Шпаргалка

Коэффициент Пирсона pearsonr в scipy.stats
Коэффициент Спирмена spearmanr в scipy.stats
Коэффициент Кенделла kendalltau и weightedtau в scipy.stats
Хи-квадрат тест scipy.stats.chisquare (на вход подаются эмпирические и теоретические частоты)
Подход HHG hyppo.independence.HHG
Подход DCorr hyppo.independence.Dcorr
Подход HSIC hyppo.independence.Hsic
Подход MGC hyppo.independence.MGC

Вариант 1

1. Исследуйте зависимость выборок X и $Y = f(X) + U$ с помощью коэффициентов Пирсона, Спирмена, Кенделла, MGC. Шум U возьмите независимыми $\mathcal{N}(0, 2)$ величинами. Рассмотрите шесть случаев: $X \sim R[0, 1]$ или $X \sim R[-10, 10]$, а $f(x) = 2x + 5$ или $f(x) = ||x| - 1|$ или $f(x) = \sin x$. Попробуйте объемы выборок 100 и 50. Почему коэффициенты отслеживают (или нет) данные виды зависимости?

2. Устройте равномерное распределение на кольце $1/2 < X^2 + Y^2 < 1$ (возьмите выборки объема 100 и 50). Исследуйте зависимость между X и Y с помощью коэффициента Пирсона, подхода MGC и подхода DCorr. Теперь то же самое для равномерного распределения на квадрате $-1 \le X \le 1$, $-1 \le Y \le 1$. Теперь то же самое на множестве, состоящем из четырех кругов с центрами в $(\pm 1, \pm 1)$ и радиуса $1/2$.

3. Сделайте выборку (x_i, y_i, z_i), $1 \le i \le 100$, из трех независимых $\mathcal{N}(0, 1)$ распределений. Теперь посчитайте $(u_i, v_i, w_i) = (x_i, y_i, z_i) \begin{pmatrix} 1 & 0.1 & -0.1 \\ 1 & -0.7 & 0.2 \\ 1 & 0.2 & 0.8 \end{pmatrix}$. Теперь для данных (X, Y, Z, U, V, W) найдите матрицу парных корреляций. Можете использовать тепловую диаграмму. Найдите те частные корреляции, которые считаете нужными найти. Почему взяли именно эти? Какой вывод можете сделать?

4. Файл babies1.txt содержит информацию о ребенке и его матери, babiesdesc содержит описание переменных. Исследуйте зависимость переменных (включая исключенные корреляции) и сделайте выводы о том, как параметры влияют друг на друга.

5. Файл televisions.dat.txt содержит информацию следующего характера:
Country, Life expectancy, People per television, People per physician, Female life expectancy, Male life expectancy.
Исследуйте зависимость продолжительности жизни для каждого из полов от других параметров а) напрямую б) используя исключенные (скорректированные) корреляции.

Вариант 2

1. Исследуйте зависимость выборок X и $Y = f(X)+U$ с помощью коэффициентов Пирсона, Спирмена, Кенделла, MGC. Шум U возьмите независимыми $R[-2, 2]$ величинами. Рассмотрите шесть случаев: $X \sim R[0, 1]$ или $X \sim R[-10, 10]$, а $f(x) = 2x + 5$ или $f(x) = x^2$ или $f(x) = x \cos x$.

2. Устройте равномерное распределение на кольце $1/2 < X^2 + Y^2 < 1$ (возьмите выборки объема 100 и 50). Исследуйте зависимость между X и Y с помощью коэффициента Кенделла, подхода MGC и подхода HHG. Теперь то же самое для равномерного распределения на квадрате $-1 \le X \le 1$, $-1 \le Y \le 1$. Теперь то же самое на множестве, состоящем из четырех кругов с центрами в $(\pm 1, \pm 1)$ и радиуса $1/2$.

3. Сделайте выборку (x_i, y_i, z_i), $1 \le i \le 100$, из трех независимых exp(1) распределений. Теперь посчитайте $(u_i, v_i, w_i) = (x_i, y_i, z_i) \begin{pmatrix} 2 & 0.5 & 1 \\ 0.4 & -1 & 0.4 \\ 0 & 0.5 & 1 \end{pmatrix}$. Теперь для данных (X, Y, Z, U, V, W) найдите матрицу парных корреляций. Можете использовать тепловую диаграмму. Найдите те частные корреляции, которые считаете нужными найти. Почему взяли именно эти? Какой вывод можете сделать?

4. Исследуйте зависимость показателей x_1, x_2 в файле babies.txt. Есть ли между ними зависимость? Можем ли мы сделать вывод, что большой показатель x_1 повышает шансы ребенка вырасти с большим показателем x_2?

5. В файле Priem.csv приведены данные о сдаче экзаменов при поступлении на мехмат. Возьмите метод, наиболее подходящий для использования, учитывая возможные совпадения данных, найдите парные и частные (скорректированные) корреляции. Что можно сказать о зависимости?

Вариант 3

1. Исследуйте зависимость выборок X и $Y = f(X) + U$ с помощью коэффициентов Пирсона, Спирмена, Кенделла, MGC. Шум U возьмите независимыми $Laplace(1)$ величинами. Рассмотрите шесть случаев: $X \sim R[0,1]$ или $X \sim R[-10,10]$, а $f(x) = 2x + 5$ или $f(x) = \ln|x|$ или $f(x) = \sin x$. Почему коэффициенты отслеживают (или нет) данные виды зависимости?

2. Устройте равномерное распределение на кольце $1/2 < X^2 + Y^2 < 1$ (возьмите выборки объема 100 и 50). Исследуйте зависимость между X и Y с помощью коэффициента Пирсона, подхода MGC и подхода HSIC. Теперь то же самое для равномерного распределения на квадрате $-1 \le X \le 1$, $-1 \le Y \le 1$. Теперь то же самое на множестве, состоящем из четырех кругов с центрами в $(\pm 1, \pm 1)$ и радиуса $1/2$.

3. Сделайте выборку (x_i, y_i, z_i), $1 \le i \le 100$, из трех независимых $R[-1,1]$ распределений. Теперь посчитайте $(u_i, v_i, w_i) = (x_i, y_i, z_i) \begin{pmatrix} 0.5 & 0.1 & 1.5 \\ -1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.1 \end{pmatrix}$. Теперь для данных (X, Y, Z, U, V, W) найдите матрицу парных корреляций. Можете использовать тепловую диаграмму. Найдите те частные корреляции, которые считаете нужными найти. Почему взяли именно эти? Какой вывод можете сделать?

4. Файл carbon-footprint-2015.txt содержатся данные об углеродном следе различных автомобилей на различных трассах. Какие параметры наиболее тесно связаны с углеродным следом (Carbon Footprint).

5. Файл body.dat.txt содержит информацию об измерениях нескольких параметров человека. Определите, какие из них наиболее сильно зависят друг от друга, учитывая исключенные (скорректированные) корреляции. Описание дано в файле body.txt.