Некоторые ранговые критерии

Понятие ранга

Пусть $X_1, \ldots, X_n$ — выборка из распределения с.в. $\xi \in \mathbb{R}$.
$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ — вариационный ряд.

Определение 1.1

Рангом R_i наблюдения X_i называется номер этого наблюдения в вариационном ряду.

Пример 1.1

Например, пусть есть выборка:
X_i: 3, 0, -1, 2, 1 *Соответствующие ранги:*
R_i: 5, 2, 1, 4, 3

В системе $\mathbb{R}$ ранги наблюдений выборки $\mathbf{X} = (X_1, \ldots, X_n)$ находятся обращением к функции $rank(\mathbf{X})$

Определение 1.2

Пусть $R_1, \ldots, R_n$ – ранги элементов выборки $X_1, \ldots, X_n$. Статистика, являющаяся функцией рангов

$$\phi_n = \phi(R_1, \ldots, R_n)$$

называется ранговой статистикой.

Определение 1.3

Критерии, основанные на ранговых статистиках, называют ранговыми критериями

Основные преимущества ранговых критериев в том, что:

1. их распределения не зависят от распределений исходных наблюдений (distribution free criterions);
2. Их применение, как правило, не предполагает конечности моментов у наблюдаемых величин.

Ранговый критерий независимости Спирмена

Имеется n наблюдений $(X_1, Y_1), \ldots, (X_n, Y_n)$ двумерной случайной величины (ξ_1, ξ_2).
Гипотеза H_0 состоит в том, что величины ξ_1 и ξ_2 — независимые.
Обозначим через R_i ранги наблюдений X_i в выборке $X_1, \ldots, X_n$, а через S_i — ранги наблюдений Y_i в выборке $Y_1, \ldots, Y_n$.
Если наблюдения X_i не зависят от Y_i, то и ранги этих наблюдений не зависят друг от друга. Критерий Спирмена основан на выборочном коэффициенте корреляции для выборки $(R_1, S_1), \ldots, (R_n, S_n)$ рангов:

$$\rho_S = \frac{\sum_{i=1}^{n} \left[(R_i - \overline{R})(S_i - \overline{S})\right]}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}}, \tag{1}$$

где

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}.$$

Очевидно, что также $\overline{S} = \frac{n+1}{2}$

Кроме того, мы аналогично имеем

$$\sum_{i=1}^{n}(R_i - \overline{R})^2 = \sum_{i=1}^{n}(S_i - \overline{S})^2 = \frac{n(n^2-1)}{12}.$$

В результате подстановки в формулу (1) приходим к выражению:

$$\rho_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - S_i)^2. \qquad (2)$$

Поскольку ρ_S — это выборочный коэффициент корреляции, то $-1 \leq \rho_S \leq 1$.

Заметим, что если для всех i $R_i = S_i$ (т.е., полное совпадение рангов, означающее положительную монотонную зависимость), то $\rho_S = 1$.

Если же $R_i = n - S_i + 1$ (полная противоположность рангов, то есть монотонная отрицательная зависимость), то $\rho_S = -1$.

При справедливости гипотезы о независимости мы имеем

$$\mathbf{E}\rho_S = 0, \qquad \mathbf{D}\rho_S \cong \frac{1}{n-1}$$

Распределение статистики Спирмена не зависит от распределений с.в. ξ_1 и ξ_2, его квантили табулированы для $n \leq 20$ и используются для проверки гипотезы независимости.
При $n > 20$ проверку гипотезы основывают на факте асимптотической нормальности распределения ρ_S.

Теорема 1.1

Если гипотеза независимости H_0 верна, то

$$\mathbf{P}(\sqrt{n}\,\rho_S < z) \longrightarrow_{n\to\infty} \Phi(z)$$

Из теоремы вытекает следующий алгоритм проверки гипотезы H_0:

1. назначается уровень значимости α;
2. находится квантиль стандартного нормального закона:
$z_\alpha = \Phi^{-1}(1-\alpha/2)$;
3. вычисляются ранги и значение статистики Спирмена ρ_S;
4. если $\sqrt{n}\,|\rho_S| > z_\alpha$, то гипотеза H_0 отвергается.

Ранговый критерий однородности Уилкоксона

Имеются две выборки вещественнозначных наблюдений:

$X_1, \ldots, X_n \sim F_1(x)$,
$Y_1, \ldots, Y_m \sim F_2(x)$

Требуется проверить гипотезу однородности, т.е. то, что эти данные из одного и того же распределения:

$$H_0 : \quad F_1(x) \equiv F_2(x)$$

Составим объединенную выборку $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, затем упорядочим значения:

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n+m)}$$

Пусть $R_1, R_2, \ldots, R_n$ обозначают ранги величин $X_1, \ldots, X_n$ в объединенном вариационном ряду.

Рассмотрим статистику

$$T_{n,m} = R_1 + R_2 + \cdots + R_n,$$

равную сумме номеров мест, которые занимают элементы первой выборки в объединенном вариационном ряду. Можно показать, что если гипотеза однородности H_0 верна, то

$$\mathbf{E}\,T_{n,m} = \frac{n}{2}(n+m+1), \qquad \mathbf{D}\,T_{n,m} = \frac{m\,n\,(m+n+1)}{12}.$$

Распределение статистики Уилкоксона (Манна–Уитни) не зависит от распределений F_1 и F_2 и табулировано при $m+n \le 50$, $\min(m,n) \ge 5$. При бóльших значениях объемов выборок используют предельное распределение. А именно: рассмотрим нормированную статистику:

$$T^0_{n,m} = \frac{T_{n,m} - \mathbf{E}\,T_{n,m}}{\sqrt{\mathbf{D}\,T_{n,m}}} = \frac{T_{n,m} - \frac{n}{2}(n+m+1)}{\sqrt{\frac{m\,n}{12}(m+n+1)}}.$$

Асимптотический вариант критерия Уилкоксона–Манна–Уитни основан на следующем факте.

Теорема 1.2

Если гипотеза однородности H_0 верна, то

$$\mathbf{P}(T^0_{n,m} < z) \longrightarrow_{m,\, n\to\infty} \Phi(z)$$

Из теоремы вытекает следующий алгоритм проверки гипотезы H_0:

1. назначается уровень значимости α;
2. находится квантиль стандартного нормального закона: $z_\alpha = \Phi^{-1}(1-\alpha/2)$;
3. вычисляются ранги R_i в объединенной выборке и значение статистики $T_{n,m}$;
4. если $\left|T^0_{n,m}\right| = \left|\frac{T_{n,m}-\frac{n}{2}(n+m+1)}{\sqrt{\frac{m\,n}{12}(m+n+1)}}\right| > z_\alpha$, то гипотеза однородности H_0 отвергается.

Задание

1. Смоделировать выборку $(X_1, Y_1), \ldots, (X_n, Y_n)$ «наблюдений » независимых X и Y. Проверить независимость по критерию Спирмена. Предусмотреть зависимость (например, сделать часть элементов выборки Y-ков функциями от соответствующих X). Проверить гипотезу о независимости.

2. Смоделировать две выборки $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$, вначале — из одного и того же распределения. Проверить однородность с помощью критерия Уилкоксона.
Затем смоделировать неоднородность (например, заменить в одной из выборок часть данных на данные из другого распределения).
Проверить гипотезу об однородности.