4 Проверка независимости

4.1 Теория

Гипотеза независимости проверяется для двумерных выборок $(X_i, Y_i) \sim H$ и имеет вид

$$\mathrm{H}_0 : H(x, y) = F(x)G(y),$$

где $F(x)$, $G(y)$ – маргинальные распределения. Начнем с общей альтернативы $\mathrm{H}_1 : H(x, y) \neq F(x)G(y)$ для некоторых x, y.

4.1.1 Общая альтернатива

Прежде всего, заметим, что многие из наших прежних подходов остаются действенными.

1. Критерий хи-квадрат предлагает рассматривать статистику

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{1}{\frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n}} \left(\nu_{i,j} - \frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n}\right)^2, \quad \nu_{\cdot,j} = \sum_{i=1}^{m} \nu_{i,j}, \quad \nu_{i,\cdot} = \sum_{j=1}^{k} \nu_{\cdot,j}.$$

   Эту величину предлагается сравнивать с квантилью $\chi^2_{(k-1)(m-1)}$ распределения. Как мы видим, этот критерий тот же самый, что и критерий однородности для той же таблицы сопряженности. Аналогично дело обстоит с критерием отношения правдоподобия.

   Как обычно, мы дискретизируем значения переменных X, Y если они не дискретны, считаем количество попаданий в соответствующие ячейки по паре переменных и применяем к ним критерий хи-квадрат.

   Реализация используется все та же, что и ранее.

2. Критерий Смирнова можно адаптировать для гипотезы независимости со статистикой в форме

$$D_n = \sqrt{n} \sup_{x,y} |\widehat{H}_n(x, y) - \widehat{F}_n(x)\widehat{G}_n(y)|.$$

   При верной гипотезе $\mathrm{H}_0 : H(x, y) = F(x)G(y)$ данная статистика имеет некоторое распределение, которое не зависит от F и G. Таким образом, уровень значимости критерия можно определять методом Монте-Карло.

   В Python его, видимо, нет, реализуйте его методом Монте-Карло.

3. Более эффективным оказывается подход Секея и Риццо, предложенный уже в 21 веке. Их статистику можно представить в виде

$$D_n = n \int_{\mathbb{R}^2} |\widehat{\psi}_H(s, t) - \widehat{\psi}_F(s)\widehat{\psi}_G(t)|^2 \omega(s, t) ds dt,$$

   где $\widehat{\psi}_H$ – выборочная х.ф. вектора (X_i, Y_i), $\widehat{\psi}_F$, $\widehat{\psi}_G$ – выборочные х.ф. отдельных выборок, ω – некоторая весовая функция. Увы, распределение статистики зависит от H даже при верной гипотезе и определяется перестановочным методом. Данный критерий есть здесь.

4.1.2 Частная альтернатива

Рассмотрим также некоторые критерии, которые используют для более узкого спектра альтернатив. Описать конкретную альтернативу здесь не так просто, поэтому скажем условно – альтернатива "Если X большой, то и Y в среднем тоже".

1. Критерий Пирсона предлагает смотреть на коэффициент корреляции

$$\rho_P(X,Y) = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{S_X S_Y}.$$

При гипотезе независимости $\sqrt{n}\rho_P$ стремится к величине $Z \sim \mathcal{N}(0,1)$. Реализация есть здесь.

2. Критерий Спирмена предлагает считать тот же коэффициент для рангов (R_i, S_i), где R_i – ранг X_i среди $X_1, \ldots, X_n$, T_i – ранг Y_i среди $Y_1, \ldots, Y_n$:

$$\rho_S(X,Y) = \frac{\overline{RT} - \overline{R}\,\overline{T}}{n^{-1}\sqrt{\sum_{i=1}^n (R_i - \overline{R})^2 \sum_{i=1}^n (T_i - \overline{T})^2}}.$$

При гипотезе независимости $\sqrt{n-1}\rho_S$ стремится к величине $Z \sim \mathcal{N}(0,1)$. Критерий реализован здесь.

3. Коэффициент Кендалла предлагает рассматривать среди всех пар (X_i, Y_i), (X_j, Y_j) пары пар, для которых $X_i \le X_j$, $Y_i \le Y_j$ или $X_i > X_j, Y_i > Y_j$. Такие пары назовем *согласованными*. Пусть число согласованных пар N, а несогласованных – $M = C_n^2 - N$. Тогда

$$\rho_K = \frac{M - N}{M + N}.$$

При этом ρ_K при гипотезе имеет распределение, не зависящее от F, G, причем

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}}\rho_K \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Конечно, можно упростить коэффициент до $9n/4$, но утверждается, что так точность аппроксимации выше. Критерий реализован здесь.

4.1.3 Коэффициенты корреляции

Отметим, что коэффициент корреляции зачастую используют не только для проверки гипотезы, но и для характеристики зависимости. Так для набора многомерных данных (записывая их по строкам) строят таблицу коэффициентов корреляции столбцов, из чего представляют, какие признаки связаны, а какие нет.

