Методы современной прикладной статистики
5. Корреляционный анализ

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Постановка задачи

Пусть значения признаков X и Y измерены на n объектах, т.е. имеются выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$. Хотим измерить величину зависимости между признаками X и Y.

Но независимость признаков проверять чрезвычайно сложно (потому что независимость – континуальное свойство), поэтому вместо этого проверяют их некоррелируемость.

Коэффициент корреляции

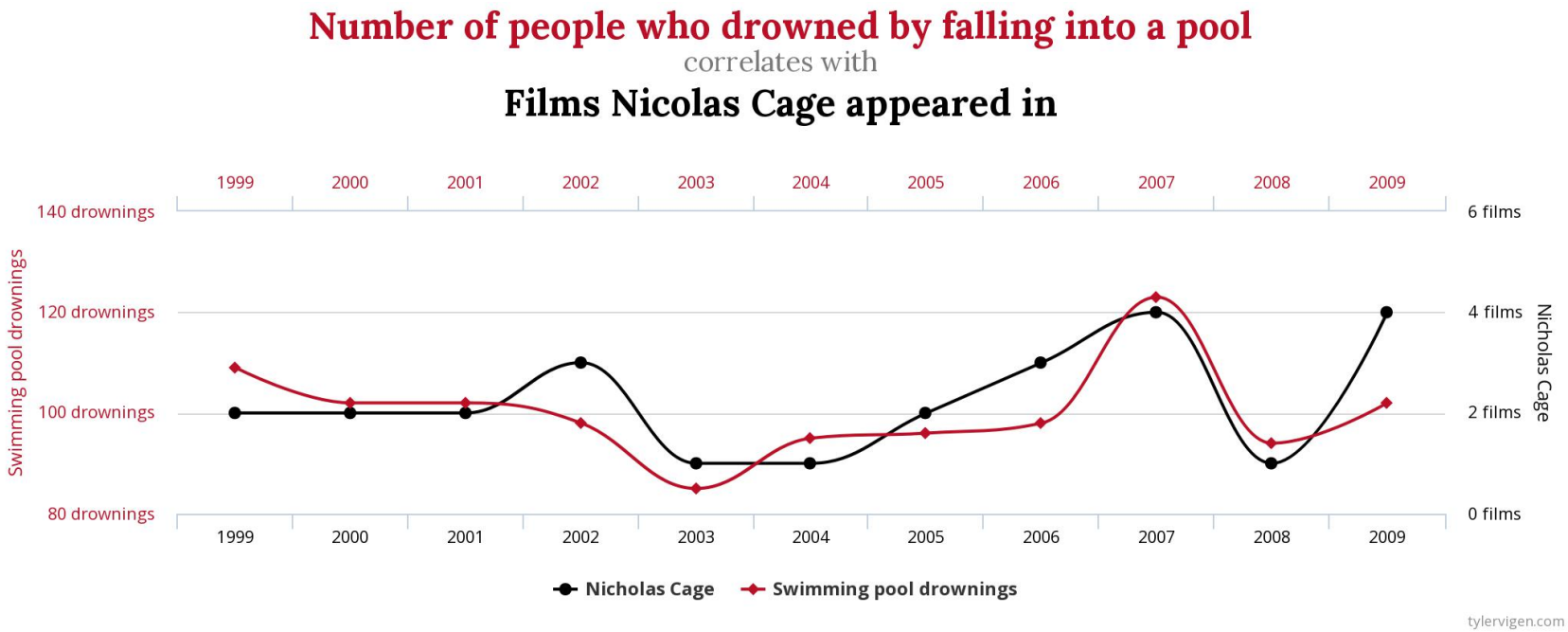
Напомним, коэффициент корреляции Пирсона случайных величин X и Y равен

$$\rho(X, Y) = \frac{\mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)}{\sqrt{\mathrm{D}X\mathrm{D}Y}}.$$

Основное свойство: если коэффициент корреляции не равен нулю, то величины зависимы, обратное неверно.
Кроме того, $\rho(X, Y) \in [-1, 1]$.

Корреляция и причинность

Корреляция – статистическая взаимосвязь между случайными величинами. Не стоит думать, что она является достаточным условием причинно-следственной связи.

