Глава 6
Непараметрические методы математической статистики

6.1. Основные понятия

Классические методы решения основных задач математической статистики, например таких, как оценивание параметров распределения, проверка большинства статистических гипотез, корреляционный, регрессионный и дисперсионный анализ опираются на предположение о нормальности исследуемых генеральных совокупностей. Однако часто это предположение не выполняется. Например, в социологических «измерениях» результаты опросов могут иметь форму ответов «да» или «нет» и представляются в виде частот разных ответов. Традиционные методы математической статистики нельзя использовать для обработки таких данных. В подобных случаях используются *непараметрические методы*, т.е. методы, независящие от вида распределений генеральных совокупностей. При использовании этих методов может потребоваться лишь минимальная априорная информация об этих распределениях, например, предположение о непрерывности или симметрии функции распределения.

Непараметрические методы, как правило, используют не числовые значения элементов выборки, а её *структурные свойства*, например порядок этих элементов. При этом часть находящейся в выборке информации теряется. Но данный недостаток компенсируется ослаблением ограничений на вид закона распределения и упрощением вычислений.

6.2. Критерий знаков

Определение 6.1. Говорят, что две генеральные совокупности X и Y *однородны*, если их функции распределения совпадают: $F_X(x) \equiv F_Y(y)\big|_{y=x}$.

Определение 6.2. Выборки $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$ из генеральных совокупностей X и Y называются *попарно связанными*, если их элементы x_i и y_i при каждом i получены в одинаковых условиях. Например, если значения x_i и y_i - результаты измерения одного и того же параметра i-го объекта из n объектов двумя приборами.

Критерий знаков используется для проверки по попарно связанным выборкам $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$ гипотезы об однородности двух *непрерывных* генеральных совокупностей X и Y : $H_0 = \left\{F_X(x) \equiv F_Y(y)\big|_{y=x}\right\}$.

Если гипотеза H_0 верна, то совокупности X и Y взаимозаменяемы и для разностей элементов $X_i - Y_i$ случайных выборок $\{X_i\}$ и $\{Y_i\}$ положительные и отрицательные значения должны быть равновероятными. Вероятности нулевых значений $X_i - Y_i$ равны нулю в силу непрерывности совокупностей X и Y. Если же оказывается, что $x_i = y_i$, например, из-за ошибок округления, то пара (x_i, y_i) из рассмотрения исключается. В любом случае объём выборок будем считать равным n.

Итак, если генеральные совокупности X и Y однородны, то

$$P\{X_i - Y_i > 0\} = P\{X_i - Y_i < 0\} = \frac{1}{2},\ i = 1, 2, \ldots, n. \qquad (6.1)$$

С учётом (6.1) гипотезу H_0 об однородности X и Y можно сформулировать в виде

$$H_0 = \{p = 1/2\}, \text{ где } p = P\{X_i - Y_i > 0\} \tag{6.2}$$

при одной из альтернатив:

$$H_1^{(1)} = \{p > 1/2\},\ H_1^{(2)} = \{p < 1/2\},\ H_1^{(3)} = \{p \neq 1/2\}. \tag{6.3}$$

В качестве статистики $Z = Z\left(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}\right)$ критерия возьмём число знаков "+" в последовательности разностей $X_i - Y_i$. Поскольку элементы случайных выборок всегда считаются независимыми, знаки разностей $X_i - Y_i$ независимы, поэтому

$$Z \sim B(n, 1/2) \,|\, H_0. \tag{6.4}$$

Значит, критерий проверки H_0 на уровне значимости α можно сформулировать так. Пусть $Z_{\text{выб}} = r$ - число знаков "+" в последовательности $x_i - y_i$. Тогда гипотеза $H_0 = \{p = 1/2\}$ отклоняется, если:

а) $2^{-n}\sum_{i=r}^{n} C_n^i \le \alpha$ при альтернативе $H_1^{(1)} = \{p > 1/2\}$; (6.5)

б) $2^{-n}\sum_{i=0}^{r} C_n^i \le \alpha$ при альтернативе $H_1^{(2)} = \{p < 1/2\}$; (6.6)

в) $2^{-n}\sum_{i=0}^{r} C_n^i \le \frac{\alpha}{2}$ или $2^{-n}\sum_{i=r}^{n} C_n^i \le \frac{\alpha}{2}$ при альтернативе $H_1^{(3)} = \{p \neq 1/2\}$. (6.7)

Если неравенства (6.5) – (6.7) при соответствующих альтернативах не выполняются, гипотеза H_0 принимается как не противоречащая выборке на уровне значимости α.

