Лекция 12

Корреляционный анализ, часть II:
таблицы сопряжённости,
критерий независимости хи-квадрат
и коэффициент Крамера.

В прошлый раз

Мы рассмотрели коэффициенты корреляции Пирсона и Спирмена и научились с их помощью проверять гипотезу о независимости.

► коэффициент Пирсона измеряет тесноту линейной связи между количественными признаками;

► коэффициент Спирмена измеряет тесноту монотонной связи между количественными или порядковыми признаками.

Вопросы

► Как измерить тесноту связи между номинальными признаками?

► Как измерить тесноту немонотонной связи?

Напоминалка: независимость дискретных случайных величин

Дискретные с.в. X и Y независимы тогда и только тогда, когда

$$\mathrm{P}\left(\{X=x\}\cap\{Y=y\}\right)=\mathrm{P}\left(X=x\right)\mathrm{P}\left(Y=y\right)\quad\forall\, x,y\in R$$

Пример. Проверим независимость с.в. X и Y, совместное распределение которых задано таблицей:

$X\backslash Y$	-1	0	1
0	0.1	0.2	0.2
2	0.1	0.3	0.1

Напоминалка: независимость дискретных случайных величин

Дискретные с.в. X и Y независимы тогда и только тогда, когда

$$\mathrm{P}\left(\{X=x\}\cap\{Y=y\}\right)=\mathrm{P}\left(X=x\right)\mathrm{P}\left(Y=y\right)\quad\forall\, x, y\in R$$

Пример. Проверим независимость с.в. X и Y, совместное распределение которых задано таблицей:

$X\backslash Y$	-1	0	1	
0	0.1	0.2	0.2	→ 0.5
2	0.1	0.3	0.1	
	↓ 0.2			

Проверим первую клетку $\left(x=0,\ \ y=-1\right)$:

$P\left(\{X=0\}\cap\{Y=-1\}\right)=0.1=0.5\times0.2$ => выполнено

Напоминалка: независимость дискретных случайных величин

Дискретные с.в. X и Y независимы тогда и только тогда, когда

$$\mathrm{P}\big(\{X=x\}\cap\{Y=y\}\big)=\mathrm{P}\big(X=x\big)\mathrm{P}\big(Y=y\big)\quad\forall\, x,y\in R$$

Пример. Проверим независимость с.в. X и Y, совместное распределение которых задано таблицей:

$X\backslash Y$	-1	0	1	
0	0.1	0.2	0.2	→ 0.5
2	0.1	0.3	0.1	
	↓ 0.2	↓ 0.5		

Проверим первую клетку $\big(x=0,\ \ y=-1\big)$:

$P\big(\{X=0\}\cap\{Y=-1\}\big)=0.1=0.5\times0.2$ => выполнено.

Идём дальше $\big(x=0,\ \ y=0\big)$:

$P\big(\{X=0\}\cap\{Y=0\}\big)=0.2\neq0.5\times0.5$ => не выполнено.

Значит, X и Y не независимы.

Таблица сопряжённости

Выборка $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Конечное число возможных значений:

$$X_i \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_r \\ p_1 & p_2 & \ldots & p_r \end{pmatrix}; \qquad Y_i \sim \begin{pmatrix} y_1 & y_2 & \ldots & y_s \\ q_1 & q_2 & \ldots & q_s \end{pmatrix}.$$

Совместное распределение в выборке можно представить в виде таблицы:

$X \setminus Y$	y_1	y_2	...	y_s	Σ
x_1	O_{11}	O_{12}	...	O_{1s}	R_1
x_2	O_{21}	O_{22}	...	O_{2s}	R_2
...	...	...	...	...	...
x_r	O_{r1}	O_{r2}	...	O_{rs}	R_r
Σ	C_1	C_2	...	C_s	n

O_{ij} — наблюдаемые частоты

Таблица сопряжённости

Выборка $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Конечное число возможных значений:

$$X_i \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_r \\ p_1 & p_2 & \ldots & p_r \end{pmatrix}; \quad Y_i \sim \begin{pmatrix} y_1 & y_2 & \ldots & y_s \\ q_1 & q_2 & \ldots & q_s \end{pmatrix}.$$