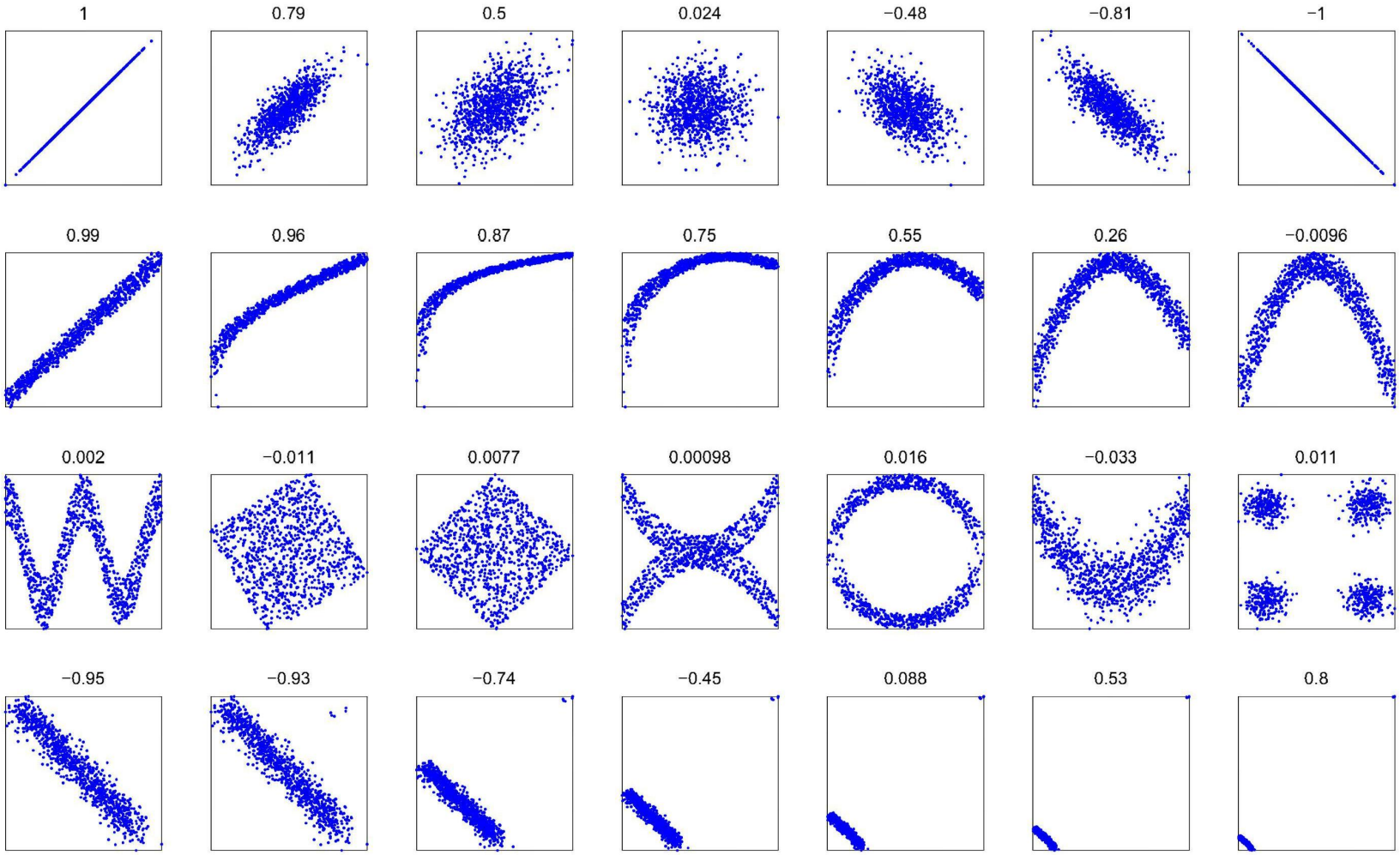


Коэффициент корреляции Пирсона

Пусть имеются две выборки одинакового размера $(X_1, \ldots, X_n)$ и $(Y_1, \ldots Y_n)$. Оценкой $\rho(X, Y)$ служит выборочный **коэффициент корреляции Пирсона**

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{j=1}^{n}(Y_j - \overline{Y})^2}}.$$

Коэффициент корреляции Пирсона



Значения коэффициента Пирсона на разных данных.

