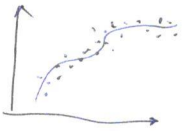
ИСПС | Корреляционный анализ |

Пусть мы изучаем некоторые объекты, кот. обладают несколькими признаками. Часто возникает задача изучения взаимосвязей этих признаков (и дальнейшие задачи отделения наиболее "внятельных" признаков и т.д.)
Чаще всего встречаются 2 вида зависимостей (естественно, случаи упрощенные)

а) объекты образуют "облако" эллиптич. типа

б) объекты располаг. в окрест. некоторой кривой



В случае а) оба признака являются полноценными, и изучению подлежит уровень зависимости (корреляции) между ними.
Случай б) соответствует "функциональной" зависимости между признаками, испорченной шумом. (этот случай изучается методами регресс. анализа)
А сегодня будем изучать первый случай с помощью методов корреляц. анализа

Сначала рассмотрим гипотезу независимости признаков ξ и η:

$H_0: F_{\xi,\eta}(x,y) = F_\xi(x) \cdot F_\eta(y)$, где $F_{\xi,\eta}(x,y) - $ ф. р. (ξ,η); $F_\xi(x)$ и $F_\eta(x) -$ ф. р. ξ и η соответственно

Будем считать, что есть данные, и признаки ξ и η реализованы в виде выборок $(X_1 \dots X_n)$ и $(Y_1 \dots Y_n)$
Первая мысль – воспольз. обычным коэфф. корреляции $\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{DX}\sqrt{DY}}$ (точнее, его выборочной характеристикой). (Увы, если $\rho = 0$, это не значит, что с.в. независимы. То же самое верно и для выборочного коэфф. корреляции. Подробнее об этом – на семинарах)

Коэфф. корреляции Пирсона

Опр. $\hat\rho = \dfrac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\left(\sum_{i=1}^n (X_i - \overline{X})^2 \cdot \sum_{i=1}^n (Y_i - \overline{Y})^2\right)^{1/2}}$.

Св-ва 1) Если $EX_i^2 < +\infty$ и $EY_i^2 < +\infty$, то $\hat\rho \xrightarrow{P} \rho$

2) Если H_0 верна (а, точнее, H_0' – гипотеза о некоррелируемости выборок), то $\dfrac{\hat\rho\sqrt{n-2}}{\sqrt{1-\hat\rho^2}} \to T_{n-2}$ (распред. Стьюдента)

3) Плохо реагирует на выбросы.

4) Для гаусс. выборок гипотезы H_0 и H_0' эквивалентны, для них $\hat\rho$ чаще всего и применяется.

Коэфф. корр. Спирмэна

Пусть R_i – ранг наблюдения X_i в выборке $(X_1 \dots X_n)$, т.е. $X_{(R_i)} = X_i$, порядк. статистика
S_j – ранг Y_j.

Опр. Коэфф. корр. Спирмэна: $\rho_S = \dfrac{\sum_{i=1}^n (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^n (R_i - \overline{R})^2 \sum_{j=1}^n (S_j - \overline{S})^2}}$, где $\overline{R} = \frac{1}{n}\sum_{i=1}^n R_i = \frac{n+1}{2} = \overline{S}$.

Св-ва ρ_S: 1) $\rho_S = 1 - \frac{6}{n^3-n}\sum_{i=1}^{n}(R_i - S_i)^2$

2) При верной H_0 $E\rho_S = 0$; $D\rho_S = \frac{1}{n-1}$

3) $-1 \le \rho_S \le 1$ (из Коши-Бун.), причём $\rho_S = 1$, если $R_i = S_i$ и $\rho_S = -1$, если $R_i + S_i = n+1$

4) Распред. ρ_S не зависит от F_x и F_y, его распред. табулировано.

5) При верной H_0: $\frac{\rho_S}{\sqrt{D\rho_S}} \xrightarrow{d} N(0,1)$, приближением можно пользоваться при $n \ge 50$

6) Если $n \le 50$, то лучше брать $\widetilde{\rho_S} = \frac{1}{2}\rho_S\left(\sqrt{n-1} + \sqrt{\frac{n-2}{1-\rho_S^2}}\right)$ (поправка Имана). Тогда критерий: отвергать H_0, если $\widetilde{\rho_S} \notin (z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}})$, где $z_\gamma = \frac{1}{2}(x_\gamma + y_\gamma)$, x_γ и y_γ – квантили $N(0,1)$ и T_{n-2} соответств.

7) Устойчив к выбросам (как и τ Кэндалла).

Коэфф. корреляции Кэндалла (τ Кэндалла)

Опр. Пары (X_i, Y_i) и (X_j, Y_j) – согласованы, если $\operatorname{sign}(X_i - X_j)(Y_i - Y_j) = 1$.

