Лекция 8. Непараметрические критерии однородности и независимости

Буре В.М., Грауэр Л.В.

ШАД

Санкт-Петербург, 2013

Содержание

Критерии однородности Вилкоксона и Манна-Уитни

Пусть имеются две независимые выборки $X_{[n]} = (X_1, \ldots, X_n)$ и $Y_{[m]} = (Y_1, \ldots, Y_m)$ из двух генеральных совокупностей с непрерывными функциями распределения равными соответственно F и G. Сформулируем гипотезы:

- $H_0: \ F(x) = G(x)$ для всех $x \in \mathbb{R}$.
- $H_1: \ F(x) \geqslant G(x)$ для всех $x \in \mathbb{R}$ — правосторонняя альтернативная гипотеза.
- $H_1': \ F(x) \leqslant G(x)$ для всех $x \in \mathbb{R}$ — левосторонняя альтернативная гипотеза.
- $H_1'': \ F(x) \neq G(x)$ для всех $x \in \mathbb{R}$ — двусторонняя альтернативная гипотеза (т.е. выполнена H_1 или H_1').

Критерий однородности Вилкоксона

Без ограничения общности будем считать, что $m \leqslant n$. Составим объединенную выборку $Z_{[n+m]} = (X_{[n]}, Y_{[m]})$. Построим вариационный ряд объединенной выборки:

$$z_{(1)} < z_{(2)} < \ldots < z_{(m+n)}.$$

Если распределения генеральных совокупностей непрерывны, то совпадения возможны только с нулевой вероятностью. В дальнейшем будем предполагать, что совпадений нет.

В литературе [1], [2] имеются поправочные формулы, когда совпадения возникают, например, вследствие округления. Однако, следует заметить, что при большом числе совпадений рассматриваемые критерии неприменимы.

Найдем, какие места занимают в вариационном ряду, построенном по объединенной выборке, элементы выборки $Y_{[m]}$. Назовем эти номера *рангами* элементов выборки $Y_{[m]}$ в объединенной выборке $Z_{[n+m]}$:

$$rank(Y_1) = s_1, \;\; rank(Y_2) = s_2, \;\; \ldots, \;\; rank(Y_m) = s_m.$$

Рассмотрим статистку критерия:

$$W = \sum_{i=1}^{m} s_i.$$

Очевидно, что минимальным значением статистики Вилкоксона может быть величина $W_{\min} = m(m+1)/2$, максимальным — $W_{\max} = mn + m(m+1)/2$.
Поэтому статистика W находится в промежутке

$$[m(m+1)/2; mn + m(m+1)/2].$$

Распределение статистики Вилкоксона W является симметричным относительно середины данного промежутка при условии справедливости нулевой гипотезы H_0.

Для больших объемов выборок существуют аппроксимации статистики W, которую можно найти, например, в [1].

Если справедлива альтернативная гипотеза H_1, то чаще будут встречаться события $x_i < y_j$, то есть, распределение статистики W перестанет быть симметричным относительно середины и будет сдвинуто вправо.
Если справедлива альтернативная гипотеза H_1', то распределение статистики W будет сдвинуто влево, так как чаще будут выполняться события $x_i > y_j$.
Если выбрана альтернативой гипотеза H_1, то критическая область для нулевой гипотезы H_0 будет иметь вид:

$$S = \left[c_1, mn + \frac{m(m+1)}{2}\right].$$

Если выбрана альтернативой гипотеза H_1', то критическая область для нулевой гипотезы H_0 будет иметь вид:

$$S = \left[\frac{m(m+1)}{2}, c_2\right].$$

Если выбрана альтернативой гипотеза H_1'', то критическая область для нулевой гипотезы H_0 будет иметь вид:

$$S = \left[\frac{m(m+1)}{2}, c_3\right] \cup \left[c_4, mn + \frac{m(m+1)}{2}\right].$$

При этом, константы c_1, c_2, c_3, c_4 следует находить по таблицам распределения статистики Вилкоксона W, рассчитанным при условии справедливости нулевой гипотезы H_0 для разных m и n ([1]).

В качестве искомых констант выбираются квантили распределения. При этом, константы c_1 и c_2 симметричны относительно середины промежутка $[m(m+1)/2, mn + m(m+1)/2]$. Также симметрично относительно середины этого промежутка расположены константы c_3 и c_4.