Критерий Стьюдента

Если выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots Y_n)$ – нормальные, то равенство нулю коэффициента корреляции $\rho(X_1, Y_1)$ эквивалентно независимости выборок.

При верности гипотезы $H_0 : \rho(X_1, Y_1) = 0$ выполнено $\frac{\widehat{\rho}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}} \sim St(n-2)$ (при $n > 10$). Отсюда получаем критерий проверки гипотезы H_0 против альтернативы $H_1 : \rho(X_1, Y_1) \neq 0$:

$$\textit{если } \frac{\widehat{\rho}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}} \notin (t_{\alpha/2}, t_{1-\alpha/2}), \textit{ то отвергнуть } H_0,$$

где t_γ – γ-квантиль распределения Стьюдента с $n-2$ степенями свободы.

Коэффициент корреляции Пирсона

Недостатки коэффициента корреляции Пирсона:

- Крайне чувствителен к выбросам (см. квартет Энскомба);
- Служит эффективной мерой только для линейной зависимости;
- В случае отличия выборок от нормальных распределение статистики критерия может отличаться от распределения Стьюдента.

Замечание. При использовании коэффициента Пирсона нормальное распределение данных является желательным, но не строго обязательным условием. Если данные хорошо сгруппированы, т.е. не имеют тяжелых хвостов и выбросов, то коэффициентом Пирсона можно пользоваться.

Коэффициент корреляции Спирмена

Если распределение выборки не является близким к нормальному, то лучше пользоваться ранговыми коэффициентами корреляции.

Напомним, ранг R_i наблюдения X_i – это его номер в вариационном ряду выборки $(X_1, \ldots, X_n)$, т.е. $X_i = X_{(R_i)}$.

Пусть R_i – ранг X_i в выборке $(X_1, \ldots, X_n)$, S_j – ранг Y_j в выборке $(Y_1, \ldots, Y_n)$. Тогда **коэффициент корреляции Спирмена**

$$\rho_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{j=1}^{n}(S_j - \overline{S})^2}},$$

где $\overline{R} = \overline{S} = \frac{n+1}{2}$.

