Эконометрика-1

Филатов Александр Юрьевич

(Главный научный сотрудник, доцент ШЭМ ДВФУ)

alexander.filatov@gmail.com

http://vk.com/alexander.filatov, **http://vk.com/baikalreadings**

Лекции 2.1-2.2

Множественные связи.

Порядковые и категоризованные переменные

Линейная зависимость от нескольких объясняющих переменных

2

Парные коэффициенты корреляции $r_{yx}^{(i)}$ не учитывают влияние на эту связь других переменных $x^{(j)}$. Следовательно, **необходим измеритель связи, очищенный от опосредованного влияния других переменных**, т.е. дающий оценку тесноты связи между y и $x^{(i)}$ при условии, что остальные переменные зафиксированы на некотором постоянном уровне.

Предположение: простой (линейный) характер влияния всех остальных переменных на y: $y = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)} + \varepsilon(X)$.

Обозначим для удобства $\boldsymbol{y \equiv x^{(0)}}$.

$$R = \begin{pmatrix} 1 & r_{01} & r_{02} & \ldots & r_{0p} \\ r_{10} & 1 & r_{12} & \ldots & r_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{p0} & r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}$$

R_{ij} – алгебраическое дополнение для r_{ij} в определителе корреляционной матрицы.

$R_{ij} = (-1)^{i+j} \det A_{ij}$, матрица A_{ij} получена из R вычеркиванием i-строки и j-столбца.

Частные (очищенные) коэффициенты корреляции

$r_{ij(-ij)} = \dfrac{-R_{ij}}{\left(R_{ii} R_{jj}\right)^{1/2}}$ – частный коэффициент корреляции, коэффициент кор-реляции между переменными $x^{(i)}$ и $x^{(j)}$ при фиксиро-ванных значениях всех остальных переменных.

Случай трех переменных:

$$\hat{R} = \begin{pmatrix} 1 & r_{01} & r_{02} \\ r_{01} & 1 & r_{12} \\ r_{02} & r_{12} & 1 \end{pmatrix} \qquad r_{01(2)} = \frac{r_{01} - r_{02} r_{12}}{\sqrt{\left(1 - r_{02}^2\right)\left(1 - r_{12}^2\right)}}$$

Свойства частных коэффициентов корреляции:

Проверка гипотезы о наличии/отсутствии связи, а также построение до-верительного интервала для частного коэффициента корреляции k-по-рядка (при исключении влияния k переменных) происходит по тем же формулам, что и для парного коэффициента корреляции с единственным отличием: **объем выборки уменьшается на k**.

Численные примеры

Пример 1:

n = 37 – число исследуемых предприятий легкой промышленности,

$x^{(0)} \equiv y$ – качество ткани (в баллах),

$x^{(1)}$ – среднемесячное число профилактических наладок автоматич. линии,

$x^{(2)}$ – среднемесячное число обрывов нити.

$\hat{r}_{01} = 0{,}105; \quad \hat{r}_{02} = 0{,}024; \quad \hat{r}_{12} = 0{,}996.$

$$\hat{R} = \begin{pmatrix} 1 & 0{,}105 & 0{,}024 \\ 0{,}105 & 1 & 0{,}996 \\ 0{,}024 & 0{,}996 & 1 \end{pmatrix}$$

$$r_{01(2)} = \frac{0{,}105 - 0{,}024 \cdot 0{,}996}{\sqrt{\left(1-0{,}024^2\right)\left(1-0{,}996^2\right)}} = 0{,}908,$$

$$r_{02(1)} = \frac{0{,}024 - 0{,}105 \cdot 0{,}996}{\sqrt{\left(1-0{,}105^2\right)\left(1-0{,}996^2\right)}} = -0{,}907.$$

Связь есть, что согласуется с профессиональными представлениями!

При нахождении доверительного интервала корректируем n = 37 – 1 = 36.

$$z = \text{ФИШЕР}(0{,}908) - \frac{0{,}908}{2(36-1)} = 1{,}502, \qquad z \in [1{,}161; 1{,}843],$$

$$z \in \left[1{,}502 - \frac{1{,}96}{\sqrt{36-3}}; 1{,}502 + \frac{1{,}96}{\sqrt{36-3}}\right], \qquad r \in [0{,}8214; 0{,}9511] \text{ при } \gamma = 0{,}95.$$

Численные примеры

Пример 2:

$n = 20$ – число лет метеонаблюдений,

$x^{(0)} \equiv y$ – урожайность кормовых трав,

$x^{(1)}$ – весеннее количество осадков,

$x^{(2)}$ – накопленная за весну сумма активных (выше +5,5°C) температур.