Общее требование заключается в том, что вероятность попадания статистики W в критическую область при условии справедливости нулевой гипотезы H_0 должна быть равна заданному значению α:

$$P_0\{W \in S\} = \alpha.$$

Критерий Манна-Уитни

Будем проверять те же нулевую и альтернативные гипотезы, что и в критерии Вилкоксона.
Запишем статистику критерия Манна-Уитни:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I\{X_i < Y_j\},$$

где

$$I\{X_i < Y_j\} = \begin{cases} 1, & X_i < Y_j; \\ 0, & X_i > Y_j. \end{cases}$$

Находим в таблице [3] критические значения распределения уровня α статистики Манна-Уитни U при условии справедливости гипотезы однородности H_0: $U_{left} = U_{\alpha,n,m}$, $U_{right} = V_{\alpha,n,m}$ и сравниваем с ними статистику критерия.
Если $U \geqslant U_{right}$, то нулевая гипотеза отвергается в пользу правосторонней альтернативной гипотезы H_1.
Если $U \leqslant U_{left}$, то нулевая гипотеза отвергается в пользу левосторонней альтернативной гипотезы H_1'.

Для случая двусторонней альтернативы H_1'' находятся точки $U_{right} = V_{\alpha/2,n,m}$ и $U_{left} = U_{\alpha/2,n,m}$, и нулевая гипотеза H_0 отвергается при выполнении любого из неравенств: $U \geqslant U_{right}$, $U \leqslant U_{left}$.

Взаимосвязь критериев Вилкоксона и Манна-Уитни

Статистики Вилкоксона и Манна-Уитни связаны между собой следующим соотношением:

$$W = U + \frac{m(m+1)}{2}.$$

Для доказательства справедливости приведенной формулы заметим, что значение статистики Вилкоксона W, по существу, равно количеству неравенств вида: $\{X_i < Y_j\}$ плюс сумма рангов элементов выборки $Y_{[m]}$ в самой выборке $Y_{[m]}$. Последняя сумма представляет собой арифметическую прогрессию чисел от 1 до m шагом 1. Полученная формула позволяет пересчитывать значение одной статистики в другую и пользоваться тем распределением, которое более удобно для вычислений.

Непараметрические критерии анализа парных повторных наблюдений

Рассмотрим две зависимые выборки: $X_{[n]} = (X_1, \ldots, X_n)$ и $Y_{[n]} = (Y_1, \ldots, Y_n)$ из генеральных совокупностей с функциями распределения равными соответственно $F(x)$ и $G(x)$, которые считаем непрерывными.

Выборки имеют одинаковый объем, и зависимость носит следующий характер. Внутри каждой из выборок элементы независимы, но X_i и Y_i — зависимые наблюдения. Чаще всего i означает номер объекта, а X_i и Y_i — два наблюдения над одним и тем же объектом до и после некоторого воздействия.

Сформулируем гипотезы:

- $H_0: \ F(x) = G(x)$ для всех $x \in \mathbb{R}$.
- $H_1: \ F(x) \geqslant G(x)$ для всех $x \in \mathbb{R}$.
- $H_1': \ F(x) \leqslant G(x)$ для всех $x \in \mathbb{R}$.
- $H_1'': \ F(x) \neq G(x)$ для всех $x \in \mathbb{R}$.

Критерий знаков

Составим разности $z_i = X_i - Y_i$. Случайные величины z_i, $i = 1, \ldots, n$, взаимно независимы и одинаково распределены. Переформулируем гипотезы:

- $H_0: \; P\{z_i < 0\} = P\{z_i > 0\} = 1/2.$
- $H_1: \; P\{z_i < 0\} > P\{z_i > 0\}.$
- $H_1': \; P\{z_i < 0\} < P\{z_i > 0\}.$
- $H_1'': \; P\{z_i < 0\} \neq P\{z_i > 0\}.$

Так как в гипотезах фигурируют вероятности, то можем рассматривать схему Бернулли, где событие $z_i < 0$ означает успех.