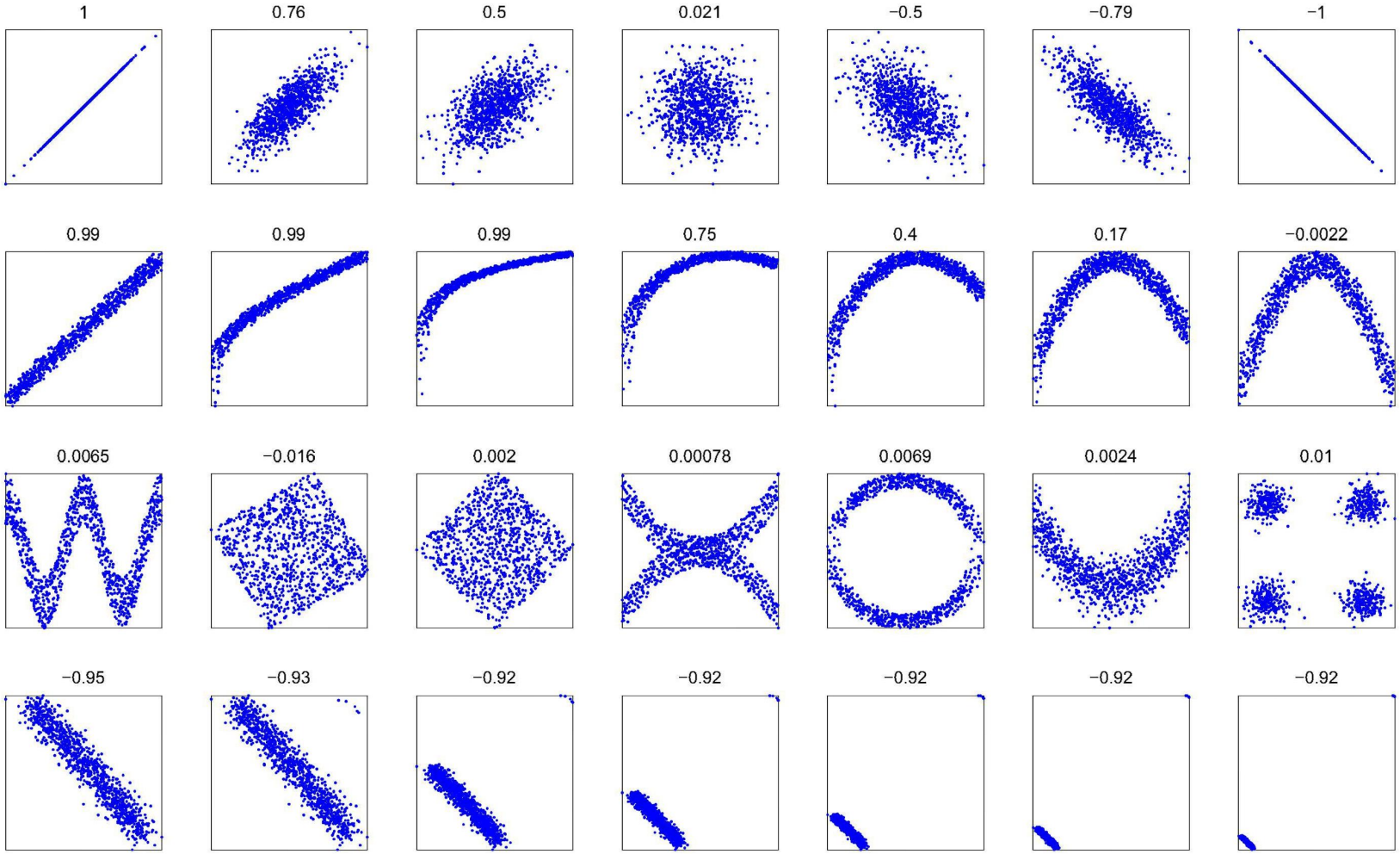
Свойства

1. Коэффициент корреляции Спирмена – мера монотонной корреляции между признаками X и Y. Он равен коэффициенту корреляции Пирсона между рангами наблюдений.
2. $\rho_S = 1 - \frac{6}{n^3-n}\sum_{i=1}^{n}(R_i - S_i)^2$.
3. При верной гипотезе H_0 о независимости признаков $E\rho_S = 0, D\rho_S = \frac{1}{n-1}$ и $\sqrt{n-1}\rho_S \approx N(0,1)$ $(n > 10)$.
4. Более точное приближение: возьмем

$$\widetilde{\rho}_S = \frac{1}{2}\rho_S\left(\sqrt{n-1} + \sqrt{\frac{n-2}{1-\rho_S^2}}\right),$$

тогда отвергаем H_0, если $\widetilde{\rho}_S \notin (z_{\alpha/2}, z_{1-\alpha/2})$, где $z_\gamma = (x_\gamma + y_\gamma)/2$, x_γ и y_γ – γ-квантили $N(0,1)$ и $St(n-2)$ соответственно.

Коэффициент корреляции Спирмена



Значения коэффициента Спирмена на разных данных.

Коэффициент корреляции Кендалла

Будем говорить, что пары (X_i, Y_i) и (X_j, Y_j) согласованы, если

$$sgn((X_i - X_j)(Y_i - Y_j)) = 1.$$

Обозначим S – число согласованных пар и R – число несогласованных, обозначим также