$\hat{r}_{01} = 0{,}80, \quad \hat{r}_{02} = -0{,}40, \quad \hat{r}_{01} = -0{,}56.$

$$\hat{R} = \begin{pmatrix} 1 & 0{,}80 & -0{,}40 \\ 0{,}80 & 1 & -0{,}56 \\ -0{,}40 & -0{,}56 & 1 \end{pmatrix}$$

$$r_{01(2)} = \frac{0{,}80 - 0{,}56 \cdot 0{,}40}{\sqrt{\left(1 - 0{,}56^2\right)\left(1 - 0{,}40^2\right)}} = 0{,}759,$$

$$r_{02(1)} = \frac{-0{,}40 + 0{,}80 \cdot 0{,}56}{\sqrt{\left(1 - 0{,}80^2\right)\left(1 - 0{,}56^2\right)}} = 0{,}097.$$

Связь со второй переменной не отрицательная, а слабая положительная, что согласуется с профессиональными представлениями!

$r_{01(2)} \in [0{,}448;\, 0{,}898]$, $r_{02(1)} \in [-0{,}376;\, 0{,}526]$ при $\gamma = 0{,}95$.

Множественный коэффициент корреляции

Множественный коэффициент корреляции – коэффициент корреляции между *y* и линейной функцией регрессии, т.е. между *y* и наилучшей ли-нейной комбинацией переменных $x^{(1)},\ldots,x^{(p)}$ – той, для которой значение коэффициента корреляции максимально.

$$R_{y.X} = r\left(y, f(X)\right) = r\left(y,\ \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}\right)$$

Свойства множественного коэффициента корреляции:

1. При предположении о линейности связи

$$K_d(y;X) = R^2_{y.X} = 1 - \frac{D\varepsilon}{Dy}.$$

2. Вычисление множественного коэффициента корреляции по корреля-ционной матрице:

$$R^2_{y.X} = 1 - \frac{|R|}{|R_{00}|}.$$

Множественный коэффициент корреляции

Свойства множественного коэффициента корреляции:

3. Вычисление МКК по частным коэффициентам корреляции:

$$R^2_{y.X} = 1 - \left(1 - r^2_{01}\right)\left(1 - r^2_{02(1)}\right)\left(1 - r^2_{03(12)}\right)\ldots\left(1 - r^2_{0p\left(123\ldots(p-1)\right)}\right)$$

$$\hat{R}^2_{y.\left(x^{(1)}x^{(2)}\right)} = 1 - \left(1 - \hat{r}^2_{01}\right)\left(1 - \hat{r}^2_{02(1)}\right) = 1 - \left(1 - 0{,}105^2\right)\left(1 - (-0{,}9068)^2\right) = 0{,}8243.$$

$$\hat{R}^2_{y.\left(x^{(1)}x^{(2)}\right)} = 1 - \left(1 - \hat{r}^2_{02}\right)\left(1 - \hat{r}^2_{01(2)}\right) = 1 - \left(1 - 0{,}024^2\right)\left(1 - 0{,}9079^2\right) = 0{,}8243.$$

4. МКК мажорирует все парные и частные КК, характеризующие статистическую связь: $R^2_{y.X} \geq r^2_{0j\left(I_j\right)}$, где I_j – любое подмножество $\{1,\ldots,p\}$, не содержащее j.
5. Присоединение новой переменной не может уменьшить величины R вне зависимости от порядка присоединения:

$$R^2_{y.x^{(1)}} \leq R^2_{y.\left(x^{(1)},x^{(2)}\right)} \leq R^2_{y.\left(x^{(1)},x^{(2)},x^{(3)}\right)} \leq \ldots \leq R^2_{y.\left(x^{(1)},x^{(2)},\ldots x^{(p)}\right)}.$$

Проверка гипотезы о наличии множественной линейной связи

Гипотеза о статистической независимости y и $x^{(1)},\ldots, x^{(p)}$ H_0: $R_{y.X} = 0$.

