Корреляционный анализ

Основные задачи

1) Существует ли связь между факторами и откликом?
2) Оценить силу связи
3) Выявить факторы, оказывающие наибольшее влияние на отклик

Основное числовые характеристики

1) Выборочный коэффициент корреляции Пирсона
2) Ранговые коэффициенты корреляции
3) Выборочное корреляционное отношение
4) Множественный и частный коэффициенты корреляции

Шкалы измерений

Количественные признаки

1) Абсолютная шкала (начало отсчета+единицы измерения)

2) Шкала отношений (во сколько раз)

3) Интервальная шкала (на сколько единиц)

Качественные признаки

1) Порядковая шкала (ординальная)

2) Номинальная (шкала наименований)

Коэффициент корреляции Пирсона

ТВ

$$r_{xy} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

МС

$$\tilde{r}_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

1) X, Y независимы $\Rightarrow$

$$\Rightarrow r_{xy} = 0$$

2) $\left| r_{xy} \right| \le 1$

3) $Y = a + bX \iff$

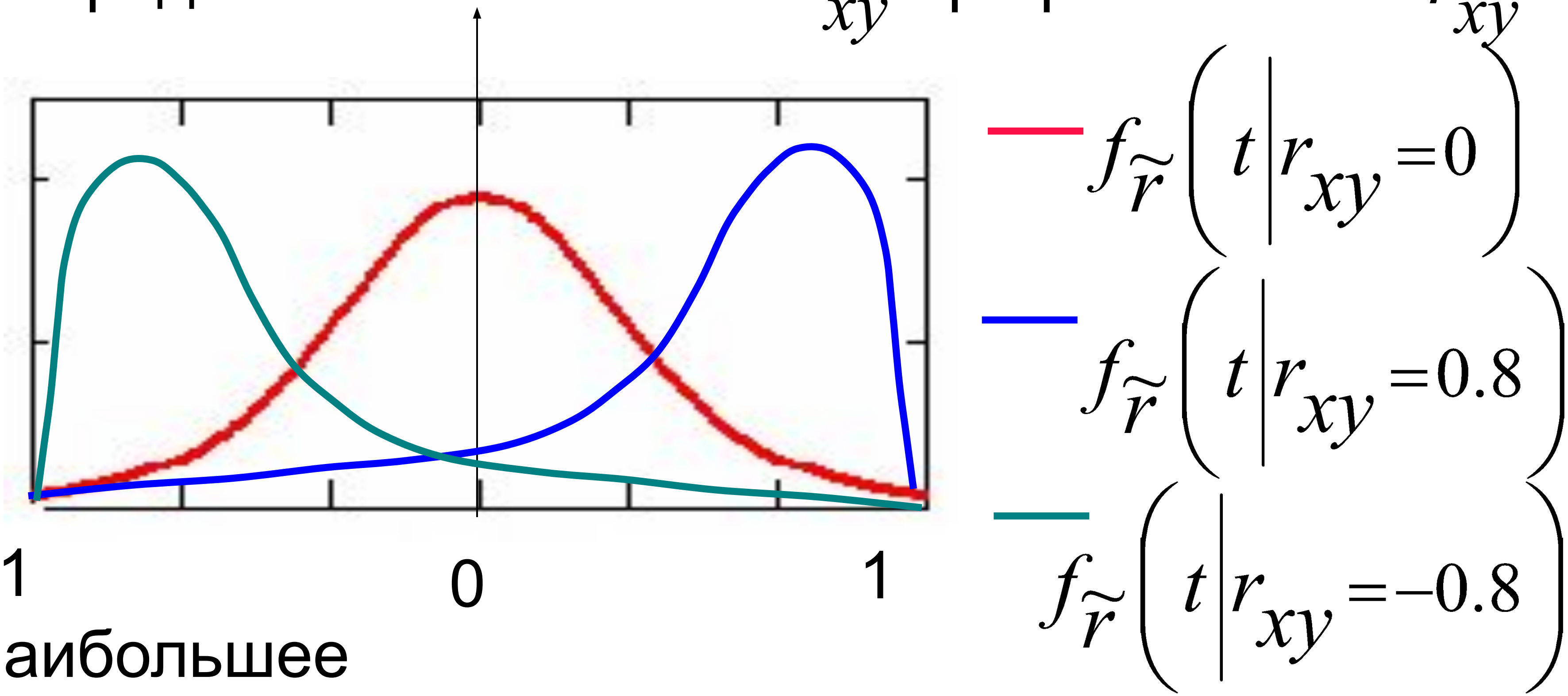
$$\iff \left| r_{xy} \right| = 1$$

Для 2НГС оценка асимптотически несмещенная, асимптотически эффективная

$$\boxed{\tilde{r}_{xy} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x} \cdot \bar{y}}{s_X s_Y}}$$