$$T = S - R = \sum_{i<j} sgn((X_i - X_j)(Y_i - Y_j)).$$

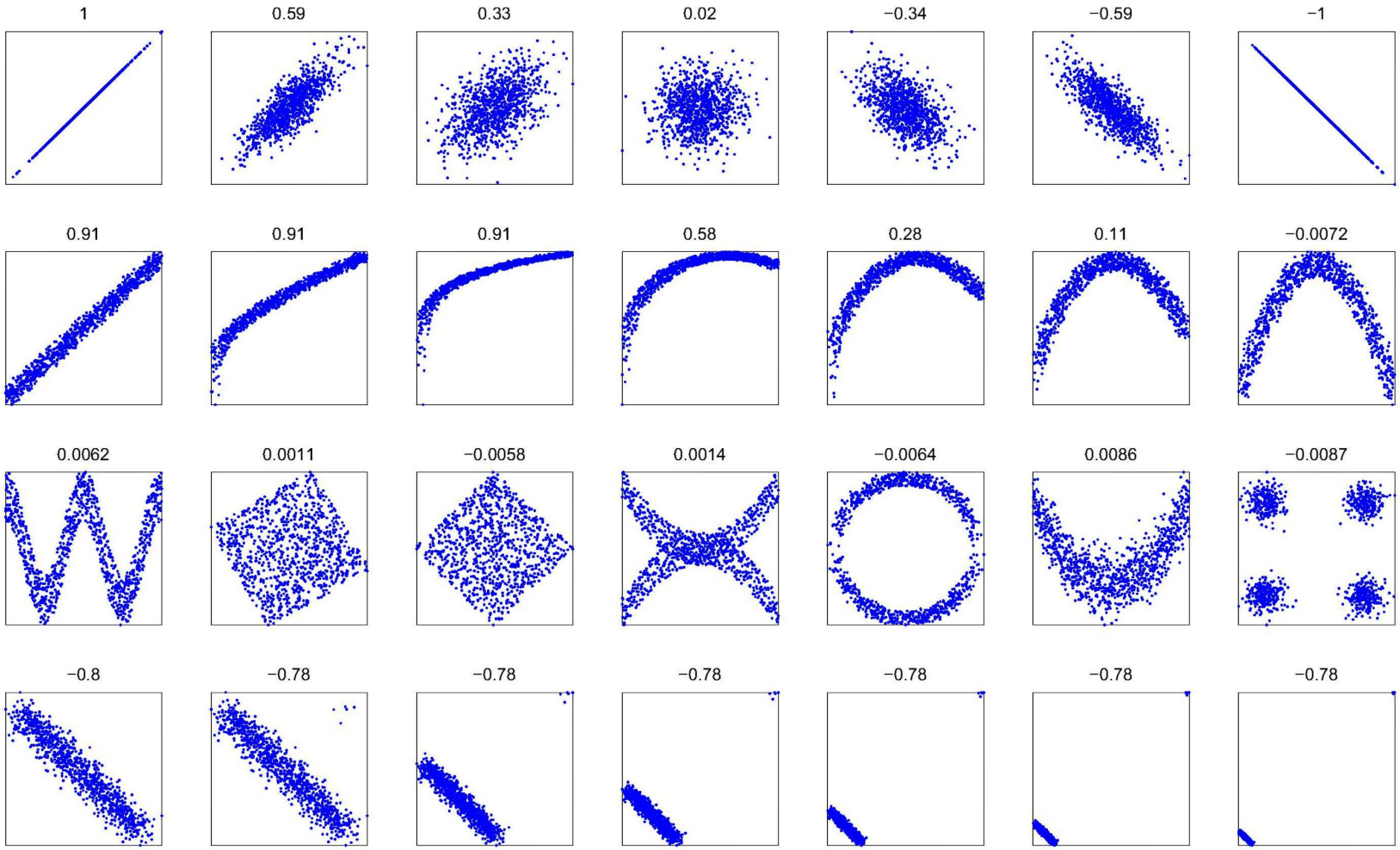
Легко видеть, что $|T| \leq \frac{n(n-1)}{2}$. **Коэффициент корреляции Кендалла** определяется как

$$\tau = \frac{2}{n(n-1)} T.$$

Свойства τ Кендалла

1. $\tau = 1 - \frac{4}{n(n-1)} R$;
2. при верной гипотезе H_0 $\frac{\tau}{D\tau} \approx N(0,1)$ (при $n > 10$), где $D\tau = \frac{2(2n+5)}{9n(n-1)}$;
3. $corr(\rho_S, \tau) > 0.99$ при $n > 4$; если их значения не близки к 1, то, как правило, $\rho_S \approx 1.5\tau$;
4. τ чаще используется, чем ρ_S, потому что его проще пересчитывать при увеличении количества наблюдений;
5. и ρ_S, и τ являются устойчивыми к выбросам.

τ Кендалла



Значения τ Кендалла на различных данных.

Частная корреляция

Бывает так, что наблюдаемая линейная (монотонная) связь между признаками X и Y вызвана влиянием третьего признака Z, и тогда можно выяснить величину зависимости X и Y без учета влияния Z с помощью частной корреляции:

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

где ρ_{XY} – коэффициент корреляции Пирсона или Спирмена признаков X и Y.