1. Выбираем уровень значимости α.

Типичные значения $\boldsymbol{\alpha = 0{,}05}$; 0,1; 0,01, 0,001.

2. Вычисляем эмпирическое значение критерия:

$$F_{\text{эмп}} = \frac{\hat{R}^2_{y.X}}{1-\hat{R}^2_{y.X}} \frac{n-p-1}{p}.$$

3. Вычисляем критическую точку:

$F_{\text{крит}} = F(\alpha;\, p;\, n-p-1) =$ FРАСПОБР $(\alpha;\, p;\, n-p-1)$.

4. Сравниваем эмпирическое и критическое значение и делаем вывод:

Если $F_{\text{эмп}} > F_{\text{крит}}$, то гипотеза H_0 об отсутствии множественной линейной связи отвергается при уровне значимости α, связь есть.

Корреляционный анализ порядковых переменных

$x^{(1)},\ldots,x^{(p)}$ – порядковые переменные (обозначающие порядковое место в ряду, отсортированному по соответствующему показателю).

Типовые задачи:

1. Анализ структуры упорядочений.

1. Точки разбросаны равномерно, нет согласованности между переменными.
2. Часть из p переменных близки между собой.
3. Часть из n объектов близки между собой.

2. Анализ совокупной согласованности переменных.

Исследование степени согласованности мнений экспертов.

3. Построение единого группового упорядочения объектов, т.е. ранжировки $x^{(0)}$, минимально удаленной от $x^{(1)},\ldots,x^{(p)}$.

Объединенные ранги:

Если есть неразличимые по некоторому свойству объекты, им всем приписывается единый ранг, равный среднему арифметическому.

Ранговый коэффициент корреляции Спирмена

Базовая формула:

$$\hat{r}_{kj}^{(S)} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} \left(x_i^{(k)} - x_i^{(j)} \right)^2 .$$

Формула для случая объединенных рангов:

$$\hat{r}_{kj}^{(S)} = \frac{1/6\left(n^3 - n\right) - \sum_{i=1}^{n} \left(x_i^{(k)} - x_i^{(j)} \right)^2 - T^{(k)} - T^{(j)}}{\sqrt{\left(1/6\left(n^3 - n\right) - 2T^{(k)}\right)\left(1/6\left(n^3 - n\right) - 2T^{(j)}\right)}},$$

$$T^{(k)} = \frac{1}{12} \sum_{t=1}^{m^{(k)}} \left(\left(n_t^{(k)}\right)^3 - n_t^{(k)} \right),$$

$m^{(k)}$ – число групп объединенных рангов,

$n_t^{(k)}$ – число элементов в каждой групп.

Свойства коэффициента Спирмена:

$\hat{r}_{kj}^{(S)} = 1$, если $x_i^{(k)} = x_i^{(j)}$, $i = 1,\dots,n$,

$\hat{r}_{kj}^{(S)} = -1$, если $x_i^{(k)} + x_i^{(j)} = n + 1$, $i = 1,\dots,n$,

$\hat{r}_{kj}^{(S)} \in \left(-1; 1\right)$ в остальных случаях.

Численные примеры

Пример 1:

10 инвестиционных проектов, проранжированных 2 экспертами.

$$x^{(1)} = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10)$$

$$x^{(2)} = (2\ 3\ 1\ 4\ 6\ 5\ 9\ 7\ 8\ 10)$$

$$\hat{r}_{12}^{(S)} = 1 - \frac{6}{990}\left(1 + 1 + 4 + \ldots + 0\right) =$$

$$= 1 - \frac{84}{990} = 0{,}915.$$

Пример 2:

10 стран, проранжированных по уровню жизни и качеству институтов.