Замечания. 1. Подсчёт вероятностей $2^{-n}\sum_{i=0}^{r} C_n^i$ и $2^{-n}\sum_{i=r}^{n} C_n^i$ при реализации данного критерия проверки может вызвать затруднения. Поэтому с учётом интегральной теоремы Муавра-Лапласа используют нормальную аппроксимацию для биномиального распределения (6.4), считая, что приближённо $Z \sim N\left(n/2, \sqrt{n}/2\right)$ или $\frac{Z - n/2}{\sqrt{n}/2} \sim N(0,1)$. Тогда

$2^{-n}\sum_{i=0}^{r} C_n^i = P\{Z \le r\} \approx \Phi\left(\frac{r - n/2}{\sqrt{n}/2}\right)$ или, более точно,

$$2^{-n}\sum_{i=0}^{r} C_n^i \approx \Phi\left(\frac{r - n/2 + 1/2}{\sqrt{n}/2}\right) \tag{6.8}$$

и

$$2^{-n}\sum_{i=r}^{n} C_n^i = P\{Z \ge r\} \approx 1 - \Phi\left(\frac{r - n/2}{\sqrt{n}/2}\right), \tag{6.9}$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ функция распределения стандартного нормального закона.

2. Часто более удобно использовать другие статистики, имеющие распределение Фишера. Тогда критерий проверки гипотезы H_0 на уровне значимости α принимает следующий вид. Гипотеза H_0 отклоняется, если

а) $Z_{1\,\text{выб}} = \frac{r}{n - r + 1} \ge F_{k_1, k_2, 1-\alpha}$, где $k_1 = 2(n - r + 1)$, $k_2 = 2r$ при $H_1 = H_1^{(1)}$; (6.10)

б) $Z_{2\,\text{выб}} = \dfrac{n-r}{r+1} \geq F_{k_1,k_2,1-\alpha}$, где $k_1 = 2(r+1)$, $k_2 = 2(n-r)$ при $H_1 = H_1^{(2)}$; (6.11)

в) если верно (6.10) или (6.11) при замене α на $\alpha/2$ для $H_1 = H_1^{(3)}$. (6.12)

Здесь $F_{k,l,p}$ - квантиль порядка *p* распределения Фишера $F(k,l)$.

Пример 6.1. Есть основания предполагать, что один из двух приборов для измерения ёмкости конденсатора имеет систематическую ошибку. Ёмкости 10 конденсаторов измерили обоими приборами, результаты измерений приведены ниже: в первой строке – показания первого, а во второй строке – второго прибора.

C_1, пф	55	52	95	75	80	65	54	63	85	70
C_2, пф	57	53	90	78	80	63	55	62	86	72

Можно ли на уровне значимости $\alpha = 0{,}1$ утверждать, что второй прибор даёт завышенные показания?

◀Проверяется основная гипотеза $H_0 = \{p = 1/2\}$ при альтернативе $H_1^{(2)} = \{p < 1/2\}$.

Применим критерий знаков. Запишем последовательность знаков разностей $C_1 - C_2$:

$$-,\ -,\ +,\ -,\ 0,\ +,\ -,+,\ -,\ -.$$

Число ненулевых разностей $n = 9$, число положительных разностей $r = 3$.

а) *Первый способ*. Учитывая (6.6) и (6.8), находим:

$$2^{-9}\sum_{i=0}^{3} C_9^i \approx \Phi\left(\frac{3-9/2+1/2}{\sqrt{9}/2}\right) = \Phi(-0{,}667) = 1-\Phi(0{,}667) = 0{,}2514 > \alpha = 0{,}1.$$

Таким образом, гипотеза H_0 принимается: различия в показаниях приборов вызваны случайными ошибками.

б) *Второй способ*. С учётом (6.11) получаем:

$$Z_{2\,\text{выб}} = \frac{9-3}{3+1} = 1{,}5 < F_{8,12,0.9} = 2{,}24,$$

т.е. гипотеза H_0 принимается.►

Критерий знаков применяется и для проверки *одновыборочной гипотезы о сдвиге*. Пусть закон распределения непрерывной генеральной совокупности X зависит от параметра θ и удовлетворяет следующим требованиям.

а) Для каждого значения параметра θ функция распределения $F_X(x,\theta)$ получается из функции $F_X(x,0)$ сдвигом по оси *Ox* на величину θ, т.е.

$$F_X(x,\theta) = F_X(x-\theta),$$

см. рис. 6.1. Как правило, θ является математическим ожиданием величины X, а в случае его отсутствия - медианой или модой.

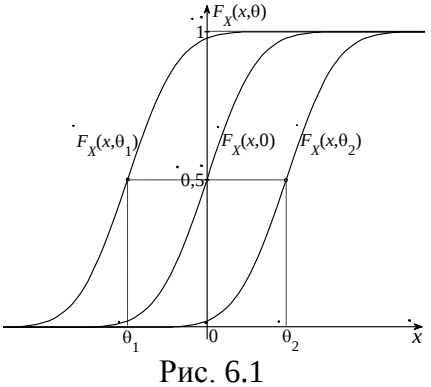

Рис. 6.1

б) Распределение X имеет единственную медиану, равную θ.