Статистика критерия:

$$L = \sum_{i=1}^{n} I\{z_i < 0\}.$$

При выполнении гипотезы H_0 статистика L подчиняется биномиальному распределению с вероятностью успеха $p = 1/2$.
Критическая область для нулевой гипотезы при выборе

- альтернативной гипотезы H_1 имеет вид: $(c_1, n]$;
- альтернативной гипотезы H_1' имеет вид: $[0, c_2)$;
- альтернативной гипотезы H_1'' имеет вид: $[0, c_3) \cup (c_4, n]$.

Для нахождения констант c_1, c_2, c_3, c_4 можно использовать таблицы вероятностей биномиального распределения, следуя общему правила: попадание статистики критерия L в критическую область при условии выполнения гипотезы H_0 равно α.

Критерий знаковых ранговых сумм Вилкоксона

Составим разности $z_i = X_i - Y_i$. Предполагаем, что величины z_i не зависят друг от друга. Рассмотрим случаи, когда $z_i < 0$ и $z_i > 0$.
Рассмотрим гипотезы:

- $H_0: \ P\{z_i < 0\} = P\{z_i > 0\} = 1/2.$
- $H_1: \ P\{z_i < 0\} > P\{z_i > 0\}.$
- $H_1': \ P\{z_i < 0\} < P\{z_i > 0\}.$
- $H_1'': \ P\{z_i < 0\} \neq P\{z_i > 0\}.$

Построим вариационный ряд из модулей разностей:
$|z_1|, \ldots, |z_n|$.
Сопоставим каждому элементу вариационного ряда ранг:
$s_1 = rank(|z_1|), \ldots, s_n = rank(|z_n|)$.

Составим ранговую статистику

$$U = \sum_{i=1}^{n} \Psi_i s_i, \text{ где } \Psi_i = \begin{cases} 1, & z_i > 0; \\ 0, & z_i < 0. \end{cases}$$

Используем таблицы [3] для нахождения критических значений статистики U и сравним их с полученным значением статистики.
При выборе левосторонней альтернативы H_1 критическая область имеет вид: $[0, c_1]$.
При выборе правосторонней альтернативы H_1' критическая область имеет вид: $[c_2, n(n+1)/2]$.
При выборе двусторонней альтернативы H_1'' критическая область имеет вид: $[0, c_3] \cup [c_4, n(n+1)/2]$.
Вероятность попадания статистики критерия U в критическую область при условии выполнения гипотезы H_0 равно α.

Критерий однородности Колмогорова–Смирнова

Пусть имеется две выборки $X_{[n]} = \{X_1, \ldots, X_n\}$ и $Y_{[m]} = \{Y_1, \ldots, Y_m\}$ из генеральных совокупностей ξ и η соответственно.
Объемы выборок могут быть различны, но, не нарушая общности, предположим, что $m \leqslant n$.
Функции распределения этих генеральных совокупностей равны $F(x)$ и $G(x)$ соответственно. Наложим дополнительное ограничение: функции распределения $F(x)$ и $G(x)$ непрерывны.

Критерий Колмогорова–Смирнова проверяет гипотезу о равенстве функций распределения двух генеральных совокупностей ξ и η, из которых извлечены выборки $X_{[n]}$ и $Y_{[m]}$ соответственно:
$H_0 : F(x) = G(x)$ для всех $x \in \mathbb{R}$,
при альтернативной $H_1 : F(x) \neq G(x)$.

Критерий основан на использовании эмпирических функций распределения $F_n^*(x)$ и $G_m^*(x)$.

Теорема 1

Пусть

$$D_{m,n} = \sup_{x \in \mathbb{R}} \left| G_m^*(x, Y_{[m]}) - F_n^*(x, X_{[n]}) \right|.$$

Если истинная функция распределения $F_0(x) = F(x) = G(x)$ *непрерывна, тогда*

$$P_0 \left\{ \sqrt{\frac{mn}{m+n}} D_{m,n} \leqslant z \right\} \longrightarrow K(z) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 z^2}, \qquad (1)$$