Коэффициент корреляции Пирсона

Распределение статистики $\tilde{r}_{xy}$ при различных r_{xy}



Наибольшее значение статистики $\tilde{r}_{xy}$ при значении $r_{xy} = 0$ $(\alpha = 0.05)$

$n-2$	5	10	20	30	50
$\tilde{r}_{\max}$	0.75	0.58	0.42	0.35	0.27

Коэффициент корреляции Пирсона

Задача 1 (о значимости связи)
(2НГС+некоррелированность)≡(независимость)

Гипотеза	$H_o:\ r_{xy}=0$
Требования	$2НГС,\quad n\geq 50$
Оценки	$\overline{x},\overline{y},s_x,s_y,\widetilde{r}_{xy}$
Статистика	$\dfrac{\widetilde{r}_{xy}\sqrt{n-2}}{\sqrt{1-\widetilde{r}_{xy}^{2}}}$
Распределение	$T_{(n-2)}$

Коэффициент корреляции Пирсона

Задача 1 (о значимости)

(2НГС+некоррелированность)≡(независимость)

Г	$H_o: \ r_{xy} = 0$
Т	$2НГС, \quad n \ge 50$
О	$\bar{x}, \bar{y}, s_x, s_y, \tilde{r}_{xy}$
С	$\dfrac{\tilde{r}_{xy}\sqrt{n-2}}{\sqrt{1-\tilde{r}_{xy}^2}}$
Р	$T_{(n-2)}$

Пример. $n = 66, \ \tilde{r}_{xy} = 0.4$

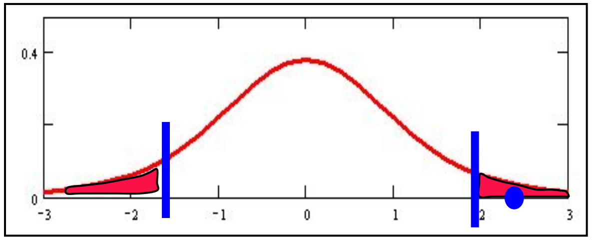
$H_o: \ r_{xy} = 0$
$H_1: \ r_{xy} \ne 0$

1) $\alpha = 0{,}05$

2) $t_{табл} = t_{(0.025; \ 60)} = 2.0$

3) $t_{набл} = \dfrac{0{,}4 \cdot 8}{\sqrt{1 - 0{,}16}} \approx 3{,}6$

4) $\left| t_{набл} \right| \ge t_{табл} \ \Rightarrow \ H_1$



Коэффициент корреляции Спирмена

а) ГС не является нормальной
б) X,Y – в порядковой шкале
в) n<50

$$\tilde{r}_{xy} = \frac{\frac{1}{n}\sum x_i y_i - \overline{x}\cdot\overline{y}}{s_x s_y}$$

$$\boxed{r_s = 1 - \frac{6\sum d_i^2}{n^3 - n}}$$

$$d_i = x_i' - y_i'$$

$$\Leftarrow \quad \Leftarrow \left\{ \begin{matrix} x_i \boxtimes x_i' \\ y_i \boxtimes y_i' \end{matrix} \right\} \Leftarrow \Downarrow$$

$$\begin{cases} H_0 : r_s = 0 \\ H_1 : r_s \neq 0 \end{cases}$$

H_0: связь незначима
H_1: связь значима

1) X, Y независимы $\Rightarrow r_s = 0$

2) $|r_s| \le 1$

3) X, Y – монотонная связь

$$\Leftrightarrow |r_s| = 1$$

$$\frac{r_s\sqrt{n-2}}{\sqrt{1 - r_s^2}} \sim T_{(n-2)}$$

Коэффициент корреляции Спирмена

Пример. *Y* – вес монеты (грамм), *X*- возраст (лет)

x_i	5	9	14	17	23	31	35	42	46	50
y_i	2.82	2.85	2.80	2.80	2.79	2.78	2.77	2.79	2.75	2.72

x_i'	1	2	3	4	5	6	7	8	9	10
y_i'	9	10	7.5	7.5	5.5	4	3	5.5	2	1
d_i	-8	-8	-4.5	-3.5	-0.5	2	4	2.5	7	9