Проверяемая гипотеза относительно параметра θ формулируется так:

$$H_0 = \{\theta = \theta_0\}, \quad (6.13)$$

где θ_0 - заданное значение. В качестве альтернативы используется одна из гипотез

$$H_1^{(1)} = \{\theta > \theta_0\},\ H_1^{(2)} = \{\theta < \theta_0\},\ H_1^{(3)} = \{\theta \neq \theta_0\} \tag{6.14}$$

Критерий проверки основан на том, что если гипотеза H_0 верна, то с учётом условий а) – б) генеральная совокупность $X - \theta_0$ имеет медиану, равную нулю. Поэтому для случайной выборки $\{X_i - \theta_0\}$ имеем:

$$P\{X_i - \theta_0 > 0\} = P\{X_i - \theta_0 < 0\} = 1/2 \,|\, H_0,\ i = 1, 2, \ldots, n, \tag{6.15}$$

т.е. выполняются равенства, подобные (6.1).

Обозначим $p = P\{X_i - \theta_0 > 0\}$. Тогда с учётом (6.15) гипотезу H_0 из (6.13) можно сформулировать в виде (6.2), а альтернативы (6.14) – в форме (6.3). Таким образом, для проверки одновыборочной гипотезы о сдвиге применим критерий знаков, основанный на статистике $Z = Z(\mathbf{X}^{(n)})$, равной числу знаков "+" в последовательности разностей $X_i - \theta_0$.

Пример 6.2. По выборке
3.43, 3.38, 3.76, 3.79, 3.18, 3.48, 3.44, 3.64, 3.70, 3.75, 3.27, 3.67, 3.65, 3.16, 3.11, 3.49, 3.95, 3.34, 3.58, 3.22
объёма $n = 20$ на уровне значимости $\alpha = 0,05$ проверить гипотезу $H_0 = \{\theta = \theta_0 = 3,3\}$ против альтернативы $H_1^{(3)} = \{\theta \neq \theta_0\}$.

◄Перейдём к выборке $x_i - \theta_0 = x_i - 3,3$:
0.13, 0.08, 0.46, 0.49, -0.11, 0.18, 0.14, 0.34, 0.40, 0.45, -0.02, 0.37, 0.35, -0.13, -0.18, 0.19, 0.65, 0.04, 0.28, -0.07.

Число положительных значений в этой выборке $r = 15$.

а) *Первый способ*. С учётом (6.6) и (6.8), находим:

$$2^{-20}\sum_{i=0}^{15} C_{20}^i \approx \Phi\left(\frac{15 - 20/2 + 1/2}{\sqrt{20}/2}\right) = \Phi(2,46) = 0,9931 > \alpha/2 = 0,025;$$

$$2^{-20}\sum_{i=15}^{20} C_n^i \approx 1 - \Phi\left(\frac{15 - 20/2}{\sqrt{20}/2}\right) = 1 - \Phi(2,24) = 1 - 0,9875 = 0,0128 < \alpha/2 = 0,025.$$

С учётом последнего неравенства гипотезу $H_0 = \{\theta = 3,3\}$ отклоняем.

б) *Второй способ*. Согласно (6.10) – (6.12) находим: $k_1 = 2(n - r + 1) = 12$, $k_2 = 2r = 30$, поэтому $Z_{1\text{ выб}} = \dfrac{r}{n - r + 1} = 2,5 \geq F_{k_1, k_2, 1-\alpha/2} = F_{12,\,30,\,0.975} = 2,41$.

Отсюда следует, что гипотезу $H_0 = \{\theta = 3,3\}$ следует отклонить.►

6.3. Критерий серий Вальда-Вольфовица

Критерий серий применяется для проверки гипотезы H_0, утверждающей, что две выборки объёмами n_1 и n_2 извлечены из однородных генеральных совокупностей. Приведём алгоритм проверки.

1. Результаты измерений объединяются и строится вариационный ряд объединённой выборки.

2. Каждой выборке присваивается свой символ (*код*) (например, «1» - первой и «0» - второй выборке). По вариационному ряду строится *последовательность кодов* в соответствии с расположением элементов выборок в этом ряду.

3. Подсчитывается N число *серий* в последовательности кодов. Серией называется группа одинаковых кодов, ограниченная противоположными кодами с двух сторон или с одной стороны, если эта группа расположена в начале или в конце вариационного ряда.

4. Статистикой критерия является число N серий в вариационном ряду. Критическая область определяется неравенствами $N \le N_1$ и $N \ge N_2$, где для граничных значений существуют специальные таблицы. При больших объёмах выборок ($n_1 > 20$ и (или) $n_2 \ge 20$) для проверки гипотезы H_0 можно использовать статистику

$$Z = \frac{\left| N - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right) \right| - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2 \left(2n_1 n_2 - n_1 - n_2 \right)}{\left(n_1 + n_2 \right)^2 \left(n_1 + n_2 - 1 \right)}}}. \qquad (6.16)$$

Если гипотеза H_0 верна, то Z приближённо распределена по закону $N(0,1)$.

Пример 6.3. Однотипные микросхемы производятся на двух заводах и поставляются потребителю партиями по 1000 шт. Входному контролю подвергли 15 партий с первого и 21 партию – со второго завода. Количества забракованных схем в партиях приведены ниже.