Статистика Смирнова определяется следующей формулой:

$$D_{m,n} = \sup_{|x|<\infty} |G_m^*(x) - F_n^*(x)| \tag{2}$$

На практике значение статистики $D_{m,n}$ рекомендуется вычислять по формулам:

$$D_{m,n}^+ = \max_{1\leqslant r\leqslant m} \left[\frac{r}{m} - F_n^*(y_{(r)})\right] = \max_{1\leqslant s\leqslant n} \left[G_m^*(x_{(s)}) - \frac{s-1}{n}\right], \tag{3}$$

$$D_{m,n}^- = \max_{1\leqslant r\leqslant m} \left[F_n^*(y_{(r)}) - \frac{r-1}{m}\right] = \max_{1\leqslant s\leqslant n} \left[\frac{s}{n} - G_m^*(x_{(s)})\right], \tag{4}$$

$$D_{m,n} = \max(D_{m,n}^+, D_{m,n}^-), \tag{5}$$

где $X_{(s)}$ и $Y_{(r)}$ — элементы вариационных рядов $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$ и $Y_{(1)} \leqslant Y_{(2)} \leqslant \ldots \leqslant Y_{(m)}$, построенных по выборкам $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$.

При справедливости нулевой гипотезы и неограниченном увеличении объемов выборок исправленная статистика

$$\sqrt{\frac{mn}{m+n}}D_{m,n} \tag{6}$$

асимптотически подчиняется распределению Колмогорова с функцией распределения $K(z)$ из правой части (1).
Таблицу значений функции распределения Колмогорова можно найти в [1].
Критическая область для гипотезы H_0 при использовании статистики 6 имеет вид: $S = (k_{1-\alpha}, \infty)$, где $k_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения Колмогорова 1.

Критерий однородности хи-квадрат

С помощью критерия χ^2 можно анализировать однородность любого конечного числа выборок.
Пусть имеется s независимых выборок, содержащих соотвественно $n_1, n_2, \ldots, n_s$ элементов: $\xi_1 : X^1_{[n_1]}, \ldots, \xi_s : X^s_{[n_s]}$. Сформулируем гипотезы:

- H_0 — выборки взяты из одной и той же совокупности $F_{\xi_1} = \ldots = F_{\xi_s} = F_\xi$,
- H_1 — выборки взяты из разных генеральных совокупностей.

Каждую выборку разобьем на k групп Δ_i, $i = 1, \ldots, k$.
Пусть n_{ij} — число элементов j-ой выборки, попавших в множество Δ_i, $i = 1, \ldots, k$, $j = 1, \ldots, s$.

Пусть верятность попадания случайной величины ξ в множество Δ_i равна p_i: $p_i = P(\xi \in \Delta_i)$, $i = 1, \ldots, k$.

Пусть $n_j = \sum_{i=1}^{k} n_{ij}$ — общее число элементов j-ой выборки, $j = 1, \ldots, s$.

Если гипотеза H_0 верна, то относительная частота $\frac{n_{ij}}{n_j}$ попадания элементов j-ой выборки в множество Δ_i будет близка к вероятности p_i. По методу ξ^2 для одной j-ой выборки статистикой критерия является величина

$$\sum_{i=1}^{k} \frac{n_j}{p_i} \left(\frac{n_{ij}}{n_j} - p_i \right)^2 = \sum_{i=1}^{k} \frac{(n_{ij} - n_j p_i)^2}{n_j p_i},$$

а для всех выборок

$$\sum_{j=1}^{s} \sum_{i=1}^{k} \frac{(n_{ij} - n_j p_i)^2}{n_j p_i}. \tag{7}$$

Вероятности p_i, $i = 1, \ldots, k$, неизвестны. Их оценки находим методом максимума правдоподобия, объединяя все выборки в одну объемом

$$n = \sum_{j=1}^{s} \sum_{i=1}^{k} n_{ij}.$$

Учтем связь

$$p_1 + \ldots + p_k = 1. \tag{8}$$

Пусть $\nu_i = \sum_{j=1}^{s} n_{ij}$ — общее число элементов всех выборок, попавших в множество Δ_i, $i = 1, \ldots, k$.