Совместное распределение в выборке можно представить в виде таблицы:

$X \setminus Y$	y_1	y_2	...	y_s	Σ
x_1	O_{11}	O_{12}	...	O_{1s}	R_1
x_2	O_{21}	O_{22}	...	O_{2s}	R_2
...	...	...	...	...	...
x_r	O_{r1}	O_{r2}	...	O_{rs}	R_r
Σ	C_1	C_2	...	C_s	n

O_{ij} — наблюдаемые частоты

Пример: распределение шведских призывников 1926 г по цвету волос и бровей:

цвет бровей \ цвет волос	светлые или рыжие	темные	Σ
светлые или рыжие	30 472	3 238	33 710
темные	3 364	9 468	12 832
Σ	33 836	12 706	46 542

Критерий независимости хи-квадрат

Выборка из i.i.d. пар $(X_1, Y_1), \ldots, (X_n, Y_n)$ с распределением

$$X_i \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_r \\ p_1 & p_2 & \ldots & p_r \end{pmatrix}; \quad Y_i \sim \begin{pmatrix} y_1 & y_2 & \ldots & y_s \\ q_1 & q_2 & \ldots & q_s \end{pmatrix}.$$

Гипотезы:

$$\mathrm{H}_0: X_i \text{ и } Y_i \text{ независимы}$$
$$\mathrm{H}_\mathrm{A}: X_i \text{ и } Y_i \text{ зависимы}$$

Идея критерия: сравнивать наблюдаемые и ожидаемые частоты.

наблюдаемые частоты

$X \setminus Y$	y_1	y_2	...	y_s	Σ
x_1	O_{11}	O_{12}	...	O_{1s}	R_1
x_2	O_{21}	O_{22}	...	O_{2s}	R_2
...	...	...	...	...	...
x_r	O_{r1}	O_{r2}	...	O_{rs}	R_r
Σ	C_1	C_2	...	C_s	n

ожидаемые частоты

$X \setminus Y$	y_1	y_2	...	y_s	Σ
x_1					R_1
x_2		???			R_2
...					...
x_r					R_r
Σ	C_1	C_2	...	C_s	n

Ожидаемые частоты в критерии независимости

$X \setminus Y$	y_1	y_2	...	y_s	Σ
x_1					R_1
x_2		???			R_2
...					...
x_r					R_r
Σ	C_1	C_2	...	C_s	n

Оценим частные распределения X и Y:

$$\hat{p}_i = \hat{\mathrm{P}}(X = x_i) = \frac{R_i}{n}$$

$$\hat{q}_j = \hat{\mathrm{P}}(Y = y_j) = \frac{C_j}{n}$$

Оценка вероятности попасть в отдельную клетку:

$$\hat{\mathrm{P}}\left(\{X = x_i\} \cap \{Y = y_j\}\right) = \hat{p}_i \hat{q}_j = \frac{R_i C_j}{n^2}$$

Ожидаемая частота:

$$E_{ij} = n \cdot \frac{R_i C_j}{n^2} = \frac{R_i C_j}{n}$$

Критерий независимости хи-квадрат

Выборка из i.i.d. пар $(X_1, Y_1), \ldots, (X_n, Y_n)$ с распределением

$$X_i \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_r \\ p_1 & p_2 & \ldots & p_r \end{pmatrix}; \quad Y_i \sim \begin{pmatrix} y_1 & y_2 & \ldots & y_s \\ q_1 & q_2 & \ldots & q_s \end{pmatrix}.$$

$p_i > 0, \ i = 1, \ldots, r$
$q_j > 0, \ j = 1, \ldots, s$

это разные i

Гипотезы:

H_0: X_i и Y_i независимы
H_A: X_i и Y_i зависимы

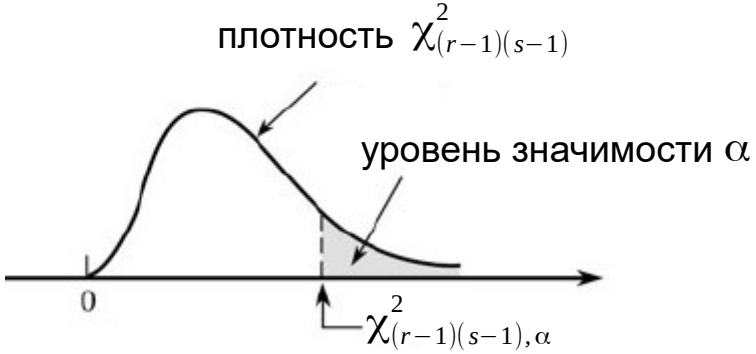
Статистика:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \underset{asy}{\overset{\mathrm{H}_0}{\sim}} \chi^2_{(r-1)(s-1)}$$

Решающее правило:

$\chi^2 > \chi^2_{(r-1)(s-1),\alpha}$ => H_0 отвергается в пользу H_A, выявлена связь между признаками

$\chi^2 < \chi^2_{(r-1)(s-1),\alpha}$ => нет оснований отвергнуть H_0, связь не выявлена.



Пример: руки, ноги и пол

Результаты обследования стоп у 40 мужчин и 87 женщин правшей*:

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	2	10	28	*40*
Женский	55	18	14	*87*
Σ	*57*	*28*	*42*	*127*

Проверим гипотезу о независимости пола и соотношения размеров стоп на уровне 1%.

* Levy J., Levy J.M. (1978). Human Lateralization from Head to Foot: Sex-Related Factors. *Science*, 200(4347): 1291–1292.

Пример: руки, ноги и пол

Результаты обследования стоп у 40 мужчин и 87 женщин правшей*:

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	2	10	28	*40*
Женский	55	18	14	*87*
Σ	*57*	*28*	*42*	*127*

Проверим гипотезу о независимости пола и соотношения размеров стоп на уровне 1%.

Ожидаемые частоты:

$$E_{11} = \frac{40 \times 57}{127} = 17.95; \quad E_{12} = \frac{40 \times 28}{127} = 8.82; \ldots$$

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	17.95	8.82	13.23	*40*
Женский	39.05	19.18	28.77	*87*
Σ	*57*	*28*	*42*	*127*

* Levy J., Levy J.M. (1978). Human Lateralization from Head to Foot: Sex-Related Factors. *Science*, 200(4347): 1291–1292.

Пример: руки, ноги и пол

Результаты обследования стоп у 40 мужчин и 87 женщин правшей*:

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	2	10	28	*40*
Женский	55	18	14	*87*
Σ	*57*	*28*	*42*	*127*

Проверим гипотезу о независимости пола и соотношения размеров стоп на уровне 1%.

Ожидаемые частоты:

$$E_{11} = \frac{40 \times 57}{127} = 17.95; \quad E_{12} = \frac{40 \times 28}{127} = 8.82; \ldots$$

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	17.95	8.82	13.23	*40*
Женский	39.05	19.18	28.77	*87*
Σ	*57*	*28*	*42*	*127*

Статистика:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(2-17.95)^2}{17.95} + \frac{(10-8.82)^2}{8.82} + \ldots + \frac{(14-28.77)^2}{28.77} = 45$$

Пример: руки, ноги и пол

Результаты обследования стоп у 40 мужчин и 87 женщин правшей*:

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	2	10	28	*40*
Женский	55	18	14	*87*
Σ	*57*	*28*	*42*	*127*

Проверим гипотезу о независимости пола и соотношения размеров стоп на уровне 1%.

