Сложная гипотеза. Критерии независимости

Для проверки сложных гипотез используются Критерий Спирмена, Критерии независимости.

Критерий Спирмена. Пусть имеется *случайная выборка* $(X_1, Y_1), \ldots, (X_n, Y_n)$ из *генеральной совокупности* двумерной непрерывной случайной величины (X, Y) с функцией распределения $F(t,\tau)$, а $F_X(t)$ и $F_Y(\tau)$ — функции распределения случайных величин X и Y соответственно. Если случайные величины X и Y имеют нормальные распределения, то для проверки *статистической гипотезы* об их независимости

$$H_0\colon\ F(t,\tau) = F_X(t)\,F_Y(\tau) \tag{5.18}$$

можно использовать процедуру, связанную с вычислениями *выборочного коэффициента корреляции* (см. формулу (6.12)).

Если же о распределениях непрерывных случайных величин X и Y ничего не известно, то для проверки *основной гипотезы* (5.18) при *альтернативной гипотезе*

$$H_1\colon\ F(t,\tau) \neq F_X(t)\,F_Y(\tau)\ \text{ для некоторых } (t,\,\tau) \in \mathbb{R}^2$$

используют ***ранговый критерий Спирмена***, основанный на следующем понятии.

Определение 5.1. ***Рангом*** $R_i(\vec{z}_N)$ ***элемента*** z_i числовой ***последовательности*** $\vec{z}_N = (z_1, \ldots, z_N)$ называют его порядковый номер в *вариационном ряду* $z_{(1)}, \ldots, z_{(N)}$.

Согласно определению, $R_i(\vec{z}_N)$ — это число элементов последовательности $z_1, \ldots, z_N$, не бóльших чем z_i, которое можно записать следующим образом:

$$R_i(\vec{z}_N) = 1 + \sum_{k=1}^{N} \eta(z_i - z_k),$$

где $\eta(t)$ — *функция Хевисайда*. Ранг любого элемента последовательности $\vec{z}_N$ — это натуральное число в диапазоне от 1 до N, причем ранг наименьшего элемента последовательности равен 1, а ранг наибольшего — N.

Пример 5.5. Рассмотрим выборку $\vec{z}_4 = (3{,}8,\ 4{,}7,\ -2{,}6,\ 17{,}3)$. Ее вариационный ряд имеет вид $-2{,}6;\ 3{,}8;\ 4{,}7;\ 17{,}3$. Поэтому $R_1(\vec{z}_4) = 2$, $R_2(\vec{z}_4) = 3$, $R_3(\vec{z}_4) = 1$, $R_4(\vec{z}_4) = 4$. #

Определение 5.2. ***Рангом элемента*** Z_i ***случайной выборки*** $\vec{Z}_N = (Z_1, \ldots, Z_N)$ называют случайную величину

$R_i(\vec{Z}_N)$, реализация которой $R_i(\vec{z}_N)$ есть ранг реализации z_i случайной величины Z_i в вариационном ряду $z_{(1)}, \ldots, z_{(N)}$.

Обозначим через $R_i = R_i(\vec{X}_n)$ — ранг элемента X_i случайной выборки $X_1, \ldots, X_n$, а через $S_i = S_i(\vec{Y}_n)$ — ранг элемента Y_i случайной выборки $Y_1, \ldots, Y_n$.

Ранговым коэффициентом корреляции Спирмена назовем случайную величину

$$\rho(\vec{X}_n, \vec{Y}_n) = \frac{\sum\limits_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum\limits_{i=1}^{n}(R_i - \overline{R})^2 \sum\limits_{i=1}^{n}(S_i - \overline{S})^2}}, \qquad (5.19)$$

где

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i, \qquad \overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i.$$

Согласно определению, $R_i(\vec{z}_N)$ — это число элементов последовательности $z_1, \ldots, z_N$, не бóльших чем z_i, которое можно записать следующим образом:

$$R_i(\vec{z}_N) = 1 + \sum_{k=1}^{N} \eta(z_i - z_k),$$

где $\eta(t)$ — *функция Хевисайда.* Ранг любого элемента последовательности $\vec{z}_N$ — это натуральное число в диапазоне от 1 до N, причем ранг наименьшего элемента последовательности равен 1, а ранг наибольшего — N.

Пример 5.5. Рассмотрим выборку $\vec{z}_4 = (3{,}8,\ 4{,}7,\ -2{,}6,\ 17{,}3)$. Ее вариационный ряд имеет вид $-2{,}6;\ 3{,}8;\ 4{,}7;\ 17{,}3$. Поэтому $R_1(\vec{z}_4) = 2$, $R_2(\vec{z}_4) = 3$, $R_3(\vec{z}_4) = 1$, $R_4(\vec{z}_4) = 4$. #

Определение 5.2. ***Рангом элемента Z_i случайной выборки*** $\vec{Z}_N = (Z_1, \ldots, Z_N)$ называют случайную величину $R_i(\vec{Z}_N)$, реализация которой $R_i(\vec{z}_N)$ есть ранг реализации z_i случайной величины Z_i в вариационном ряду $z_{(1)}, \ldots, z_{(N)}$.