Обозн. S – число согласов. пар, R – число несогласов., $T = S - R = \sum_{i<j}\operatorname{sign}(X_i - X_j)(Y_i - Y_j)$

Легко видеть, что $-\frac{n(n-1)}{2} \le T \le \frac{n(n-1)}{2}$.

Опр. τ Кэндалла: $\tau = T \Big/ \frac{n(n-1)}{2}$.

Св-ва τ: 1) $\tau = 1 - \frac{4}{n(n-1)}R$

2) Если H_0 верна, то $E\tau = 0$ и $\frac{\tau}{\sqrt{D\tau}} \xrightarrow{d} N(0,1)$, где $D\tau = \frac{2(2n+5)}{9n(n-1)}$.

3) Верны ф-лы: $\tau = 1 - \frac{4}{n^2-n}\sum_{i<j} I(T_i > T_j)$ и $\rho_S = 1 - \frac{12}{n^3-n}\sum_{i<j}(j-i)\cdot I(T_j < T_i)$, где T_i вводится так: если $R_k = i$, то обознач. $T_i = S_k$.

4) ρ_S и τ сильно коррел., $\rho(\rho_S, \tau) > 0{,}98$ при $n > 5$.

Обобщенный коэфф. корреляции

Удобен для реализации на комп. коэфф. корреляции.

Определим отображение $C(\underset{\text{выборка}}{X}) = \|c_{ij}(X_i, X_j)\|_{i,j=1\dots n}$, где $c_{ij} = -c_{ji}$ и $c_{ii} = 0$.

Опр. Обобщ. коэфф. корр.: $\hat{\tau} = \frac{\sum_{i<j} c_{ij}(X)\cdot c_{ij}(Y)}{\sqrt{\sum_{i<j} c_{ij}^2(X)\cdot\sum_{i<j} c_{ij}^2(Y)}}$

Тогда 1) при $c_{ij}(X) = X_j - X_i$ $\hat{\tau} = \hat{\rho}$

2) при $c_{ij}(X) = R_j - R_i$ и $c_{ij}(Y) = S_j - S_i$ $\hat{\tau} = \rho_S$

3) при $c_{ij}(X) = \operatorname{sign}(R_j - R_i)$ $\hat{\tau} = \tau$.

Таблицы сопряжённости

Решаем задачу выявл. статистич. связи для сгрупп. данных (т.е. разбитых на категории). Данные: $\|\nu_{ij}\|_{\substack{i=1\dots n\\ j=1\dots m}}$, где ν_{ij} – числовые данные об объектах с призн. i,j.

Пример

пол \ возраст	≤18	18-35	≥35
М	0.2	0.5	0.9
*	1.6	2.5	0.4

← Потребление булочек в день

Опишем 2 схемы, кот. приводят к таблицам сопряжённости:

Схема I Пусть $(\nu_{i1},\dots\nu_{im})$; $i=1\dots n$ – независимые выборки из полиномиальных распред. с вероятн. q_{ij} и заранным числом наблюд. n_i, $\sum\limits_{j=1}^{m} q_{ij}=1$ и $\sum\limits_{j=1}^{m}\nu_{ij}=n_i$. Т.е. наблюдаем n_i незав. экспериментов

$P(\text{эксперимент завершается } j\text{-тым исходом в } i\text{-той серии}) = q_{ij}$; $\nu_{ij} = \#\{\text{экспер., завершив. } j\text{-тым исходом}\}$

(Ср. с критерием хи-квадрат, то же самое делали).

Эта схема позволяет проверить однородность нескольких одномер. распределений (т.е. проверить гипотезу о их равенстве).

Ставится эта гипотеза так: $H_I: q_{ij}=q_{\cdot j}\ \forall i$, где $q_{\cdot j}=\frac{1}{n}\sum\limits_{i=1}^{n} q_{ij}$

Схема II Предполагается, что $(\nu_{11}\dots\nu_{nm})$ имеют полином. распред. с вероятн. $(p_{11}\dots p_{nm})$ и фиксир. числом наблюдений $N=\sum\limits_{ij}\nu_{ij}$, $\sum\limits_{ij}p_{ij}=1$

Проверяем гипотезу независимости $H_{II}: p_{ij}=r_i\cdot s_j$, где $r_i=\sum\limits_{k=1}^{m}p_{ik}$, $s_j=\sum\limits_{e=1}^{n}p_{ej}$.

Почему эта схема для нас важна. Пусть есть 2 выборки $X(X_1\dots X_N)$ и $Y(Y_1\dots Y_N)$. Хотим проверить их независимость, т.е. что $F_{XY}(s,t)=F_X(s)\cdot F_Y(t)$.