Ожидаемые частоты:

$$E_{11} = \frac{40 \times 57}{127} = 17.95; \quad E_{12} = \frac{40 \times 28}{127} = 8.82; \ldots$$

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	17.95	8.82	13.23	*40*
Женский	39.05	19.18	28.77	*87*
Σ	*57*	*28*	*42*	*127*

Статистика:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(2-17.95)^2}{17.95} + \frac{(10-8.82)^2}{8.82} + \ldots + \frac{(14-28.77)^2}{28.77} = 45$$

$$\chi^2_{(r-1)(s-1),\alpha} = \chi^2_{(2-1)(3-1),0.01} = 9.21$$

Вывод: $\chi^2 > 9.21$ => гипотеза о независимости отвергается, связь есть.

Заметьте

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	2	10	28	*40*
Женский	55	18	14	*87*
Σ	*57*	*28*	*42*	*127*

Проверим гипотезу о независимости пола и соотношения размеров стоп на уровне 1%.

Основную гипотезу (независимость пола и соотношения стоп) можно было сформулировать иначе:

► Распределение правшей по соотношению стоп одинаково среди мужчин и женщин.

► Доли женщин среди «левоногих», «равноногих» и «правоногих» совпадают.

Критерием хи-квадрат можно проверять совпадение распределений в разных выборках!

Остатки

«Сырые» остатки (raw residuals): $e_{ij} = O_{ij} - E_{ij}$

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$
Мужской	–15.95	1.18	14.77
Женский	15.95	–1.18	–14.77

Остатки

«Сырые» остатки (raw residuals): $e_{ij} = O_{ij} - E_{ij}$

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$
Мужской	–15.95	1.18	14.77
Женский	15.95	–1.18	–14.77

Стандартизованные остатки, или остатки Пирсона: $e^*_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$
(standardized residuals, Pearson residuals)

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$
Мужской	–3.76	0.40	4.06
Женский	2.55	–0.27	–2.75

Иногда стандартизованным остатком называют чуть более сложную штуку.

Коэффициент Крамера

$$V = \sqrt{\frac{\chi^2}{n \min(r-1, s-1)}}$$

$$0 \le V \le 1$$

признаки независимы даже в выборке

наиболее тесная связь *(это какая?)*

Пол \ соотношение стоп	$L > R$	$L = R$	$L < R$	Σ
Мужской	2	10	28	*40*
Женский	55	18	14	*87*
Σ	*57*	*28*	*42*	*127*

$$V = \sqrt{\frac{45}{127 \times \min(2-1,\ 3-1)}} = 0.595$$

Коэффициент Крамера для шведских призывников

$$V = \sqrt{\frac{\chi^2}{n \min(r-1, s-1)}}$$

$$0 \le V \le 1$$

признаки независимы даже в выборке

наиболее тесная связь *(это какая?)*

цвет бровей \ цвет волос	светлые или рыжие	темные	$\sum$
светлые или рыжие	30 472	3 238	33 710
темные	3 364	9 468	12 832
$\sum$	33 836	12 706	46 542

$$V = 0.644$$

Когда $V = 1$?

Либо каждому значению X соответствует только оно значение Y, либо наоборот:

$X \setminus Y$	y_1	y_2
x_1	0	100
x_2	50	0

$X \setminus Y$	y_1	y_2	y_3
x_1	0	100	50
x_2	50	0	0

А тут $V = 0$:

$X \setminus Y$	y_1	y_2
x_1	10	50
x_2	40	200

Критерий независимости хи-квадрат и коэффициент Крамера пригодны для выявления *любого* вида статистической связи.

Зачем тогда нужно что-то ещё?

Применимость коэффициентов корреляции к разным типам признаков

	количественные	**порядковые**	**номинальные**
r **Пирсона**	да	нет	нет
r^s **Спирмена**	да	да	нет
V **Крамера**	да	да	да

Тут стоит сделать уточнение...

Уточнение: двоичные признаки

$X \setminus Y$	y_1	y_2
x_1	a	b
x_2	c	d

Упражнение 1. Докажите равенства:

$$r_{X,Y} = r^S_{X,Y}$$

$$V = |r_{X,Y}| = |r^S_{X,Y}|$$

Упражнение 2 *(разогрев)*. Докажите, что

$$r_{X,Y} = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}.$$

Мораль: для двоичных признаков имеет смысл обычный коэффициент корреляции.