Частная корреляция

В случае, если мы хотим выяснить зависимость признаков X_i и X_j без влияния множества признаков $M \backslash \{X_i, X_j\}$, то пользуемся следующим методом:

$$\rho_{X_i X_j | M \backslash \{X_i, X_j\}} = -\frac{r_{ij}}{\sqrt{r_{ii} r_{jj}}},$$

где Σ – обратимая матрица выборочных корреляций множества признаков M, $R = \Sigma^{-1}$, $R = \|r_{ij}\|$.

Частная корреляция

Проверим гипотезу H_0 некоррелированности (независимости) признаков X_1 и X_2 без влияния множества признаков $(X_3, \ldots, X_M)$. Обозначим $\rho = \rho_{X_1 X_2 | X_3 \ldots X_M}$. Тогда статистика

$$T = \frac{\rho\sqrt{n - M}}{\sqrt{1 - \rho^2}}$$

приближаема распределением Стьюдента с $n - M$ степенями свободы.

Здесь n по-прежнему число объектов, т.е. число элементов в выборках $X_1, X_2, \ldots$ Если основой для частного коэффициента корреляции выступает коэффициент Пирсона, то требуем, чтобы $(X_1, \ldots, X_M) \sim N(\mu, A)$.

Множественная корреляция

Для того, чтобы оценить величину линейной (монотонной) связи между признаками X и (Y, Z), пользуются множественным коэффициентом корреляции

$$\rho_{XYZ} = \frac{\rho_{XY}^2 + \rho_{XZ}^2 - 2\rho_{XY}\rho_{YZ}\rho_{XZ}}{1 - \rho_{YZ}^2},$$

где, как и ранее, ρ_{XY} – коэффициент корреляции Пирсона или Спирмена признаков X и Y.
Если имеем больше 2 дополнительных признаков, то используем коэффициент

$$\rho_{X,M} = c^T R c,$$

где M – множество дополнительных признаков, Σ – матрица их выборочных корреляций, $R = \Sigma^{-1}$, а c – вектор корреляций признака X с дополнительными.

Множественная корреляция

Для проверки гипотезы H_0 о некоррелированности (независимости) признака X с множеством дополнительных признаков M используется статистика

$$F = \frac{\rho^2_{X,M}}{1 - \rho^2_{X,M}} \cdot \frac{n - m}{m - 1},$$

где m – число дополнительных признаков. Данная статистика при верности H_0 хорошо приближается распределением Фишера с $m - 1$ и $n - m$ степенями свободы соответственно.

Коэффициент конкордации

Другим методом оценки зависимости нескольких выборок являются коэффициенты конкордации. В отличие от коэффициентов множественной корреляции, в данном методе оценивается не корреляция одного признака от остальных, а общая корреляция признаков между собой.

Определим ранговый коэффициент конкордации Кендалла как

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

где k – количество выборок, n – число наблюдений в них, R_{ij} – ранг i-того наблюдения в j-той выборке.

Свойства

1. Как и коэффициенты множественой корреляции, $W \in [0, 1]$.
2. Пусть $\overline{\rho}_S$ – среднее арифметическое коэффициентов Спирмена по всем $\frac{k(k-1)}{2}$ парам выборок, тогда $W = \frac{k-1}{k}\overline{\rho}_S + \frac{1}{k}$.
3. При $n > 10$ и верной гипотезе о некоррелированности выборок распределение $k(n-1)W$ приближается хи-квадрат распределением с $n-1$ степенью свободы.

Таблицы сопряженности

Обсудим задачу выявления статистической зависимости между категориальными признаками. Пусть признак X принимает m значений, признак Y – k значений.

Образуем таблицу $\|\nu_{ij}\|$, где ν_{ij} – количество случаев, когда признак X принял i-тое значение, а признак Y – j-тое. Такие таблицы называются таблицами сопряженности.