1-й завод: 31, 26, 33, 11, 13, 5, 18, 1, 2, 16, 17, 23, 20, 21, 9;
2-й завод: 12, 7, 4, 8, 3, 6, 10, 25, 22, 24, 15, 19, 14, 36, 34, 32, 27, 29, 30, 35, 28.

Можно ли на уровне значимости $\alpha = 0{,}1$ считать, что качество продукции заводов существенно различается?

◄Проверяется гипотеза H_0: обе выборки получены из одной генеральной совокупности. Альтернативная гипотеза: H_1: выборки взяты из разных генеральных совокупностей.

Присвоим первой выборке код «1» и код «0» - второй группе. Последовательность кодов для вариационного ряда объединённой выборки имеет вид:

1 1 0 0 1 0 0 0 1 0 1 0 1 0 0 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0.

Число серий $N = 22$, объёмы выборок $n_1 = 15$ и $n_2 = 21$. Для проверки гипотезы используем статистику Z. Выборочное значение $Z_{\text{выб}} \approx 1{,}04$. В данном случае доверительная область $G_{1-\alpha} = \left(-u_{1-\alpha/2}; u_{1-\alpha/2} \right) = \left(-1{,}645; 1{,}645 \right)$. Поскольку $Z_{\text{выб}} \in G_{1-\alpha}$, значит гипотеза H_0 принимается: качество продукции на заводах не различается значимо.►

6.4. Критерий Вилкинсона-Манна-Уитни

Как и критерий серий, данный критерий применяется для проверки гипотезы H_0 о том, что две выборки объёмами n_1 и n_2 однородны или получены из одной генеральной совокупности.

Назовём *рангом* $R(a_k)$ элемента a_k числовой последовательности $\{a_i\}_{i=1}^{n}$ его порядковый номер в этой последовательности, т.е. $R(a_k) = k$.

Опишем алгоритм проверки гипотезы H_0.

1. Выборки объединяются, и строится вариационный ряд объединённой выборки.

2. Каждому элементу объединённой выборки ставится в соответствие его *ранг* – номер по порядку в вариационном ряду. Если несколько элементов ряда совпадают, то им присваиваются одинаковые ранги, равные среднему арифметическому их номеров.

3. Вычисляются суммы R_1 и R_2 рангов первой и второй выборок соответственно. Далее, находятся значения

$$w_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \text{ и } w_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{6.17}$$

(для правильно найденных значений w_1 и w_2 выполняется равенство $w_1 + w_2 = n_1 n_1$).
Выборочное значение статистики W критерия проверки находится по формуле

$$W_{\text{выб}} = \min\{w_1, w_2\}. \tag{6.18}$$

Критические значения для $W_{\text{выб}}$ находят из специальных таблиц. При объёмах выборок $n_1 > 8$ и $n_2 > 8$ для проверки гипотезы H_0 можно использовать статистику

$$Z = \frac{W - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}. \tag{6.19}$$

Если гипотеза H_0 верна, то Z приближённо распределена по закону $N(0,1)$.

Пример 6.4. Проверить гипотезу H_0 из примера 6.3 с помощью критерия Манна-Уитни на уровне значимости $\alpha = 0{,}1$.

◄Последовательность кодов вариационного ряда объединённой выборки найдена в примере 6.3:

1 1 0 0 1 0 0 0 1 0 1 0 1 0 0 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0.

Ей соответствует суммы рангов элементов выборок: $R_1 = 246$ и $R_2 = 420$.

Находим значения $w_1 = 189$ и $w_2 = 126$ из (6.17).

Отсюда видно, что $W_{\text{выб}} = \min\{w_1, w_2\} = 126$.

Так как $n_1 = 15 > 8$ и $n_2 = 21 > 8$, используем статистику Z из (6.19).

Вычисляем её выборочное значение: $Z_{\text{выб}} = -1{,}01$.

Поскольку $|Z_{\text{выб}}| = 1{,}01 < u_{1-\alpha/2} = u_{0{,}95} = 1{,}645$, гипотеза H_0, как и в примере 6.3, принимается.►

6.5. Непараметрический корреляционный анализ

Для проверки гипотезы о независимости двух генеральных совокупностей, имеющих нормальное распределение, используют *параметрический критерий*, основанный на выборочном коэффициенте корреляции, см. п. 3.1. Рассмотрим подобные (*непараметрические*) критерии, пригодные для совокупностей с произвольными законами распределения.

6.5.1. Критерий квадрантной корреляции

Этот упрощённый критерий позволяет по выборке $\{(x_i, y_i)\}_{i=1}^{n}$ проверить, имеется ли статистическая зависимость между генеральными совокупностями X и Y следующим образом.

1. Строим корреляционное поле, т.е. в системе координат Oxy отмечаем точки (x_i, y_i).