$$x^{(1)} = (\ 1\ \ (2{,}5\ 2{,}5)(4{,}5\ 4{,}5)(6{,}5\ 6{,}5)\ 8\ (9{,}5\ 9{,}5))$$

$$x^{(2)} = ((1{,}5\ 1{,}5)(4{,}5\ 4{,}5\ 4{,}5\ 4{,}5)(8\ \ 8\ \ 8)\ 10)$$

$$T^{(1)} = \frac{1}{12}\left(2^3 - 2\right)\cdot 4 = \frac{24}{12} = 2,$$

$$T^{(2)} = \frac{1}{12}\left(\left(2^3 - 2\right) + \left(4^3 - 4\right) + \left(3^3 - 3\right)\right) = 7{,}5,$$

$$\hat{r}_{12}^{(S)} = \frac{\frac{990}{6} - \left(0{,}25 + 1 + \ldots + 0{,}25\right) - 2 - 7{,}5}{\sqrt{\left(\frac{990}{6} - 2\cdot 2\right)\left(\frac{990}{6} - 2\cdot 7{,}5\right)}} = 0{,}911.$$

Недостатки коэффициента Спирмена:

1. Недостаточная изученность статистических свойств.
2. Невозможность построения частных коэффициентов корреляции.
3. Необходимость полного пересчета при добавлении объекта.

Ранговый коэффициент корреляции Кендалла

Базовая формула:

$$\hat{r}_{kj}^{(K)} = 1 - \frac{4v\left(x^{(k)}, x^{(j)}\right)}{n(n-1)}, \quad v\left(x^{(k)}, x^{(j)}\right) -$$

минимальное число обменов соседних элементов переменной $x^{(j)}$ для ее приведения к виду $x^{(k)}$.

Свойства коэффициента Кендалла:

$\hat{r}_{kj}^{(K)} = 1$, если $x_i^{(k)} = x_i^{(j)}$, $i = 1,\ldots,n$,

$\hat{r}_{kj}^{(K)} = -1$, если $x_i^{(k)} + x_i^{(j)} = n+1$, $i = 1,\ldots,n$, $v = n(n-1)/2$,

$\hat{r}_{kj}^{(K)} \in (-1; 1)$ в остальных случаях.

Расчет числа обменов неудобен, *v* – также число инверсий (число расположенных в разном порядке пар элементов из $x^{(k)}$ и $x^{(j)}$.

Удобно произвести сортировку данных по одной из переменных!

$$\begin{pmatrix}4\\3\\2\\1\end{pmatrix} \to \begin{pmatrix}4\\3\\1\\2\end{pmatrix} \to \begin{pmatrix}4\\1\\3\\2\end{pmatrix} \to \begin{pmatrix}1\\4\\3\\2\end{pmatrix} \to \begin{pmatrix}1\\4\\2\\3\end{pmatrix} \to \begin{pmatrix}1\\2\\4\\3\end{pmatrix} \to \begin{pmatrix}1\\2\\3\\4\end{pmatrix} \Leftrightarrow \begin{pmatrix}1\\2\\3\\4\end{pmatrix}\begin{pmatrix}4\\3\\2\\1\end{pmatrix} \begin{matrix}3\\2\\1\\0\end{matrix}$$

$v = 6,$

$$\hat{r}_{12}^{(K)} = 1 - \frac{4 \cdot 6}{4 \cdot 3} = -1.$$

Ранговый коэффициент корреляции Кендалла

Формула для случая объединенных рангов:

$$\hat{r}_{kj}^{*(K)} = \frac{\hat{r}_{kj}^{(K)} - \dfrac{2\left(u^{(k)} + u^{(j)}\right)}{n(n-1)}}{\sqrt{\left(1 - \dfrac{2u^{(k)}}{n(n-1)}\right)\left(1 - \dfrac{2u^{(j)}}{n(n-1)}\right)}}, \qquad u^{(k)} = \frac{1}{2}\sum_{t=1}^{m^{(k)}} n_t^{(k)}\left(n_t^{(k)} - 1\right).$$

Пример 1:

$$x^{(1)} = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10)$$
$$x^{(2)} = (2\ 3\ 1\ 4\ 6\ 5\ 9\ 7\ 8\ 10)$$
$$1\ 1\ 0\ 0\ 1\ 0\ 2\ 0\ 0\ 0$$

$$\nu\left(x^{(1)}, x^{(2)}\right) = 5,$$

$$\hat{r}_{12}^{(K)} = 1 - \frac{4 \cdot 5}{90} = 0{,}778.$$

Пример 2:

$$x^{(1)} = (\ 1\ \ (2{,}5\ 2{,}5)(4{,}5\ 4{,}5)(6{,}5\ 6{,}5)\ 8\ (9{,}5\ 9{,}5))$$
$$x^{(2)} = ((1{,}5\ 1{,}5)(4{,}5\ 4{,}5\ 4{,}5\ 4{,}5)(8\ \ 8\ \ 8)\ \ 10)$$

$$u^{(1)} = \frac{1}{2}(4 \cdot 2 \cdot 1) = 4, \quad u^{(2)} = \frac{1}{2}(2 \cdot 1 + 4 \cdot 3 + 3 \cdot 2) = 10,$$

$$\hat{r}_{12}^{(K)} = \frac{1 - \dfrac{2 \cdot (4 + 10)}{90}}{\sqrt{\left(1 - \dfrac{2 \cdot 4}{90}\right)\left(1 - \dfrac{2 \cdot 10}{90}\right)}} = 0{,}818.$$

Проверка гипотезы о наличии связи между порядковыми переменными

Связь есть, если $\left|\hat{r}^{(S)}\right| > t\left(\frac{\alpha}{2}; n-2\right)\sqrt{\frac{1-\left(\hat{r}^{(S)}\right)^2}{n-2}}$

или $\left|\hat{r}^{(K)}\right| > u\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{2(2n+5)}{9n(n-1)}}.$

$0{,}915 > \text{СТЬЮДРАСПОБР}(0{,}025; 8)\sqrt{\left(1-0{,}915^2\right)/8}$, **0,915 > 0,392**.

$0{,}778 > \text{НОРМСТОБР}(0{,}975)\sqrt{\frac{2\cdot(2\cdot 10+5)}{9\cdot 10\cdot(10-1)}}$, **0,778 > 0,487**.

Неравенства утверждают, что связь есть при уровне значимости $\alpha = 0{,}05$.

Доверительный интервал для коэффициента Кендалла

$$r_{kj}^{(K)} \in \left[\hat{r}_{kj}^{(K)} - u\left(\frac{1+\gamma}{2}\right)\sqrt{\frac{2}{n}\left(1-\left(\hat{r}_{kj}^{(K)}\right)^2\right)};\ \hat{r}_{kj}^{(K)} + u\left(\frac{1+\gamma}{2}\right)\sqrt{\frac{2}{n}\left(1-\left(\hat{r}_{kj}^{(K)}\right)^2\right)}\right]$$

Интервал приближенный, формулу использовать для больших выборок!

Связь между несколькими порядковыми переменными

Коэффициент конкордации:

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

n – число объектов,
m – число переменных,
$k_1,\ldots,k_m$ – номера переменных.

$$\hat{W}(m)=\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{m}x_i^{(k_j)}-\frac{m(n+1)}{2}\right)^2}{\frac{1}{12}m^2\left(n^3-n\right)-m\sum_{j=1}^{m}T^{(k_j)}}$$

– при наличии объединенных рангов.

Свойства коэффициента конкордации:

$W(m) \in [0;1]$,
$W(m) = 1$ при полном совпадении переменных,
$W(m) = 0$, когда распределение случайно.

Коэффициент конкордации не может быть отрицательным:

если $x_i^{(2)}=n-x_i^{(1)}+1$, $x_i^{(3)}=n-x_i^{(1)}+1$, то $x_i^{(2)}=x_i^{(3)}$.

Численный пример

Ранжировка 10 инвестиционных проектов, осуществленная 3 экспертами.

$$
x^{(1)} \quad x^{(2)} \quad x^{(3)} \quad \Sigma \quad \Sigma - m\bar{x}
$$

$$
\begin{pmatrix}1\\4{,}5\\2\\3\\4{,}5\\7{,}5\\6\\9\\7{,}5\\10\end{pmatrix}
\begin{pmatrix}2{,}5\\1\\2{,}5\\4{,}5\\4{,}5\\8\\9\\6{,}5\\10\\6{,}5\end{pmatrix}
\begin{pmatrix}2\\1\\4{,}5\\4{,}5\\4{,}5\\4{,}5\\8\\8\\8\\10\end{pmatrix}
\begin{pmatrix}5{,}5\\6{,}5\\9\\12\\13{,}5\\20\\23\\23{,}5\\25{,}5\\26{,}5\end{pmatrix}
\begin{pmatrix}-11\\-10\\-7{,}5\\-4{,}5\\-3\\3{,}5\\6{,}5\\7\\9\\10\end{pmatrix}
$$