Таблицы сопряженности

Для проверки гипотезы H_0 независимости признаков X и Y используется вариант критерия хи-квадрат. Определим $m_i = \sum_{j=1}^{k} \nu_{ij}$, $k_j = \sum_{i=1}^{m} \nu_{ij}$, а $n = \sum_{ij} \nu_{ij}$ – общее количество объектов. Рассмотрим статистику

$$\chi = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(\nu_{ij} - \frac{m_i k_j}{n}\right)^2}{m_i k_j}.$$

При верной H_0 и при $\frac{m_i k_j}{n} \geq 5$ для всех i, j распределение статистики χ приближается распределением хи-квадрат с $(m-1)(k-1)$ степенями свободы.

Замечания

С помощью предложенного критерия можно проверить гипотезу о независимости 2 произвольных выборок $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$. Действительно, выберем 2 разбиения $\mathbb{R}$ $\{B_i\}_{i=1}^m$ и $\{C_j\}_{j=1}^k$ и положим $\nu_{ij} = \sharp\{u : X_u \in B_i, Y_u \in C_j\}$.

Также с помощью критерия можно проверить гипотезу о однородности k выборок $(X_1^1, \ldots, X_{n_1}^1), \ldots, (X_1^k, \ldots, X_{n_k}^k)$ (т.е. гипотезу о том, что все k выборок одинаково распределены). Для этого рассмотрим $\{B_i\}_{i=1}^m$ – разбиение $\mathbb{R}$. Положим $\nu_{ij} = \sharp\{u : X_u^j \in B_i\}$, тогда малость статистики χ будет означать, что распределение членов выборок по множествам B_i не зависит он номера выборки, что и говорит об одинаковой распределенности данных выборок.

Таблицы сопряженности 2×2

Отдельный случай таблиц сопряженности – таблицы 2×2, т.е. когда оба признака являются бинарными.

	$X = 0$	$X = 1$	Σ
$Y = 0$	a	b	$a + b$
$Y = 1$	c	d	$c + d$
Σ	$a + c$	$b + d$	n

Как и ранее, хотим проверить гипотезу о независимости признаков X и Y.

Коэффициент контингенции

Рассмотрим статистику

$$V = \frac{n\left(|ad - bc| - n/2\right)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Если $V > u_{1-\alpha}$, где $u_{1-\alpha}$ – квантиль ξ^2-распределения с 1 степенью свободы, то отвергаем гипотезу о независимости признаков X и Y.

Точный критерий Фишера

Критерием контингенции можно пользоваться в случае, если $n \geq 40$ и $a, b, c, d \geq 5$. В остальных случае следует применять точный критерий Фишера.

Пусть количество объектов по строкам и столбцам таблицы фиксировано. Тогда вероятность получить ровно такую таблицу 2×2 равна

$$P(X, Y) = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_{n}^{a+c}} = \frac{(a+b)!(a+c)!(b+d)!(c+d)!}{n!a!b!c!d!}.$$

Точный критерий Фишера

P-значение критерия определяется как сумма по всем возможным вариантам таблицы с такими же суммами по строкам и столбцам, имеющим вероятность не более чем $P(X, Y)$. В частности, если $ad < bc$, то p-значение ищется по формуле

$$p = \sum_{i=0}^{a} \frac{C_{a+b}^{i} C_{c+d}^{a+c-i}}{C_{n}^{a+c}}.$$

Парадокс хи-квадрат (Симпсона)

Эксперимент: пациенты принимают препарат или плацебо, и по окончании курса определяется, выздоровели они или нет. Есть ли связь между выздоровлением и приемом препарата?

Мужчины	Выздор.	Нет
Препарат	700	800
Плацебо	80	130

Женщины	Выздор.	Нет
Препарат	150	70
Плацебо	300	280

Для мужчин: $V = 5.45$, $p = 0.0195$.
Для женщин: $V = 17.55$, $p = 2.8 \cdot 10^{-5}$.

М+Ж	Выздор.	Нет
Препарат	850	870
Плацебо	380	410

Суммарно: $V = 0.37$, $p = 0.53$.

Парадокс хи-квадрат (Симпсона)

Причины несогласованности выводов – большие отличия в размерах групп пациентов, принимающих плацебо и препарат: основной вклад в выводы вносят женщины, принимавшие плацебо, и мужчины, принимавшие препарат.

Чтобы такого не происходило, плацебо и препарат должны (примерно) поровну распределяться по всем анализируемым подгруппам.

Finita!