2. По выборкам $\{x_i\}$ и $\{y_i\}$ находим выборочные медианы $\tilde{h}_x$ и $\tilde{h}_y$. Напомним, что для нечётного объёма выборки $n = 2k+1$ выборочная медиана равна значению элемента

вариационного ряда выборки с номером $k+1$, а при $n=2k$ - полусумме значений элементов этого ряда с номерами k и $k+1$. Далее, прямыми $x=\tilde{h}_x$ и $y=\tilde{h}_y$ делим плоскость *Oxy* на 4 квадранта. При чётном n по обе стороны от каждой из этих прямых будет расположено по одинаковому числу точек. Если же n нечётно, то исключаем точку (x_i, y_i), через которую проходит прямая $y=\tilde{h}_y$, и получаем в результате чётное n.

3. Подсчитываем количества точек (x_i, y_i) в отдельных квадрантах (эти количества одинаковы в первом и третьем, а также во втором и четвёртом квадрантах).

4. Если меньшее или бóльшее число точек в квадрантах достигает, соответственно, нижней или верхней границы, указанных в специальных таблицах нижних и верхних критических значений квадрантной корреляции (см. табл. 6.1), или выходит за пределы этих границ, то гипотеза $H_0=\{\rho[X,Y]=0\}$ отклоняется в пользу альтернативы $H_1=\{\rho[X,Y]\neq 0\}$.

Таблица 6.1.

n	Критическое число точек				n	Критическое число точек			
	Нижнее		Верхнее			Нижнее		Верхнее	
	Уровень значимости α					Уровень значимости α			
	0,05	0,01	0,05	0,01		0,05	0,01	0,05	0,01
8–9	0	-	4	-	74–75	13	12	24	25
10–11	0	0	5	5	76–77	14	12	24	26
12–13	0	0	6	6	78–79	14	13	25	26
14–15	1	0	6	7	80–81	15	13	25	27
16–17	1	0	7	8	82–83	15	14	26	27
18–19	1	1	8	8	84–85	16	14	26	28
20–21	2	1	8	9	86–87	16	15	27	28
22–23	2	2	9	9	88–89	16	15	28	29
24–25	3	2	9	10	90–91	17	15	28	30
26–27	3	2	10	11	92–93	17	16	29	30
28–29	3	3	11	11	94–95	18	16	29	31
30–31	4	3	11	12	96–97	18	17	30	31
32–33	4	3	12	13	98–99	19	17	30	32
34–35	5	4	12	13	100–101	19	18	31	32
36–37	5	4	13	14	110–111	21	20	34	35
38–39	6	5	13	14	120–121	24	22	36	38
40–41	6	5	14	15	130–131	26	24	39	41
42–43	6	5	15	16	140–141	28	26	42	44
44–45	7	6	15	16	150–151	31	29	44	46
46–47	7	6	16	17	160–161	33	31	47	49
48–49	8	7	16	17	170–171	35	33	50	52
50–51	8	7	17	18	180–181	37	35	53	55
52–53	8	7	18	19	200–201	42	40	58	60
54–55	9	8	18	19	220–221	47	44	63	66
56–57	9	8	19	20	240–241	51	49	69	71
58–59	10	9	19	20	260–261	56	54	74	76

60–61	10	9	20	21	280–281	61	58	79	82
62–63	11	9	20	22	300–301	66	63	84	87
64–65	11	10	21	22	320–321	70	67	90	93
66–67	12	10	21	23	340–341	75	72	95	98
68–69	12	11	22	23	360–361	80	77	100	103
70–71	12	11	23	24	380–381	84	81	106	109
72–73	13	12	23	24	400–401	89	86	111	114

Пример 6.5. Корреляционное поле выборки $\{(x_i, y_i)\}_{i=1}^{n}$ показано на рис. 6.2. На уровне значимости $\alpha = 0,05$ проверить гипотезу $H_0 = \{\rho[X,Y] = 0\}$ при двухсторонней альтернативе $H_1 = \{\rho[X,Y] \neq 0\}$.

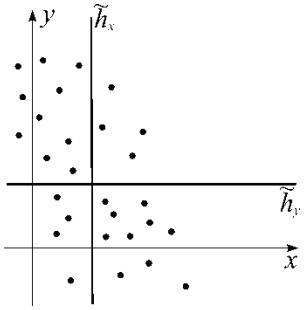

Рис. 6.2

◄В данном случае $n = 28$, минимальное и максимальное количества точек в квадрантах равны, соответственно, 4 и 10. Нижнее и верхнее критические значения из табл. 6.1 для $n = 28$ и $\alpha = 0,05$ составляют, соответственно, 3 и 11. Поэтому гипотеза H_0 принимается: значимая статистическая связь между X и Y не подтвердилась.►

6.5.2. Критерий Спирмена независимости двух генеральных совокупностей

Пусть $\{(x_i, y_i)\}$ - выборка объёма *n* из двумерной генеральной совокупности (X,Y). Поставим в соответствие каждому элементу (x_i, y_i) этой выборки пару (t_i, s_i) рангов $t_i = R(x^{(i)})$ и $s_i = R(y^{(i)})$ элементов x_i и y_i в вариационных рядах $\{(x^{(i)})\}$ и $\{(y^{(i)})\}$. Выборочный коэффициент корреляции (3.1), найденный по выборке $\{(t_i, s_i)\}$, т.е.

$$r_s = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2 \sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

называется *выборочным коэффициентом ранговой корреляции Спирмена* ρ_s.