Рассмотрим функцию правдоподобия $L = c \prod_i p_i^{\nu_i}$ и ее логарифм

$$\ln L = \sum_{i=1}^{k} \nu_i \ln p_i + \ln c.$$

Образуем функцию Лагранжа, учитывая связь (8)

$$\Lambda = \ln L + \lambda \left(1 - \sum_{i=1}^{k} p_i\right) = \sum_{i=1}^{k} \nu_i \ln p_i + \lambda \left(1 - \sum_{i=1}^{k} p_i\right) + \ln c.$$

Найдем максимум Λ

$$\frac{\delta \Lambda}{\delta p_l} = \frac{\nu_l}{p_l} - \lambda = 0, \quad p_l = \frac{\nu_l}{\lambda}, \quad l = 1, \ldots, k.$$

Из уравнения связи (8) получаем $\lambda = n$. Тогда

$$\hat{p}_i = \frac{\nu_i}{n}, \quad i = 1, \ldots, k.$$

Подставляя полученные оценки в (7) вместо вероятностей p_i получаем

$$\chi^2 = n \sum_{j=1}^{s} \sum_{i=1}^{k} \frac{(n_{ij} - n_j \nu_i / n)^2}{n_j \nu_i} = n \left(\sum_{j=1}^{s} \sum_{i=1}^{k} \frac{(n_{ij}^2)^2}{n_j \nu_i} - 1 \right) \qquad (9)$$

Статистика (9) асимпотитически при $n \to \infty$ распределена по закону χ^2 с числом степеней свободы $r = (s-1)(k-1)$ (см [5]).
Критическая область для гипотезы H_0 при использовании статистики 9 имеет вид: $S = (\chi^2_{1-\alpha}, \infty)$, где $\chi^2_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения χ^2.

В случае проверки гипотезы об однородности двух выборок ($s = 2$) статистика принимает вид

$$\chi^2 = n_1 n_2 \sum_{i=1}^{k} \frac{1}{\nu_i} \left(\frac{n_{i1}}{n_1} - \frac{n_{i2}}{n_2} \right)^2 = \sum_{i=1}^{k} \frac{1}{n_{i1} + n_{i2}} \left(n_{i1} \sqrt{\frac{n_2}{n_1}} - n_{i2} \sqrt{\frac{n_1}{n_2}} \right).$$

Число степеней свободы статистики χ^2 равно $r = k - 1$.

Замечание 1

Проверка гипотезы о независимости качественных признаков A и B с помощью критерия χ^2 подробно рассмотрена в Лб.

Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона применяется к данным, измеренным в шкале отношений.

Определение 1

Шкалой отношений называют такую шкалу с непрерывным множеством числовых значений, в которой о двух сопоставляемых объектах можно сказать не только, одинаковы они или различны, не только, в каком из них признак выражен сильнее, но и во сколько раз сильнее этот признак выражен.

Предположим, что есть генеральная совокупность, каждый элемент которой обладает двумя количественными признаками. Если случайным образом извлекать объекты, то пусть ξ — значение, которое принимает первый признак, η — значение, которое принимает второй признак. Величины ξ и η — случайные.

Корреляция случайных величин ξ и η выражается следующей формулой:

$$\rho(\xi,\eta) = \frac{cov(\xi,\eta)}{\sqrt{D\xi}\sqrt{D\eta}}.$$

Если случайные величины ξ и η независимы, то корреляция равна нулю. Обратное утверждение, вообще говоря, неверно.
Если $|\rho| = 1$, то существует линейная связь между величинами ξ и η такая, что $\eta = a + b\xi$.

Получим оценку коэффициента корреляции — выборочный коэффициент корреляции Пирсона, который определяется выражением:

$$r_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y},$$

где $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.
Здесь предполагается, что задана двумерная выборка: (X_1, Y_1), ..., (X_n, Y_n).

Пусть совместное распределение случайных величин ξ и η является **нормальным**.
Если вектор $(\xi, \eta)^T$ подчиняется совместному нормальному распределению с вектором математических ожиданий $a = (a_1, a_2)^T$ и ковариационной матрицей

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

то корреляция случайных величин ξ и η равна нулю тогда и только тогда, когда эти случайные величины взаимно независимы, $\sigma_1^2 = D\xi$, $\sigma_2^2 = D\eta$.
Сформулируем гипотезы:

- H_0: $\rho = 0$ — гипотеза о независимости.
- H_1: $\rho > 0$.
- H_1': $\rho < 0$.
- H_1'': $\rho \neq 0$ — двусторонняя альтернатива.