$d_i = x_i' - y_i'$ $\quad \sum d_i^2 = 317$ $\quad r_s = -0.92$

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

Коэффициент корреляции Спирмена

Пример. Y – вес монеты (грамм), X- возраст (лет)

x_i	5	9	14	17	23	31	35	42	46	50
y_i	2.82	2.85	2.80	2.80	2.79	2.78	2.77	2.79	2.75	2.72

Вывод: между возрастом и весом монеты значимая монотонная связь (обратная)

H_0: связь незначима

H_1: связь значима

$$r_s = -0.92$$

$$t_{набл} = \frac{-0.92 \cdot \sqrt{8}}{\sqrt{1-0,85}} \approx -6.17$$

$$\frac{r_s \sqrt{n-2}}{\sqrt{1-r_{s9}^2}} \sim T_{(n-2)}$$

$$t_{табл} = t_{(0.025;\ 8)} = 2.31$$

Корреляционное отношение

Пример. Корреляционная таблица 1

X \ Y	3	5	7	n_x	$\overline{y}_x$
2	8	4	0	12	3.67
4	6	0	3	9	4.33
6	4	3	2	9	4.56
n_y	18	7	5	$n=30$	$\overline{y}=4.13$

$$\boxed{\tilde{\eta}_{yx}^2=\frac{s_{\text{межгр}}^2}{s_{\text{общ}}^2}}$$

$$s_{\text{межгр}}^2=\frac{1}{n}\sum n_x\left(\overline{y}_x-\overline{y}\right)^2$$

$$s_{\text{общ}}^2=\frac{1}{n}\sum n_y\left(y-\overline{y}\right)^2$$

Корреляционное отношение

ТВ

$$\eta_{yx}^2 = \frac{\sigma^2(y|x)}{\sigma_y^2}$$

$$\boxed{\tilde{\eta}_{yx} = \sqrt{\frac{s^2_{\text{межгр}}}{s^2_{\text{общ}}}}}$$

МС

$$\tilde{\eta}_{yx}^2 = \frac{s^2_{\text{межгр}}}{s^2_{\text{общ}}}$$

Свойства корреляционного отношения

1) $\eta_{yx}^2 \neq \eta_{xy}^2$

2) $0 \leq \eta_{yx} \leq 1$

3) $\eta_{yx} \geq |r_{yx}|$

4) $\eta_{yx} = 0 \Rightarrow Y, X$ независимы

5) $\eta_{yx} = 1 \Rightarrow$ связь функциональная

6) $\eta_{yx} = |r_{yx}| \Rightarrow$ связь линейная

Корреляционное отношение

Пример. Корреляционная таблица 2

X \ Y	3	5	7	n_x	$\bar{y}_x$
2	6	3	0	9	3.67
4	0	5	2	7	5.57
6	0	5	9	14	6.29
n_y	6	13	11	$n = 30$	$\bar{y} = 5.18$

	$s^2_{общ}$	$s^2_{межгр}$	η^2_{yx}	r^2_{yx}
Пример 1	2.32	0.15	0.066	0.061
Пример 2	2.18	1.29	0.594	0.588

Корреляционное отношение

<u>Задача 2. </u>Проверка линейности

Гипотеза	$H_O : \eta_{yx} = \lvert r_{yx} \rvert$
Требования	$2НГС, \quad n \geq 50$
Оценки	r_{yx}, η_{yx}
Статистика	$\dfrac{\left(\eta_{yx}^2 - r_{yx}^2\right)(n-r)}{\left(1-\eta_{yx}^2\right)(r-2)}$
Распределение	$F_{(r-2, n-r)}$