Очевидно, $\bar{t} = \bar{s} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}$ и можно показать, что

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(t_i - s_i)^2 . \qquad (6.20)$$

Без ограничения общности можно считать, что пары (x_i, y_i) пронумерованы в порядке возрастания чисел x_i. Тогда $t_i = i$, поэтому формула для коэффициента Спирмена r_s из (6.20) принимает вид

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(i - s_i)^2 . \qquad (6.21)$$

Замечания. 1. Если несколько значений x_i или (и) y_i совпадают, то им ставится в соответствие одинаковые ранги, равные среднему арифметическому порядковых номеров

этих значений в вариационном ряду.

2. Выборочный коэффициент корреляции Спирмена, как и обыкновенный выборочный коэффициент корреляции (3.1) обладает свойством $-1 \le r_s \le 1$. При этом, если между X и Y имеется произвольная строго монотонная функциональная зависимость (*не обязательно линейная*), то $r_s = 1$ для возрастающей и $r_s = -1$ для убывающей зависимости.

Рассмотрим гипотезу об отсутствии корреляции между совокупностями X и Y

$$H_0 = \{\rho(X,Y) = 0\}$$

при одной из альтернатив:

$$H_1^{(1)} = \{\rho(X,Y) > 0\},\ H_1^{(2)} = \{\rho(X,Y) < 0\},\ H_1^{(3)} = \{\rho(X,Y) \neq 0\}.$$

Для проверки гипотезы H_0 используются таблицы критических значений ранговой корреляции Спирмена.

Таблица 6.2

n	α $\left(H_1 = \{\rho > 0\}\right)$			n	α $\left(H_1 = \{\rho > 0\}\right)$		
	0.10	0.05	0.01		0.10	0.05	0.01
4	1.000	1.000	-	34	0.225	0.287	0.399
5	0.800	0.900	1.000	35	0.222	0.283	0.394
6	0.657	0.829	0.943	36	0.219	0.279	0.388
7	0.571	0.714	0.893	37	0.216	0.275	0.383
8	0.524	0.643	0.833	38	0.212	0.271	0.378
9	0.483	0.600	0.783	39	0.210	0.267	0.373
10	0.455	0.564	0.745	40	0.207	0.264	0.368
11	0.427	0.536	0.709	41	0.204	0.261	0.364
12	0.406	0.503	0.671	42	0.202	0.257	0.359
13	0.385	0.484	0.648	43	0.199	0.254	0.355
14	0.367	0.464	0.622	44	0.197	0.251	0.351
15	0.354	0.443	0.604	45	0.194	0.248	0.347
16	0.341	0.429	0.582	46	0.192	0.246	0.343
17	0.328	0.414	0.566	47	0.190	0.243	0.340
18	0.317	0.401	0.550	48	0.188	0.240	0.336
19	0.309	0.391	0.535	49	0.186	0.238	0.333
20	0.299	0.380	0.520	50	0.184	0.235	0.329
21	0.292	0.370	0.508	52	0.180	0.231	0.323
22	0.284	0.361	0.496	54	0.177	0.226	0.317
23	0.278	0.353	0.486	56	0.174	0.222	0.311
24	0.271	0.344	0.475	58	0.171	0.218	0.306
25	0.265	0.337	0.466	60	0.168	0.214	0.300
26	0.259	0.331	0.457	62	0.165	0.211	0.296
27	0.255	0.324	0.448	64	0.162	0.207	0.291
28	0.250	0.317	0.440	66	0.160	0.204	0.287
29	0.245	0.312	0.433	68	0.157	0.201	0.282
30	0.240	0.306	0.425	70	0.155	0.198	0.278
31	0.236	0.301	0.418	72	0.153	0.195	0.274
32	0.232	0.296	0.412	74	0.151	0.193	0.271
33	0.229	0.291	0.405	76	0.149	0.190	0.267

Например, в табл. 6.2 для различных объемов выборок приведены *верхние* критические значения $\rho_s(\alpha,n)$ коэффициента ранговой корреляции Спирмена ρ_s, используемые при проверке гипотезы H_0 против правосторонней альтернативы $H_1^{(1)}$ (о положительной корреляции). Если $r_s \ge \rho_s(\alpha,n)$, то H_0 *отклоняется* в пользу $H_1^{(1)}$.

Нижние критические значения (для проверки H_0 против левосторонней альтернативы $H_1^{(2)}$) равны $-\rho_s(\alpha,n)$, т.е. при $r_s \le -\rho_s(\alpha,n)$ гипотеза H_0 *отклоняется* в пользу $H_1^{(2)}$.

Если гипотеза H_0 проверяется против двусторонней альтернативы (т.е. против предположения о какой-либо связи между X и Y), то величина $\rho_s(\alpha,n)$ служит критическим значением уровня значимости 2α для статистики $|\rho_s|$. Это значит, что H_0 *отклоняется* в пользу $H_1^{(3)}$ на уровне значимости α, если $|r_s| \ge \rho_s(\alpha/2,n)$.