Справедливо утверждение. При сделанных предположениях о распределении случайного вектора $(\xi, \eta)^T$, статистика

$$t = r_{X,Y}\sqrt{n-2}/\sqrt{1-r_{X,Y}^2} \tag{10}$$

при выполнении гипотезы H_0 подчиняется распределению Стьюдента с $n-2$ степенями свободы.
При использовании статистики (10) для альтернативы H_1 критическая область для гипотезы H_0 имеет вид: $S = (t_{1-\alpha,n-2}, \infty)$,
для альтернативы H_1' критическая область для гипотезы H_0 имеет вид: $S = (-\infty, t_{\alpha,n-2})$.
Критическая область для нулевой гипотезы H_0 при альтернативе H_1'' будет иметь вид:

$$S = \left(-\infty, t_{\frac{\alpha}{2},n-2}\right) \cup \left(t_{1-\frac{\alpha}{2},n-2}, +\infty\right),$$

где $t_{\beta,n-2}$ — квантиль уровня β распределения Стьюдента с $n-2$ степенями свободы. Если значение статистики $t \in S$, то гипотеза H_0 отклоняется, если $t \notin S$, то гипотеза H_0 принимается. Величина вероятности ошибки первого рода равна α.

Коэффициенты ранговой корреляции Спирмена и Кенделла

Рассматриваемые в этой части коэффициенты вычисляются только для порядковых шкал.

Определение 2

Шкалы, в которых существенен лишь взаимный порядок, в котором следуют результаты измерений, а не их количественные значения, называют порядковыми или ординальными шкалами.

Пусть имеется два признака A и B, между которыми мы хотим установить наличие зависимости или независимости. Пусть $(X_1, \ldots, X_n)$ — измерение степени выраженности признака A, $(Y_1, \ldots, Y_n)$ — измерение степени выраженности признака B. Каждый объект характеризует пара (X_i, Y_j), $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant n$.
Для обоих коэффициентов корреляции характерно то обстоятельство, что они обнаруживают лишь монотонную зависимость признаков.

Коэффициент ранговой корреляции Спирмена

Проранжируем наблюдения $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$, ранги которых будут соответственно обозначаться $r_1, \ldots, r_n$ и $s_1, \ldots, s_n$, то есть, r_i — номер наблюдения X_i в вариационном ряду, построенном по наблюдениям $X_1, \ldots, X_n$. Аналогично, s_i — номер наблюдения Y_i в вариационном ряду, построенном по наблюдениям $Y_1, \ldots, Y_n$. Будем предполагать, что в выборках нет повторяющихся элементов. Рассмотрим статистику Спирмена $S = \sum_{i=1}^{n} (s_i - r_i)^2$. Вычислим коэффициент ранговой корреляции Спирмена:

$$\varrho = 1 - \frac{6S}{n^3 - n}.$$

Нетрудно показать, что $|\varrho| \leqslant 1$.

Если $|\varrho| = 1$, то это означает полную зависимость одного признака от другого, либо, иначе говоря, полную предсказуемость одной выборки по другой.
Если ранги признаков совпадают, то $\varrho = 1$.
Если последовательности рангов полностью противоположны, то $\varrho = -1$.
Оба случая означают полную предсказуемость одной ранговой последовательности по другой, или, другими словами, полную зависимость признаков A и B.
Сформулируем гипотезы:

- H_0: признаки A и B взаимно независимы.
- H_1: имеется монотонная положительная связь признаков.
- H_1': имеется монотонная отрицательная связь признаков.
- H_1'': имеется монотонная связь признаков.

Гипотеза H_0 соответствует отсутствию взаимосвязи между признаками или, иначе говоря, независимости признаков.
Если гипотеза H_0 справедлива, то распределение статистики ϱ симметрично и концентрируется около нуля.
При наличии зависимости распределение окажется другим. Для монотонной положительной зависимости распределение ϱ сдвинуто вправо, для монотонной отрицательной — влево.
Для проверки гипотезы H_0 необходимо обратиться к таблицам распределения коэффициента Спирмена [1], вычисленным в предположении истинности гипотезы H_0.
По заданной вероятности ошибки первого рода α необходимо найти соответствующие пороговые значения, после чего, при попадании вычисленного по наблюдениям коэффициента ϱ в критическую область следует отклонить гипотезу H_0 в пользу альтернативной гипотезы.