Корреляционное отношение

Задача 2. Проверка линейности

Г	$H_0: \eta_{yx} = \lvert r_{yx} \rvert$
Т	$2НГС, \quad n \geq 50$
О	r_{yx}, η_{yx}
С	$\dfrac{\left(\eta_{yx}^2 - r_{yx}^2\right)(n-r)}{\left(1-\eta_{yx}^2\right)(r-2)}$
Р	$F(r-2, n-r)$

$$n = 30, \quad r = 3$$

$$\eta_{yx}^2 = 0.594$$

$$r_{yx}^2 = 0.588$$

$$H_0: \eta_{yx} = \lvert r_{yx} \rvert$$

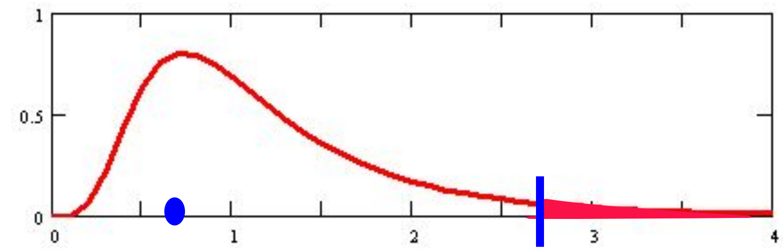
$$H_1: \eta_{yx} > \lvert r_{yx} \rvert$$

1) $\alpha = 0.05$

2) $F_{табл} = F\left(0.05; (2; 27)\right) = 3.35$

3) $F_{набл} = \dfrac{0,006 \cdot 27}{0.934 \cdot 1} \approx 0,173$

4) $F_{набл} < F_{табл} \Rightarrow H_0$



Многомерный корреляционный анализ

$$X_1, X_2, ..., X_p$$

1) Парные коэффициенты корреляции

$$Q = \begin{bmatrix} r_{11} & r_{12} & ... & r_{1p} \\ r_{21} & r_{22} & ... & r_{2p} \\ ... & ... & ... & ... \\ r_{p1} & r_{p2} & ... & r_{pp} \end{bmatrix}$$

2) Множественный коэффициент корреляции

$$R_{i,(1,...,i-1,i+1,...,p)} = \sqrt{1 - \frac{|Q|}{Q_{ii}}}$$

$$0 \le R \le 1$$

Q_{ii} - алгебраическое дополнение к r_{ii}

$$\frac{R^2(n-p)}{(1-R^2)(p-1)} \sim F_{(p-1;n-p)}$$

3) Частные коэффициенты корреляции

$$-1 \le r_{ij,(1...p)} \le 1$$

$$r_{ij,(1,...,p)} = \frac{-Q_{ij}}{\sqrt{Q_{ii} \cdot Q_{jj}}}$$

Многомерный корреляционный анализ

$$X_1, X_2, X_3$$

$$Q = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$$

1) Парные коэффициенты корреляции

2) Множественный коэффициент корреляции

$$R_{3,(1,2)} = \sqrt{\frac{r_{13}^2 - 2r_{12}r_{13}r_{23} + r_{23}^2}{1 - r_{12}^2}}$$

$$0 \le R_{3,(1,2)} \le 1$$

3) Частные коэффициенты корреляции

$$r_{13(2)} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{\left(1 - r_{12}^2\right) \cdot \left(1 - r_{32}^2\right)}}$$

$$-1 \le r_{13,(2)} \le 1$$

Многомерный корреляционный анализ

10 менеджеров оценивались 15 экспертами по 5-балльной шкале. Как связаны психологические характеристики между собой?

Результаты теста

Испыту емые п/п	тактичн ость	требова тельнос ть	критичн ость
1	70	18	36
2	60	17	29
3	70	22	40
4	46	10	12
5	58	16	31
6	69	18	32
7	32	9	13
8	62	18	35
9	46	15	30
10	62	22	36

Корреляционная матрица

	тактич ность	требов ательн ость	критичн ость
тактичность	1		
требовательность	0,86	1	
критичность	0,85	0,95	1

Частные коэффициенты корреляции

	тактич ность	требов ательн ость	критич ность
тактичность	1		
требовательность	0,32	1	
критичность	0,21	0,81	1