При $n \ge 10$ можно приближённо считать, что если H_0 верна, то следующая статистика Z имеет распределение Стьюдента:

$$Z = \frac{r_s\left(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}\right)}{\sqrt{1-r_s^2\left(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}\right)}}\sqrt{n-2} \sim St(n-2)\Big| H_0 . \qquad (6.22)$$

Поэтому возможно использование такого критерия *принятия* гипотезы H_0:

$$Z_{\text{выб}} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \in G_{1-\alpha}^{(1)} = \left(-\infty; t_{n-2,\,1-\alpha}\right) \text{ при } H_1 = H_1^{(1)} = \left\{\rho(X,Y) > 0\right\};$$

$$Z_{\text{выб}} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \in G_{1-\alpha}^{(2)} = \left(-t_{n-2,\,1-\alpha}; +\infty\right) \text{ при } H_1 = H_1^{(2)} = \left\{\rho(X,Y) < 0\right\};$$

$$Z_{\text{выб}} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \in G_{1-\alpha}^{(3)} = \left(-t_{n-2,\,1-\alpha/2}; t_{n-2,\,1-\alpha/2}\right) \text{ или } |Z_{\text{выб}}| < t_{n-2,\,1-\alpha/2},$$

где $t_{k,\,p}$ квантиль порядка p распределения $St(n-2)$.

Пример 6.6. Используя коэффициент ранговой корреляции Спирмена, проверить значимость корреляции для выборки

x	68,8	63,3	75,5	67,2	71,3	72,8	76,5	63,5	69,9	71,4
y	167	113	160	154	151	181	173	115	126	166

на уровне $\alpha = 0{,}1$ при альтернативах:

а) $H_1^{(1)} = \left\{\rho(X,Y) > 0\right\}$; б) $H_1^{(3)} = \left\{\rho(X,Y) \ne 0\right\}$.

◀Перепишем выборку, упорядочив её по неубыванию элементов x_i:

x	63,3	63,5	67,2	68,8	69,9	71,3	71,4	72,8	75,5	76,5
y	113	115	154	167	126	151	166	181	160	173

Найдём ранги для значений y_i. Для этого составим вариационный ряд $\left\{y^{(i)}\right\}$:

i	1	2	3	4	5	6	7	8	9	10
$y^{(i)}$	113	115	126	151	154	160	166	167	173	181

Таким образом, упорядоченной по элементам x_i выборке соответствует такая последовательность пар рангов и квадратов их разностей

t_i	1	2	3	4	5	6	7	8	9	10
s_i	1	2	5	8	3	4	7	10	6	9
$(t_i - s_i)^2$	0	0	4	16	4	4	0	4	9	1

Вычисляем выборочное значение коэффициента корреляции Спирмена (6.20)

$$r_s = 0,745 .$$

Далее, проверяем гипотезу H_0 двумя способами.

а) $H_1 = H_1^{(1)} = \{\rho(X,Y) > 0\}$.

Первый способ. Из табл. 6.2 находим $\rho_s(n,\alpha)$ для $n=10$, $\alpha = 0,1$: $\rho_s(10, 0.1) = 0,455$. Поскольку $r_s = 0,745 \ge \rho_s(10, 0.1) = 0,455$, гипотеза H_0 *отклоняется*: *корреляция X и Y значима*.

Второй способ. Выборочное значение статистики (6.22)

$$Z_{\text{выб}} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{0,745\cdot\sqrt{8}}{\sqrt{1-0,745^2}} \approx 3,159 ,$$

квантиль $t_{n-2,\,1-\alpha} = t_{8,\,0.9} = 1,397$, т.е. $Z_{\text{выб}} > t_{n-2,\,1-\alpha}$. Таким образом, гипотеза H_0 и на этот раз *отклоняется*.

б) $H_1 = H_1^{(3)} = \{\rho(X,Y) \ne 0\}$.

Первый способ. Из табл. 6.2 находим $\rho_s(n,\alpha/2)$ для $n=10$, $\alpha = 0,1$: $\rho_s(10, 0.05) = 0,564$. Поскольку $r_s = 0,745 \ge \rho_s(10, 0.05) = 0,564$, гипотеза H_0 *отклоняется*: *корреляция X и Y значима*.

Второй способ. Выборочное значение статистики (6.22) $Z_{\text{выб}} \approx 3,159$, квантиль $t_{n-2,\,1-\alpha/2} = t_{8,\,0.95} = 1,860$, т.е. $|Z_{\text{выб}}| > t_{n-2,\,1-\alpha/2}$. Таким образом, гипотеза H_0 и на этот раз *отклоняется*.►

6.5.3. Критерий Кендэлла независимости двух генеральных совокупностей

Наряду с коэффициентом корреляции Спирмена для обнаружения связи между генеральными совокупностями X и Y используется также *выборочный коэффициент ранговой корреляции Кендэлла* τ:

$$\tau = 1 - \frac{4k}{n(n-1)},$$

где k – число *инверсий* (нарушений порядка) в ряду рангов последовательности $\{y_i\}$ после упорядочивания чисел $\{x_i\}$.