Обозначим через $R_i = R_i(\vec{X}_n)$ — ранг элемента X_i случайной выборки $X_1, \ldots, X_n$, а через $S_i = S_i(\vec{Y}_n)$ — ранг элемента Y_i случайной выборки $Y_1, \ldots, Y_n$.

Ранговым коэффициентом корреляции Спирмена назовем случайную величину

$$\rho(\vec{X}_n, \vec{Y}_n) = \frac{\sum\limits_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum\limits_{i=1}^{n}(R_i - \overline{R})^2 \sum\limits_{i=1}^{n}(S_i - \overline{S})^2}}, \tag{5.19}$$

где

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i, \qquad \overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i.$$

Статистика (5.19) является выборочным коэффициентом корреляции последовательностей рангов $R_1, \ldots, R_n$ и $S_1, \ldots, S_n$.

Согласно определению рангов R_i, S_i, $i = \overline{1, n}$,

$$\overline{R} = \overline{S} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2},$$

и можно показать, что

$$\rho(\vec{X}_n, \vec{Y}_n) = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - S_i)^2. \tag{5.20}$$

Без ограничения общности можно считать, что значения пар наблюдений (x_i, y_i), $i = \overline{1, n}$, занумерованы в порядке возрастания их первых элементов, т.е. так, что выполняются неравенства

$$x_1 < x_2 < \ldots < x_n.$$

В этом случае реализация r_i ранга R_i равна i, $i = \overline{1, n}$, и значение $\rho(\vec{x}_n, \vec{y}_n)$ статистики $\rho(\vec{X}_n, \vec{Y}_n)$ можно вычислить по формуле

$$\rho(\vec{x}_n, \vec{y}_n) = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(i - s_i)^2, \tag{5.21}$$

где s_i — реализация ранга S_i, $i = \overline{1, n}$.

Пример 5.6. В табл. 5.1 представлены $n = 10$ значений (x_i, y_i), $i = \overline{1, 10}$, непрерывной двумерной случайной величины (X, Y). Проверим на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о независимости случайных величин X и Y.

Таблица 5.1

x_i	−1,63	1,11	1,15	−1,93	0,38	−1,08	−0,31	0,60	0,12	0,92
y_i	0,54	0,88	−1,21	0,89	−0,64	−0,21	0,08	−0,74	0,79	0,14

Строим последовательность рангов (табл. 5.2). По формуле (5.20) вычисляем реализацию статистики $\rho(\vec{X}_n, \vec{Y}_n)$

$$\rho(\vec{x}_n, \vec{y}_n) = 1 - \frac{6}{10(10^2-1)}\Big((2-7)^2 + (9-9^2) + (10-1)^2 + (1-10)^2 +$$

$$+ (6-3)^2 + (3-4)^2 + (4-5)^2 + (7-2)^2 + (5-8)^2 + (8-6)^2\Big) =$$

$$= 1 - \frac{6}{990}\Big(25 + 0 + 81 + 81 + 9 + 1 + 1 + 25 + 9 + 2\Big) \approx -0{,}4118.$$

Таблица 5.2

r_i	2	9	10	1	6	3	4	7	5	8
s_i	7	9	1	10	3	4	5	2	8	6

По таблицам распределения статистики $\rho(\vec{X}_n, \vec{Y}_n)$ рангового критерия Спирмена* находим квантили

$$\rho_{0{,}952} = 0{,}6726, \qquad \rho_{0{,}97} = 0{,}7374, \qquad \rho_{0{,}983} = 0{,}80223, \quad (5.23)$$

а квантили $\rho_{1-\alpha/2} = \rho_{0{,}975}$ нет, так как $\rho(\vec{X}_n, \vec{Y}_n)$ — дискретная случайная величина. Тем не менее, из значений квантилей (5.23) заключаем, что $|\rho(\vec{x}_n, \vec{y}_n)| < \rho_{0{,}952}$ и H_0 не отклоняется даже на большем уровне значимости.

Таблицы сопряженности признаков и критерий χ^2. Пусть имеется случайная выборка

$$(\vec{X}_n, \vec{Y}_n) = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

из генеральной совокупности двумерной дискретной случайной величины (X, Y), где случайная величина X может принимать значения $u_1, \ldots, u_r$, а случайная величина Y — значения $v_1, \ldots, v_s$. Определим случайную величину $n_{ij}(\vec{X}_n, \vec{Y}_n)$, реализация n_{ij} которой равна количеству элементов выборки $(\vec{x}_n, \vec{y}_n) = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$, совпадающих с элементом (u_i, v_j), $i = \overline{1, r}$, $j = \overline{1, s}$.