В отличие от V Крамера, он показывает направление связи.

Тест на равенство медиан

Независимые выборки:

$$\left.\begin{matrix} X_1, X_2, \dots, X_{n_X} \\ Y_1, Y_2, \dots, Y_{n_Y} \end{matrix}\right\} \text{ все независимы}$$

У всех X_i совпадают медианы, и у всех Y_i совпадают медианы.

Гипотезы:

$$H_0 : \mathrm{Med}(X_i) = \mathrm{Med}(Y_i)$$

$$H_A : \mathrm{Med}(X_i) \neq \mathrm{Med}(Y_i)$$

Тест на равенство медиан

Независимые выборки:

$$\left.\begin{matrix} X_1, X_2, \dots, X_{n_X} \\ Y_1, Y_2, \dots, Y_{n_Y} \end{matrix}\right\}$$ все независимы

У всех X_i совпадают медианы, и у всех Y_i совпадают медианы.

Гипотезы:

$$H_0 : \mathrm{Med}(X_i) = \mathrm{Med}(Y_i)$$

$$H_A : \mathrm{Med}(X_i) \neq \mathrm{Med}(Y_i)$$

Алгоритм. 1) Ищем выборочную медиану по объединённым данным $X_1, \dots, X_{n_X}, Y_1, \dots, Y_{n_Y}$.

2) Составляем таблицу:

Выборка \ Значение	< общей медианы	> общей медианы
Выборка X	O_{11}	O_{12}
Выборка Y	O_{21}	O_{22}

3) Применяем критерий независимости хи-квадрат.

Пример

Оценки за контрольную по статистике: 95 юношей и 20 девушек.

Проверим гипотезу о равенстве медиан для юношей и девушек.

Общая медиана: 23 балла из 36.

Таблица сопряжённости:

Пол \ Оценка	< 23	≥ 23	Σ
девушки	7	13	*20*
юноши	49	46	*95*
Σ	*56*	*59*	*115*

Пример

Оценки за контрольную по статистике: 95 юношей и 20 девушек.

Проверим гипотезу о равенстве медиан для юношей и девушек.

Общая медиана: 23 балла из 36.

Таблица сопряжённости:

Пол \ Оценка	< 23	≥ 23	Σ
девушки	7	13	*20*
юноши	49	46	*95*
Σ	*56*	*59*	*115*

Статистика: $\chi^2 = 1.82.$

Критическое значение: $\chi^2_{1,0.05} = 3.84.$

Вывод: $\chi^2 < 3.84$ => нет оснований отвергнуть гипотезу о равенстве медиан.

Пример

Оценки за контрольную по статистике: 95 юношей и 20 девушек.

Проверим гипотезу о равенстве медиан для юношей и девушек.

Общая медиана: 23 балла из 36.

Таблица сопряжённости:

Пол \ Оценка	< 23	≥ 23	Σ
девушки	7	13	*20*
юноши	49	46	*95*
Σ	*56*	*59*	*115*

Статистика: $\chi^2 = 1.82$.

Критическое значение: $\chi^2_{1,0.05} = 3.84$.

Вывод: $\chi^2 < 3.84$ => нет оснований отвергнуть гипотезу о равенстве медиан.

Замечание. Критерий можно использовать для:

- ► сравнения более двух выборок;
- ► сравнения различных квантилей.

Подытожим

► Совместное распределение дискретных признаков представимо в виде таблицы сопряжённости.

► Проверить гипотезу о независимости можно критерием χ^2.

► Измерить тесноту связи можно коэффициентом Крамера V.

► Значения: $0 \leq V \leq 1$

независимость в выборке

одно- или двустороннее однозначное соответствие

► Критерием хи-квадрат можно проверить

- совпадение распределений в нескольких совокупностях;
- совпадение квантилей в нескольких совокупностях.

Следующая лекция

Регрессионный анализ. Ядерная оценка регрессии. Метод наименьших квадратов.