

ФУНКЦИИ РЕГРЕССИИ НАБЛЮДЕНИЙ И ИХ РАНГОВ

Рассматриваются характеристики связи между исходными наблюдениями и их рангами. Приведены формулы для вычисления функций регрессии и коэффициента корреляции между исходными наблюдениями и их рангами. Получены приближенные формулы вычисления математических ожиданий порядковых статистик для широких классов распределений.

Пусть X случайная величина (с.в.) с функцией распределения (ф.р.) вероятностей F_X . На практике обычно ф.р. частично либо полностью неизвестна, в связи с чем возникают статистические задачи, связанные с уточнением статистической модели $(X, \mathfrak{Z})$, где X – выборочное пространство и $\mathfrak{Z} = \{F_X\}$ – множество допустимых ф.р. в условиях данного статистического эксперимента. Основой для решения статистических задач служит информация, содержащаяся в выборке $X_1, \dots, X_n$, полученной в результате n повторных и независимых наблюдений над с.в. X . Отметим, что эта информация может быть разделена на две части. Одна из них содержится в упорядоченной статистике $\vec{X}_{(.)} = (X_{(1)}, \dots, X_{(n)})$, а другая – в рангах $R_1, \dots, R_n$ наблюдений $X_1, \dots, X_n$. Статистические процедуры, построенные в рамках параметрических моделей, обычно основаны на исходных наблюдениях $X_1, \dots, X_n$, а статистические процедуры непараметрической и робастной статистики (см. например [1] и [2]), основаны либо на порядковых статистиках, либо на рангах наблюдений. Для описания свойств таких процедур в литературе (см. например [3] и [4]) подробно описаны статистические характеристики порядковых статистик и рангов наблюдений. В частности, в работе [3] показано, что если совместная плотность $f_{X_1, \dots, X_n}(x_1, \dots, x_n)$ случайных величин $X_1, \dots, X_n$ является симметричной функцией своих аргументов, то упорядоченная статистика $\vec{X}_{(.)} = (X_{(1)}, \dots, X_{(n)})$ и ранговый вектор $\vec{R} = (R_1, \dots, R_n)$ независимы. В продолжение работы авторов [5] в настоящей работе изучаются вопросы статистической связи между исходными наблюдениями и их рангами. Получены выражения для функций регрессии наблюдений и их рангов, а также предложены приближенные формулы вычисления этих функций и коэффициента корреляции для достаточно широкого класса распределений $\mathfrak{Z}_\lambda$, квантильные функции которых записываются с помощью λ -аппроксимации Тьюки [6].

Пусть $\vec{X} = (X_1, \dots, X_n)$ выборка из распределения $F_X(x)$ с плотностью $f_X(x)$, $x \in R^1$, то есть предполагаем, что $X_1, \dots, X_n$ – н.о.р. случайные величины с ф.р. F_X . Пусть далее, $\vec{X}_{(.)} = (X_{(1)}, \dots, X_{(n)})$ – упорядоченная статистика и $\vec{R} = (R_1, \dots, R_n)$ – ранговый вектор для выборки $\vec{X} = (X_1, \dots, X_n)$. Отметим, что по выборке однозначно определяются упорядоченная статистика и ранговый вектор и обратно – знание порядковых статистик и рангов наблюдений позволяет однозначно восстановить исходные наблюдения. Таким образом, меж-

ду исходной выборкой $\vec{X}$ и парой $\{\vec{X}_{(.)}, \vec{R}\}$ существует взаимно однозначное соответствие, то есть

$$(X_1, \dots, X_n) \Leftrightarrow \{(X_{(1)}, \dots, X_{(n)}), (R_1, \dots, R_n)\}. \quad (1)$$

Теорема 1. Пусть $X_1, \dots, X_n$ – н.о.р. случайные величины с ф.р. F_X и плотностью $f_X(x)$, $x \in R^1$. Пусть далее R_i ранг наблюдения X_i , $i = 1, \dots, n$ и пусть $F_{X_i R_i}(x, y)$ – совместная функция распределения, а $F_{X_i}(x)$ и $F_{R_i}(y)$ – маргинальные ф.р. случайных величин X_i и R_i , $i = 1, \dots, n$. Тогда $F_{X_i R_i}(x, y) \neq F_{X_i}(x) \cdot F_{R_i}(y)$, то есть случайные величины X_i и R_i , $i = 1, \dots, n$ – являются зависимыми случайными величинами.