2 2 $\quad T^{(1)} = \frac{1}{12}\left(2^3 - 2\right)\cdot 2 = 1,$

2 2 2 $\quad T^{(2)} = \frac{1}{12}\left(2^3 - 2\right)\cdot 3 = 1{,}5,$

4 3 $\quad T^{(3)} = \frac{1}{12}\left(\left(4^3 - 4\right) + \left(3^3 - 3\right)\right) = 7.$

$$
\hat{W}(3) = \frac{\left((-11)^2 + (-10)^2 + \ldots + 10^2\right)}{\frac{1}{12}\cdot 3^2 \cdot \left(10^3 - 10\right) - 3\cdot\left(1 + 1{,}5 + 7\right)} = 0{,}828.
$$

Проверка гипотезы о наличии связи между несколькими порядковыми переменными

Связь есть, если $m(n-1)\hat{W}(m) > \chi^2(\alpha; n-1) = ХИ2ОБР(\alpha; n-1).$

Пример 1:

$\chi^2_{\text{эмп}} = 3 \cdot 9 \cdot 0{,}828 = 22{,}35, \quad \chi^2_{\text{крит}} = ХИ2ОБР(0{,}05; 9) = 16{,}92.$

$22{,}35 > 16{,}92 \quad \Rightarrow$ связь между 3 переменными есть при $\alpha = 0{,}05$.

Пример 2:

$m = 28, \; n = 13, \; \hat{W}(28) = 0{,}08$

$\chi^2_{\text{эмп}} = 28 \cdot 12 \cdot 0{,}08 = 26{,}88, \quad \chi^2_{\text{крит}} = ХИ2ОБР(0{,}05; 12) = 21{,}03$

$26{,}88 > 21{,}03 \quad \Rightarrow$ связь между 28 переменными есть при $\alpha = 0{,}05$.

Замечание: при большом количестве переменных даже малого значения коэффициента конкордации достаточно для вывода о наличии связи.

Корреляционный анализ категоризованных переменных

$x^{(1)}, x^{(2)}$ – **категоризованные переменные** (переменные, описываемые конечным числом состояний).

пол, социальная страта, сезон, фирма-производитель,…

Таблица сопряженности:

$$
\begin{pmatrix}
 & 1 & 2 & \dots & m_2 & & \\
1 & n_{11} & n_{12} & \dots & n_{1m_2} & n_{1\bullet} = \sum n_{1j} & w_{1\bullet} = n_{1\bullet}/n \\
2 & n_{21} & n_{22} & \dots & n_{2m_2} & n_{2\bullet} = \sum n_{2j} & w_{2\bullet} = n_{2\bullet}/n \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
m_1 & n_{m_1 1} & n_{m_1 2} & \dots & n_{m_1 m_2} & n_{m_1\bullet} = \sum n_{m_1 j} & w_{m_1\bullet} = n_{m_1\bullet}/n \\
 & n_{\bullet 1} = \sum n_{j1} & n_{\bullet 2} = \sum n_{j2} & \dots & n_{\bullet m_2} = \sum n_{jm_2} & n & \\
 & w_{\bullet 1} = n_{\bullet 1}/n & w_{\bullet 2} = n_{\bullet 2}/n & \dots & w_{\bullet m_2} = n_{\bullet m_2}/n & &
\end{pmatrix}
$$

Статистическая независимость переменных: $\frac{n_{ij}}{n} \approx \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}, \quad w_{ij} \approx w_{i\bullet} w_{\bullet j}.$

Чем больше отклонение, тем больше показатель связи: $\delta_{ij} = w_{ij} - w_{i\bullet} w_{\bullet j}.$

Случаи тесной связи и независимости переменных

$x^{(1)}$ – пол (муж/жен), $x^{(2)}$ – уровень зарплаты (высокая/низкая), $n = 100$.

$$\begin{pmatrix} 50 & 0 & 50 & 0{,}5 \\ 0 & 50 & 50 & 0{,}5 \\ 50 & 50 & 100 & \\ 0{,}5 & 0{,}5 & & \end{pmatrix}, \ \Delta = \begin{pmatrix} 0{,}25 & -0{,}25 \\ 0{,}25 & 0{,}25 \end{pmatrix}.$$

