(n-2).$$

Коэффициент ранговой корреляции Кендалла

Для того чтобы оценить коэффициент ранговой корреляции Кендалла, необходимо провести ранжировку исследуемого объекта в порядке возрастания рангов по одной переменной и определить, сколько раз произошло нарушение порядка следования рангов по другой переменной (инверсия).

Инверсия – случай, когда большее число стоит слева от меньшего.

Величина К, называемая статистикой Кендалла, равна общему числу инверсий в ранговой последовательности.

Коэффициент ранговой корреляции Кендалла определяется по формуле:

$$\tau = 1 - \frac{4K}{n(n-1)}.$$

Оценка значимости коэффициента ранговой корреляции Кендалла производится по критерию согласия Стьюдента.

Наблюдаемое значение критерия определ

$$K_\tau = \tau\sqrt{\frac{9n(n-1)}{2(2n+5)}},$$

где n –число пар наблюдений.

Ранговый коэффициент корреляции Кендалла признается значимым, если

$$|K_\tau| \geq t_{1-\alpha},$$

где $t_{1-\alpha}$ определяется из выражения $\Phi(t_{1-\alpha}) = 1 - \alpha$;

$\Phi(t_{1-\alpha})$ – функция Лапласа.

Коэффициент конкордации рангов Кендалла

$$W = \frac{12\sum_{i=1}^{n}\left(\sum_{j=1}^{m} r_{ij} - \frac{m(n+1)}{2}\right)}{m^2\left(n^3 - n\right)},$$

где n – число объектов;

m – число анализируемых совокупностей.

Единственное условие для оценки коэффициента конкордации рангов Кендалла – число объектов $n \geq 7$.

$0 \leq W \leq 1$, причем $W = 1$, если все совокупности совпадают между собой по рангам.

Значимость коэффициента конкордации рангов Кендалла оценивается по критерию согласия Пирсона.

Наблюдаемое значение критерия определяется по формуле:

$K_W = m(n - 1)W$,

где n – число объектов;

m – число анализируемых совокупностей.

Коэффициент конкордации рангов Кендалла признается значимым, если

$$|K_W| \geq \chi^2_\alpha(n-1),$$

где

$\chi^2_\alpha(n-1)$ итическое значение χ^2-распределения Пирсона при уровне значимости α с числом степеней свободы (n – 1).