Доказательство. Поскольку $X_1, \dots, X_n$ – н.о.р. случайные величины, то реализация $x_1, \dots, x_n$ рангов $R_1, \dots, R_n$ наблюдений $X_1, \dots, X_n$ является перестановкой чисел $1, 2, \dots, n$ и следовательно, ранг R_i наблюдения X_i , $i = 1, \dots, n$, является дискретной случайной величиной с возможными значениями $1, 2, \dots, n$ и их вероятностями

$$P\{R_i = j\} = 1/n, \quad 1 \leq i, j \leq n. \quad (2)$$

Функция распределения вероятностей случайной величины R_i для $\forall i \in (1, \dots, n)$ записывается в виде

$$F_{R_i}(y) = n^{-1} \sum_{j=1}^n C(y-j), \quad y \in R^1, \quad (3)$$

где $C(x-x_0)$ – единичная функция Хевисайда, определяемая в виде

$$C(x-x_0) = \begin{cases} 1, & x \geq x_0, \\ 0, & x < x_0. \end{cases} \quad (4)$$

Далее, согласно определению ранга наблюдения, выполняются равенства $X_i = {}^d X_{(R_i)}$, $i = 1, \dots, n$, из которых следует, что условная плотность наблюдения X_i , при условии, что ранг этого наблюдения $R_i = j$, $1 \leq i, j \leq n$, совпадает с плотностью j -й порядковой статистики, то есть для $x \in R^1$

$$\begin{aligned} f_{X_i|R_i=j}(x|j) &= f_{X_{(j)}}(x) = \\ &= n C_{n-1}^{j-1} F_X^{j-1}(x) (1 - F_X(x))^{n-j} f_X(x), \quad 1 \leq i, j \leq n. \end{aligned} \quad (5)$$

Учитывая теперь, что (X_i, R_i) , $i = 1, \dots, n$, являются двумерными случайными величинами смешанного типа (см. [4]), где первая компонента X_i является непрерывной с.в. с ф.р. F_X , а вторая компонента R_i является дискретной с.в. с ф.р. $F_{R_i}(y)$, определенной

в (3), запишем выражение для совместной ф.р. случайных величин X_i и R_i , $i = 1, \dots, n$. Используя (3) и (5), получим

$$\begin{aligned} F_{X_i R_i}(x, y) &= F_{R_i}(y) \cdot \int_{-\infty}^x f_{X_i | R_i=j}(x|j) dx = \\ &= n^{-1} \sum_{j=1}^n C(y-j) F_{X_{(j)}}(x). \end{aligned} \quad (6)$$

Отметим, что из (6) следуют выражения для маргинальных ф.р. случайных величин X_i и R_i , $i = 1, \dots, n$. В самом деле, $F_{X_i R_i}(x, +\infty) = F_{R_i}(y)$ и

$$\begin{aligned} F_{X_i}(x) &= F_{X_i R_i}(x, +\infty) = n^{-1} \sum_{j=1}^n F_{X_{(j)}}(x) = \\ &= n^{-1} \sum_{j=1}^n n C_{n-1}^{j-1} \int_0^{F_X(x)} u^{j-1} (1-u)^{n-j} du = \\ &= \int_0^{F_X(x)} \left(\sum_{j=1}^n C_{n-1}^{j-1} u^{j-1} (1-u)^{n-j} \right) du = \int_0^{F_X(x)} du = F_X(x), \\ &i = 1, \dots, n, \quad x \in R^1. \end{aligned} \quad (7)$$

Окончательно из формулы (6) следует, что совместная ф.р. $F_{X_i R_i}(x, y)$ может быть представлена в виде произведения маргинальных ф.р. $F_{X_i}(x)$ и $F_{R_i}(y)$ тогда и только тогда, когда выполняется равенство $F_{X_{(j)}}(x) = F_X(x)$, $\forall x \in R^1$. Однако, это равенство не имеет места и, следовательно, $F_{X_i R_i}(x, y) \neq F_{X_i}(x) \cdot F_{R_i}(y)$ и это означает, что случайные величины X_i и R_i , $i = 1, \dots, n$, зависимы. Доказательство завершено.

Замечание 1. Использование понятия δ -функции Дирака, позволяет определить формальное выражение для совместной плотности двумерных случайных величин (X_i, R_i) , $i = 1, \dots, n$ смешанного типа в виде

$$f_{X_i R_i}(x, y) = n^{-1} \sum_{j=1}^n f_{X_{(j)}}(x) \delta(y-j). \quad (8)$$

Отсюда, в частности, следует формальное выражение для условной плотности дискретной случайной величины R_i , $i = 1, \dots, n$, в виде

$$\begin{aligned} f_{R_i | X_i=x}(y|x) &= n^{-1} \sum_{j=1}^n [f_{X_{(j)}}(x) / f_X(x)] \delta(y-j), \\ &1 \leq i \leq n. \end{aligned} \quad (9)$$

Приведенная теорема 1 устанавливает факт статистической связи между случайными величинами X_i и R_i , $i = 1, \dots, n$. Обсудим теперь характер статистической связи между этими случайными величинами, который определяется распределениями (2), (5) и (6). Используя эти распределения, с учетом замечания 1, получим выражения для функций регрессии наблюдения X_i на его ранг R_i , обозначим их через $M(X_i | R_i = j)$, $1 \leq i, j \leq n$, и функций регрессии ранга

R_i на наблюдение X_i , обозначим их через $M(R_i | X_i = x)$, $1 \leq i \leq n$, $x \in R^1$. Справедлива следующая теорема.