Пусть есть два разбиения $\mathbb{R}$: $\{B_i\}_{i=1}^{n}$ и $\{C_j\}_{j=1}^{m}$. Обозначим $\nu_{ij}=\#\{k: X_k\in B_i \text{ и } Y_k\in C_j\}$,

$p_{ij}:=P(X_1\in B_i,\ Y_1\in C_j)$. Если выборки X и Y независимы, то

$$p_{ij}=P(X_1\in B_i)\cdot P(Y_1\in C_j)=\Big(\sum_{k=1}^{m}P(X_1\in B_i,\ Y_1\in C_k)\Big)\cdot\Big(\sum_{e=1}^{n}P(X_1\in B_e,\ Y_1\in C_j)\Big)=$$
$$=r_i\cdot s_j\ !$$

Для проверки гипотез H_I и H_{II} применяется вариант крит. хи-квадрат: определим $\chi^2=N\cdot\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\frac{\left(\nu_{ij}-\frac{n_i\cdot m_j}{N}\right)^2}{n_i\cdot m_j}$, где $n_i=\sum\limits_{k=1}^{m}\nu_{ik}$, $m_j=\sum\limits_{e=1}^{n}\nu_{ej}$

Теорема $\chi^2\xrightarrow{d}\chi^2_{(n-1)(m-1)}$ при $N\to+\infty$.

Замечание. С помощью критерия, построенного по этой теореме, действительно можно проверить гипотезу о независимости 2 выборок, в отличие от критериев, построенных с помощью выборочных коэффициентов корреляции. Но проблема в том, что в теореме для хорошей работы критерия требуется, как и ранее, $\frac{n_i\cdot m_j}{N}\geq 5\ \forall i,\forall j$, а отвергнуть гипотезу о независимости с помощью коэф. корреляции можно на гораздо меньшем числе наблюдений.

МСПС | Множественная корреляция

Измерение стат. связи между $k \geq 3$ выборками.

Опр. (ранговый коэф. конкордации Кэндалла)

$$W = \frac{12}{k^2(n^3-n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} R_{ij} - \frac{k(n+1)}{2} \right)^2$$, где R_{ij} – ранг (от 1 до n) i-того эл-та в j-й выборке.

Св-ва:

1) $0 \leq W \leq 1$, причём $W = 1 \Leftrightarrow$ все k ранжировок совпадают.

2) Обозначим через $\overline{\rho_S}$ среднее арифм. коэф. Спирмэна по всем $\frac{k(k-1)}{2}$ парам выборок. Тогда

$$W = \frac{(k-1)\overline{\rho_S} + 1}{k}.$$

3) $k(n-1)\,W \xrightarrow[n\to\infty]{d} \chi^2(n-1)$

Частные корреляции

Пример про грузы и вес. пред. брака.

Опровер. влияние неучтён. факторов на исследуемые переменные может искажать истинную связь между перем., т.е. подсчёты могут приводить к полож. знач. парных коэф. корр., а они могут быть независ. при учёте этих факторов.

Опр. Част. коэф. корреляции между X и Y при искл. влияния сл. в. Z называется

$$\rho(X,Y|Z) \equiv \rho_{XY|Z} = \frac{\rho(X,Y) - \rho(X,Z)\,\rho(Y,Z)}{\sqrt{(1-\rho^2(X,Z))\,(1-\rho^2(Y,Z))}}$$

К этой формуле приводит попытка исключить зависимость от Z, заменив X и Y сл. в. $X' = X - \alpha Z$, $Y' = Y - \beta Z$, кот. некорр. с Z, тогда "оставшаяся" корреляция есть корреляция Y' и X'.

Для получ. оценки $\hat{\rho}_{XY|Z}$ заменяем все коэф. корр. на коэф. корр. Пирсона.

Св-во: Если X, Y, Z – выборки и нулев. корр. зависимость, то $\hat{\rho}_{XY|Z}$ будет распред. так же, как и $\hat{\rho}_{XY}$, но для выборок размера $n-1$. Тогда $\sqrt{n}\,\operatorname{arcth}\hat{\rho}_{XY|Z} \xrightarrow{d} N(0,1)$

Ранг. коэф. корр. Кэндалла τ (в отл. от ρ_S) переносится на случай част. корр. с ном. аналогич. ф-ле:

$$\tau_{XY|Z} = \frac{\tau_{XY} - \tau_{XZ}\,\tau_{YZ}}{\sqrt{(1-\tau_{XZ}^2)(1-\tau_{YZ}^2)}}.$$