Для критических значений $\rho_\tau(\alpha, n)$ коэффициента τ имеются таблицы, см., например, табл. 6.3. С их помощью проверка значимости коэффициента ранговой корреляции Кендэлла при различных альтернативных гипотезах проводится в точности так же, как это описано в п. 6.5.2 для коэффициента ранговой корреляции Спирмена.

При больших объёмах выборки *n* можно приближённо считать, что если H_0 верна, то следующая статистика Z имеет стандартное нормальное распределение:

$$Z=\sqrt{\frac{9n(n+1)}{2(2n+5)}}\tau\left(\mathbf{X}^{(n)},\mathbf{Y}^{(n)}\right)\sim N(0,1)\Big|H_0. \qquad (6.23)$$

Таблица 6.3

n	$\alpha\ \ (H_1=\{\rho>0\})$			*n*	$\alpha\ \ (H_1=\{\rho>0\})$		
	0.10	0.05	0.01		0.10	0.05	0.01
5	0.600	0.800	1.000	23	0.194	0.249	0.352
6	0.467	0.600	0.867	24	0.188	0.246	0.341
7	0.429	0.619	0.810	25	0.187	0.240	0.333
8	0.429	0.500	0.714	26	0.182	0.231	0.323
9	0.389	0.444	0.611	27	0.179	0.225	0.316
10	0.333	0.422	0.600	28	0.175	0.222	0.312
11	0.309	0.418	0.564	29	0.172	0.222	0.305
12	0.303	0.394	0.515	30	0.168	0.214	0.301
13	0.282	0.359	0.487	31	0.166	0.211	0.295
14	0.275	0.341	0.473	32	0.161	0.206	0.290
15	0.257	0.333	0.448	33	0.159	0.205	0.284
16	0.250	0.317	0.433	34	0.155	0.201	0.280
17	0.235	0.309	0.426	35	0.153	0.197	0.277
18	0.229	0.294	0.412	36	0.152	0.194	0.273
19	0.216	0.287	0.392	37	0.150	0.189	0.267
20	0.211	0.274	0.379	38	0.147	0.189	0.263
21	0.210	0.267	0.371	39	0.144	0.185	0.258
22	0.203	0.255	0.359	40	0.144	0.182	0.256

Пример 6.7. Решить пример 6.6 с использованием коэффициента ранговой корреляции Кендэлла.

◄В примере 6.6 найдены последовательности рангов для элементов выборки:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	1	2	5	8	3	4	7	10	6	9

Найдём число инверсий в последовательности чисел в нижней строке:

а) числа 1 и 2 инверсий не образует; б) число 5 образует 2 инверсии - с числами 3 и 4; в) число 8 образует 4 инверсии - с числами 3, 4, 7 и 6; г) число 7 образует 1 инверсию; д) число 10 образует 2 инверсии.

Таким образом, общее число инверсий $k=9$ и выборочный коэффициент ранговой корреляции Кендэлла равен:

$$\tau=1-\frac{4\cdot 9}{10(10-1)}=0,6.$$

Далее, проверяем гипотезу $H_0=\{\rho(X,Y)=0\}$ двумя способами.

а) $H_1=H_1^{(1)}=\{\rho(X,Y)>0\}$.

Первый способ. Из табл. 6.3 находим $\rho_\tau(\alpha, n)$ для $n=10$, $\alpha=0{,}1$: $\rho(10,\,0.1)=0{,}333$. Поскольку $\tau=0{,}6\geq\rho(10,\,0.1)=0{,}333$, гипотеза H_0 *отклоняется*: *корреляция X и Y значима*.

Второй способ. Выборочное значение статистики (6.23)

$$Z_{\text{выб}}=\sqrt{\frac{9n(n+1)}{2(2n+5)}}\tau=\sqrt{\frac{9\cdot10\cdot11}{2\cdot(20+5)}}\cdot0{,}6=2{,}67,$$

квантиль $u_{1-\alpha}=u_{0.9}=\ 1{,}282$, т.е. $Z_{\text{выб}}>u_{1-\alpha}$. Таким образом, согласно (6.24), гипотеза H_0 и на этот раз *отклоняется*.

б) $H_1=H_1^{(3)}=\{\rho(X,Y)\neq0\}$.

Первый способ. Из табл. 6.3 находим $\rho_\tau(n,\alpha/2)$ для $n=10$, $\alpha=0{,}1$: $\rho_\tau(10,\,0.05)=0{,}422$. Поскольку $\tau=0{,}6\geq\rho_\tau(10,\,0.05)=0{,}422$, гипотеза H_0 *отклоняется*: *корреляция X и Y значима*.

Второй способ. Выборочное значение статистики (6.23) $Z_{\text{выб}}=2{,}67$, квантиль $u_{1-\alpha/2}=u_{0.95}=\ 1{,}645$, т.е. $|Z_{\text{выб}}|\geq u_{1-\alpha/2}$. Таким образом, гипотеза H_0 и на этот раз *отклоняется*.►