Пример 5.8. Даны *выборка объема* $m = 25$

0,00;	−0,53;	1,47;	0,96;	3,98;	3,22;	0,25;
0,31;	−0,64;	−1,26;	−0,92;	−1,36;	0,96;	1,39;
−0,81;	1,12;	−0,62;	−0,66;	1,07;	−0,52;	0,48;
−1,00;	−0,96;	−1,43;	−1,09			

из *распределения Коши* с плотностью

$$p_X(x) = \frac{1}{\pi(1+x^2)}$$

и выборка объема $n = 28$

−0,88;	0,41;	−0,64;	−0,81;	−0,09;	−0,71;	−0,00;
0,49;	−0,65;	0,59;	0,17;	−0,46;	0,99;	−0,24;
−0,98;	−0,85;	−0,09;	−0,63;	0,68;	0,02;	−0,59;
−0,02;	−0,45;	−0,50;	0,40;	0,29;	−0,17;	−0,43

из равномерного распределения на отрезке $[-1, 1]$ с плотностью $p_Y(x)$. Проверим при помощи *критерия Смирнова статистическую гипотезу* о равенстве функций p_X и p_Y.

Объединив заданные выборки и построив вариационный ряд, по формуле (5.15) найдем соответствующие этому ряду значения δ_i, $i = \overline{1, N}$, $N = 45$:

0; 0; 0; 0; 0; 1; 0; 0; 1; 1; 1; 0; 1; 0; 1; 1; 0; 1; 0; 1;
0; 0; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 0; 1; 1; 0; 1; 0; 1; 1;
0; 1; 1; 1; 0; 0; 1; 0; 0; 0; 0; 0; 0.

Вычислив по формуле (5.16) значения s_j, $j = \overline{1, N}$, по формуле (5.17) получим $D(\vec{x}_n, \vec{y}_n) = 0{,}473$ и $\sqrt{\frac{mn}{m+n}} = 1{,}718$. Так как m и n велики, то для проверки гипотезы H_0 об однородности воспользуемся асимптотической формулой (5.13), в соответствии с которой

$$\mathbf{P}\left\{\sqrt{\frac{25 \cdot 28}{25+28}} D(\vec{x}_n, \vec{y}_n) > 1{,}718\right\} \approx 0{,}004.$$

Поэтому гипотезу об однородности следует отклонить на *уровне значимости* $\alpha \geqslant 0{,}004$.

5.16. Проверьте на уровне значимости $\alpha = 0{,}05$ при помощи критерия Колмогорова гипотезу о том, что выборка 2,1; −0,6; 0,2; 3,0; −1,0; 1,3 извлечена из распределения $N(1,1)$?

Ответ: данные не противоречат гипотезе.

5.17. Решите предыдущую задачу при помощи критерия ω^2.

Ответ: данные не противоречат гипотезе.

5.18. В экспериментах с селекцией гороха Г.И.Мендель* наблюдал частоты появления различных видов семян при скрещивании растений с круглыми желтыми семенами и растений с морщинистыми зелеными семенами. Эти данные и значения теоретических вероятностей по теории наследственности приведены в табл. 5.6. Проверьте на уровне значимости $\alpha = 0{,}1$ гипотезу H_0 о согласовании частотных данных с теоретическими вероятностями.

Таблица 5.6

Виды семян	Частота	Вероятность
Круглые и желтые	315	9/16
Морщинистые и желтые	101	3/16
Круглые и зеленые	108	3/16
Морщинистые и зеленые	32	1/16

Ответ: Гипотеза принимается.

5.19. Решите задачу 4.27, не предполагая нормальность распределения контролируемого признака.

5.20. В таблице 5.7 для каждой из девяти партий сыра приведены его жирность (в процентах) и усредненные (по 80 опрошенным респондентам) результаты опроса вкусовых качеств сыра по шестибальной системе („превосходно“ — 6 баллов, „очень хорошо“ — 5, „хорошо“ — 4, „так себе“ — 3, „плохо“ — 2, „неприемлемо“ — 1). Проверьте по результатам опроса гипотезу о связи жирности сыра и его вкусовых качеств на уровне значимости $\alpha = 0{,}05$.

Ответ: вкусовые качества сыра улучшаются с увеличением его жирности.

Таблица 5.7

Партия	Жирность, %	Результат опроса
1	44,4	2,6
2	45,9	3,1
3	41,9	2,5
4	53,3	5,0
5	44,7	3,6
6	44,1	4,0
7	50,7	5,2
8	45,2	2,8
9	60,1	3,8

5.21. Из 300 абитуриентов, поступивших в институт, 97 человек имели оценку 5 в школе и получили оценку 5 на вступительных экзаменах по тому же предмету, причем только 18 человек имели оценку 5 и в школе, и на экзамене. С уровнем значимости 0,1 проверьте гипотезу о независимости оценок 5 в школе и на экзамене.

О т в е т: гипотеза отклоняется.