- При выборе в качестве альтернативы гипотезы H_1 (положительная монотонная связь) критическую область следует выбрать в виде: $(c_1, 1]$.
- При выборе альтернативы гипотезы H_1' (отрицательная монотонная связь) критическую область следует выбрать в виде: $[-1, c_2]$.
- При выборе альтернативной гипотезы H_1'' критическая область для гипотезы H_0 имеет вид: $[-1, c_3) \cup (c_4, 1]$.

Пороговые значения c_1, c_2, c_3, c_4 определяются из статистических таблиц так, чтобы вероятность попадания в критическую область при выполнении гипотезы H_0 была равна α.

Коэффициент ранговой корреляции Кенделла

Для вычисления статистики Кенделла достаточно посчитать количество инверсий (число несогласованных пар), которое надо сделать для того, чтобы одно упорядочение объектов превратилось в другое.

Пусть есть пары наблюдений каждого из признаков (X_1, Y_1), ..., (X_n, Y_n). Упорядочим наблюдения первого признака и проранжируем их рангами от 1 до n. Затем ранжируем последовательность наблюдений второго признака, при этом объекты перенумерованы в соответствии с рангами первой совокупности. Пусть во втором наборе приписаны каждому наблюдению ранги $z_1, \ldots, z_n$, то есть, теперь все объекты характеризуются парами рангов: $(1, z_1)$, ..., (n, z_n). После перенумерования ранги измерений признаков A представляют собой новые номера самих объектов.

Пусть R — число инверсий в выборке $\{z_1, \ldots, z_n\}$. Инверсия суть нарушение порядка.
Рассмотрим коэффициент ранговой корреляции Кенделла:

$$\tau = 1 - \frac{4R}{n(n-1)} \quad \text{или}$$

$$\tau = \frac{\sum_{j=1}^{n} \sum_{i<j} (X_i - X_j)(Y_i - Y_j)}{n(n-1)}$$

Нетрудно доказать, что $|\tau| \leqslant 1$. При этом, $|\tau| = 1$ означает полную предсказуемость (зависимость) признаков.
Проверяемые гипотезы

- H_0: признаки A и B взаимно независимы.
- H_1: имеется монотонная положительная связь признаков.
- H_1': имеется монотонная отрицательная связь признаков.
- H_1'': имеется монотонная связь признаков.

По заданной вероятности ошибки первого рода α необходимо найти соответствующие пороговые значения, после чего, при попадании вычисленного по наблюдениям коэффициента τ в критическую область следует отклонить гипотезу H_0 в пользу альтернативной гипотезы.

- При выборе в качестве альтернативы гипотезы H_1 (положительная монотонная связь) критическую область следует выбрать в виде: $(c_1, 1]$.
- При выборе альтернативы гипотезы H_1' (отрицательная монотонная связь) критическую область следует выбрать в виде: $[-1, c_2]$.
- При выборе альтернативной гипотезы H_1'' критическая область для гипотезы H_0 имеет вид: $[-1, c_3) \cup (c_4, 1]$.

Пороговые значения c_1, c_2, c_3, c_4 определяются из статистических таблиц для коэффициента ранговой корреляции Кэндалла [1] так, чтобы вероятность попадания в критическую область при выполнении гипотезы H_0 была равна α.

При больших n при справедливости гипотезы H_0 случайные величины $\sqrt{n-1}\varrho$ и $\tau\sqrt{9n(n+1)/(2(2n+5))}$ приближенно распределены по стандартному нормальному закону $N(0,1)$, что позволяет проверять гипотезу H_0, пользуясь указанной асимптотикой.

Литература

- Большев Л. Н., Смирнов Н. В.Таблицы математической статистики. — М.: Изд. Наука, 1983.
- Тюрин Ю. Н., Макаров А. А.Статистический анализ опытных данных на компьютере. — Под ред. В.Э. Фигурнова. М.: ИНФРА-М, 1998.
- Холлендер М., Вулф Д.Непраметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.
- Greenwood P. E., Nikulin M. S.A Guide to Chi-Squared Testing. New York, John Wiley & Sons, Inc., 1996.
- Крамер Г.Математические методы статистики. М.: Мир, 1975