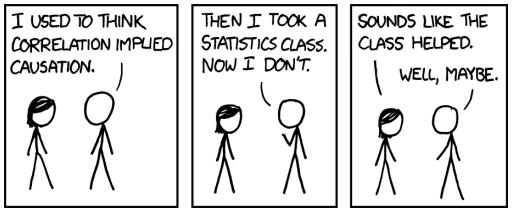

Корреляционный анализ

Чертоги разума

Задача: даны две *связанные* выборки $(X_1, Y_1), \dots, (X_n, Y_n)$. Требуется проверить гипотезу об их независимости $H_0\colon X \perp\!\!\!\perp Y$.
В непрерывном случае это тяжко, обычно проверяют их некоррелированность посредством статистик, называемых *коэффициентами корреляции*.

- Коэффициент Пирсона a.k.a выборочная корреляция выборок:

$$\widehat{\rho}(\mathbf{X}, \mathbf{Y}) = \frac{\overline{\mathbf{XY}} - \overline{\mathbf{X}} \cdot \overline{\mathbf{Y}}}{\sqrt{s^2(\mathbf{X}) \cdot s^2(\mathbf{Y})}} \xrightarrow{\mathsf{P}} \rho = \mathsf{corr}(X_1, Y_1).$$

  С помощью дельта-метода доказывается, что при верной H_0

$$\sqrt{n} \cdot \widehat{\rho}(\mathbf{X}, \mathbf{Y}) \xrightarrow{d} \mathcal{N}(0, 1).$$

  Для нормальных выборок распределение $\widehat{\rho}$ можно уточнить (см. задачу 5).

- Коэффициент Спирмена a.k.a выборочная корреляция *рангов*, то есть таких R_i и S_i, что $X_{(R_i)} = X_i$ и $Y_{(S_i)} = Y_i$:

$$\rho_S(\mathbf{X}, \mathbf{Y}) = \frac{\overline{\mathbf{RS}} - \overline{\mathbf{R}} \cdot \overline{\mathbf{S}}}{\sqrt{s^2(\mathbf{R}) \cdot s^2(\mathbf{S})}}.$$

- Коэффициент Кенделла — нормированная разность согласованных и несогласованных пар:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sign}(X_i - X_j) \cdot \operatorname{sign}(Y_i - Y_j).$$

Последние два коэффициента асимпт. нормальны в смысле, что

$$\frac{\rho_S}{\sqrt{\mathsf{D}\rho_S}}, \frac{\tau}{\sqrt{\mathsf{D}\tau}} \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{где } \mathsf{D}\rho_S = \frac{1}{n-1}, \ \mathsf{D}\tau = \frac{2(2n+5)}{9n(n-1)}$$

Для дискретного случая, когда $X_i \in \{1, \dots, k\}$, $Y_i \in \{1, \dots, m\}$, применяют *критерий независимости хи-квадрат*, основанный на статистике:

$$\chi^2(\mathbf{X}, \mathbf{Y}) = n \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(\nu_{ij} - \nu_{i\bullet}\nu_{\bullet j}/n)^2}{\nu_{i\bullet}\nu_{\bullet j}} \xrightarrow{d} \chi^2_{(k-1)(m-1)}, \quad n \to \infty$$

где $\nu_{ij} = \sum_{l=1}^{n} I(X_l = i, Y_l = j)$, $\nu_{i\bullet} = \sum_{j=1}^{m} \nu_{ij}$, $\nu_{\bullet j} = \sum_{i=1}^{k} \nu_{ij}$.

Feladatok

1. Выведете критерий независимости хи-квадрат как КОП, используя теорему Уилкса.

2. Из 360 поступающих на DS-поток юношей было принято 97, а из 82 девушек — 40. Проверить гипотезу о независимости факта поступления от пола с $\alpha = 0.01$.

3. Группа Б05-024 перед написанием контрольной по статистике решила для храбрости постоять немного в планке. Ниже приведены результаты: в первой строке записано количество минут, которое человек простоял в планке, во второй — балл за контрольную.

0.5	2.5	8.5	2	6	10	0.5	1	0.5	3	6.5	0.5	2	3.5	6.0	1.5	1
3	0	2	2	2	3	0.5	2	1	2.5	3	3	2	2	3	0	3

Посчитайте коэффициент корреляции Спирмена (можно использовать Python) и проверьте с помощью него гипотезу о том, что данные метрики независимы.

4. Докажите, что в случае, когда (X_i, Y_i) — гауссовский вектор (необязательно $X_i \perp\!\!\!\perp Y_i$):
(а) имеет место сходимость

$$\sqrt{n} \cdot (\widehat{\rho} - \rho) \xrightarrow{d} \mathcal{N}(0, (1 - \rho^2)^2);$$

(б) коэффициент Кенделла несмещённо оценивает $\frac{2}{\pi} \arcsin \rho$;
(в) коэффициент Спирмена несмещённо оценивает нечто, стремящееся к $\frac{6}{\pi} \arcsin \frac{\rho}{2}$.

5. Пусть $\widehat{\rho}$ — коэффициент корреляции Пирсона, построенный по выборкам $\mathbf{X}$ и $\mathbf{Y}$. Докажите, что есть эти выборки независимы и нормально распределены, то

$$\widehat{\rho} \cdot \sqrt{\frac{n-2}{1-\widehat{\rho}^2}} \sim T_{n-2}.$$

Указание. Найдите условное распределение $\widehat{\rho}$ при фиксированной $\mathbf{X}$ и покажите, что оно одинаково при любом условии.