В некотором смысле, коэффициенты корреляции задают геометрию пространства случайных величин. Там обычный коэффициент корреляции можно описать следующим образом:

- Рассматриваем пространство $L^2(P)$ случайных величин со скалярным произведением EXY.
- Проецируем X и Y на ортогональное дополнение к пространству константа, получаем $\widetilde{X} = X - EX$, $\widetilde{Y} = Y - EY$.
- Считаем косинус угла между полученными величинами:

$$\frac{\mathbf{E}\widetilde{X}\widetilde{Y}}{\sqrt{\mathbf{E}\widetilde{X}^2}\sqrt{\mathbf{E}\widetilde{Y}^2}}.$$

Подобную интерпретацию можно дать и двум другим коэффициентам.

Зачастую возникает известная проблема зависимости через третье – может оказаться, что $corr(X,Y)$ большая, но просто потому, что обе переменные сильно коррелируют с некоторой Z. Для снижения этого фактора используют исключенные корреляции. Данная процедура работает так.

- Рассматриваем пространство $L^2(P)$ случайных величин со скалярным произведением EXY.

- Проецируем $\widetilde{X}$ и $\widetilde{Y}$ на ортогональное дополнение к пространству $\{cZ\}$, получаем $\widehat{X}$, $\widehat{Y}$.
- Считаем косинус угла между полученными величинами:

$$\frac{\mathbf{E}\widehat{X}\widehat{Y}}{\sqrt{\mathbf{E}\widehat{X}^2}\sqrt{\mathbf{E}\widehat{Y}^2}} = \frac{\rho_{X,Y} - \rho(X,Z)\rho(Y,Z)}{\sqrt{(1-\rho(X,Z)^2)(1-\rho(Y,Z)^2)}}.$$

Такой коэффициент называется частным коэффициентом корреляции X, Y при условии Z: $\rho(X,Y|Z)$. Если я хочу посчитать $\rho(X,Y|Z,W)$, то я провожу ту же процедуру и получаю

$$\rho(X,Y|Z,W) = \frac{\rho(X,Y|W) - \rho(X,Z|W)\rho(Y,Z|W)}{\sqrt{(1-\rho(X,Z|W)^2)(1-\rho(Y,Z|W)^2)}}.$$

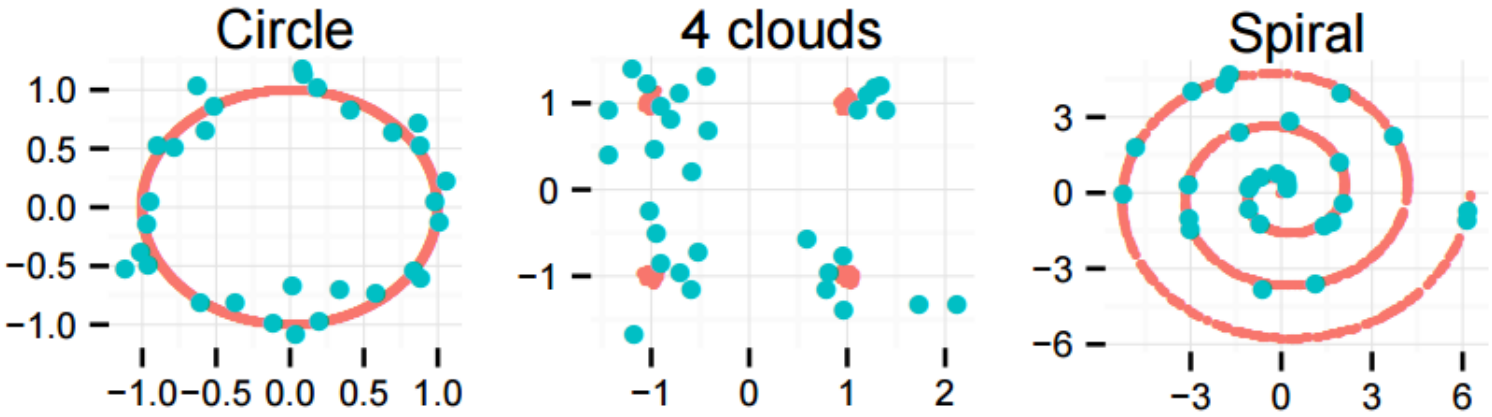
Эта процедура одна и та же для всех трех видов коэффициентов. Частные корреляции реализованы в пакете по ссылке.

4.2 Задачи

1. Найти коэффициенты корреляции баллов ЕГЭ и частные коэффициенты корреляции и сделать выводы о структуре их зависимости.

2. Для распределений, изображенных на рисунке 1 (сгенерируйте выборки самостоятельно, размер выборок 50 или 100), сравните критерии Секея-Риццо, Кендалла и хи-квадрат. Для распределения хи-квадрат ячейки предлагается выбирать, деля данные по каждой из строк на равные фрагменты.

Рис. 1: Три массива для задачи 2



3. Сравните критерии Смирнова, Пирсона, Кендалла, Спирмена и Секея-Риццо на выборках 1) $Y_i = X_i^2 + \varepsilon_i$, $X_i \sim R[-1,2]$, $\varepsilon_i \sim \mathcal{N}(0,0.5)$, 2) $Y_i = \sin X_i + \varepsilon_i$, $X_i \sim R[0,2\pi]$, $\varepsilon_i \sim \mathcal{N}(0,0.5)$.