Максимально тесная связь, знание значения одной переменной позволяет восстановить значение другой.

$$\begin{pmatrix} 25 & 25 & 50 & 0{,}5 \\ 25 & 25 & 50 & 0{,}5 \\ 50 & 50 & 100 & \\ 0{,}5 & 0{,}5 & & \end{pmatrix}, \ \Delta = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Полное отсутствие связи, знание значения одной переменной не позволяет сделать никаких выводов о значении другой.

$$\begin{pmatrix} 28 & 42 & 70 & 0{,}7 \\ 12 & 18 & 30 & 0{,}3 \\ 40 & 60 & 100 & \\ 0{,}4 & 0{,}6 & & \end{pmatrix}, \ \Delta = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Полное отсутствие связи, знание значения одной переменной не позволяет сделать никаких выводов о значении другой.

Квадратичная сопряженность – характеристика тесноты связи

Квадратичная сопряженность: два способа расчета:

$$\hat{X}^2 = n\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\delta_{ij}^2}{w_{i\bullet}w_{\bullet j}} = n\left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 1\right), \quad \hat{X}^2 \in [0; +\infty)$$

Проверка гипотезы о наличии связи:

$\hat{X}^2 > \chi^2(\alpha; (m_1 - 1)(m_2 - 1)) \quad \Rightarrow$ связь есть при уровне значимости α.

Коэффициент Крамера:

Недостатком квадратичной сопряженности является неограниченность ее значений: при $n \to \infty$ $X^2 \to \infty$. Следовательно, желательно построить другой показатель, находящийся в привычном диапазоне [0; 1].

$$\hat{C} = \sqrt{\frac{\hat{X}^2}{n \min\{m_1 - 1; m_2 - 1\}}} \in [0; 1]$$

Численный пример

Зависимость оплаты труда (низкая; средняя; высокая) **от образования** (неполное среднее; среднее; среднее специальное; высшее; высшее со степенью), $n = 300$.

$$\begin{pmatrix} \boxed{\begin{matrix} 14 & 28 & 13 & 18 & 3 \\ 11 & 48 & 35 & 51 & 17 \\ 1 & 10 & 7 & 19 & 25 \end{matrix}} & \begin{matrix} 76 & 0{,}25 \\ 162 & 0{,}54 \\ 62 & 0{,}21 \end{matrix} \\ \begin{matrix} 26 & 86 & 55 & 88 & 45 \\ 0{,}09 & 0{,}29 & 0{,}18 & 0{,}29 & 0{,}15 \end{matrix} & \begin{matrix} 300 & \\ & \end{matrix} \end{pmatrix}$$

Равномерное распределение

$$\begin{pmatrix} 7 & 22 & 14 & 22 & 11 \\ 14 & 46 & 30 & 48 & 24 \\ 5 & 18 & 11 & 18 & 9 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} 0{,}025 & 0{,}021 & -0{,}013 & -0{,}014 & -0{,}028 \\ -0{,}010 & 0{,}005 & 0{,}018 & 0{,}012 & -0{,}024 \\ -0{,}015 & -0{,}026 & -0{,}015 & 0{,}003 & 0{,}052 \end{pmatrix}$$

$$\frac{\delta_{ij}^2}{w_{i\bullet} w_{\bullet j}} = \begin{pmatrix} 0{,}028 & 0{,}006 & 0{,}000 & 0{,}003 & 0{,}021 \\ 0{,}002 & 0{,}000 & 0{,}003 & 0{,}001 & 0{,}007 \\ 0{,}012 & 0{,}011 & 0{,}006 & 0{,}000 & 0{,}088 \end{pmatrix}$$

$$\hat{X}^2 = 300 \cdot 0{,}188 = 56{,}48, \quad \hat{C} = \sqrt{\frac{56{,}48}{300 \min\{4; 2\}}} = 0{,}307.$$

$X^2_{\text{крит}} = \text{ХИ2ОБР}(0{,}001; 4 \cdot 2) = 26{,}12$, **56,48 > 26,12** $\Rightarrow$ связь есть при α=0,001.

Спасибо за внимание!

alexander.filatov@gmail.com

http://vk.com/alexander.filatov, **http://vk.com/baikalreadings**