Теорема 2. Пусть $X_1, \dots, X_n$ н.о.р. случайные величины с ф.р. F_X и плотностью $f_X(x)$, $x \in R^1$. Пусть далее R_i – ранг наблюдения X_i , $i = 1, \dots, n$. Тогда функция регрессии ранга R_i на наблюдение X_i , $\forall i \in (1, \dots, n)$, имеет вид

$$M(R_i | X_i = x) = 1 + (n-1)F_X(x), \quad x \in R^1, \quad (10)$$

а функция регрессии наблюдения X_i на его ранг R_i , $\forall i \in (1, \dots, n)$ определена лишь в дискретном наборе точек с координатами $(j, M(X_{(j)}))$ и равна

$$M(X_i | R_i = j) = M(X_{(j)}), \quad 1 \leq i, j \leq n. \quad (11)$$

Доказательство. Используя (9), получаем функцию регрессии ранга R_i на наблюдение X_i , $\forall i \in (1, \dots, n)$, в виде

$$\begin{aligned} M[R_i | X_i = x] &= \int_{-\infty}^{+\infty} y f_{R_i | X_i=x}(y|x) dy = \\ &= \int_{-\infty}^{+\infty} y \left[\sum_{j=1}^n C_{n-1}^{j-1} F_X^{j-1}(x) (1-F_X(x))^{n-j} \right] \delta(y-j) dy = \\ &= \sum_{j=1}^n C_{n-1}^{j-1} F_X^{j-1}(x) (1-F_X(x))^{n-j} \int_{-\infty}^{+\infty} y \delta(y-j) dy = \\ &= \sum_{j=1}^n j C_{n-1}^{j-1} F_X^{j-1}(x) (1-F_X(x))^{n-j} = \\ &= \sum_{k=0}^{n-1} (k+1) C_{n-1}^k F_X^k(x) (1-F_X(x))^{n-1-k} = \\ &= 1 + (n-1)F_X(x), \quad x \in R^1. \end{aligned}$$

Отметим, что эта функция не зависит от индекса i , что является следствием независимости и одинаковой распределенности случайных величин $X_1, \dots, X_n$.

Далее, согласно определению, функция регрессии наблюдения X_i на его ранг R_i , с учетом (5), $\forall i \in (1, \dots, n)$, запишется в виде

$$\begin{aligned} M(X_i | R_i = j) &= \int_{-\infty}^{+\infty} x f_{X_i | R_i=j}(x|j) dx = \\ &= \int_{-\infty}^{+\infty} x f_{X_{(j)}}(x) dx = M(X_{(j)}), \quad 1 \leq i, j \leq n. \end{aligned}$$

Отметим, что эта функция также не зависит от индекса i и определена лишь для $j = 1, \dots, n$. Доказательство завершено.

Замечание 2. Вычисление математических ожиданий порядковых статистик в общем случае для произвольной ф.р. F_X вызывает аналитические трудности. В связи с этим составлены таблицы математических ожиданий порядковых статистик для различных объемов выборки n (обычно для $n \leq 20$) и для раз-

личных, наиболее употребительных на практике, распределений вероятностей. Наиболее полные таблицы приведены в [7]. В тех случаях, когда таблицы недоступны, или ими нельзя воспользоваться из-за их ограниченности, используют различные аппроксимационные формулы. В частности, для этой цели может быть использована λ -аппроксимация Тьюки для обратной функции F_X^{-1} ф.р. F_X , которая записывается в виде

$$\mathfrak{F}_\lambda = \{ F_X : F_X^{-1}(u) \approx \lambda_1 + \lambda_2^{-1} [u^{\lambda_3} - (1-u)^{\lambda_4}] \}, \quad (12)$$

где $\lambda_1, \dots, \lambda_4$ заданные постоянные. Методы вычисления этих постоянных, обеспечивающие нужную для практики точность при аппроксимации обратной функции $F_X^{-1}(u)$, $0 \leq u \leq 1$, приведены в [6]. Использование (12) позволяет записать приближенную формулу для вычисления $M(X_i | R_i = j)$, $1 \leq i, j \leq n$ в виде

$$\begin{aligned} M(X_i | R_i = j) &= M(X_{(j)}) = \\ &= n C_{n-1}^{j-1} \int_0^1 F_X^{-1}(u) u^{j-1} (1-u)^{n-j} du \approx \\ &\approx n C_{n-1}^{j-1} \left[\lambda_1 B(j, n) + \lambda_2^{-1} B(\lambda_3 + j, n - j + 1) - \right. \\ &\quad \left. - \lambda_2^{-1} B(j, \lambda_4 + n - j + 1) \right], \\ &1 \leq i, j \leq n, \end{aligned} \quad (13)$$

где $B(m, n)$ – бета функция.

Пример. Приведем вычисление функций регрессии для гауссовского случая, то есть для $F_X(x) = \Phi(x)$, где $\Phi(x)$ – стандартная нормальная функция распределения вероятностей. В этом случае λ -аппроксимация Тьюки для обратной функции Φ^{-1} ф.р. $\Phi(x)$ запишется в виде

$$\Phi_\lambda^{-1}(u) \approx [u^{0,135} - (1-u)^{0,135}] / 0,1975,$$

что соответствует следующему набору параметров: $\lambda_1 = 0$, $\lambda_3 = \lambda_4 = 0,135$, $\lambda_2 = 0,1975$. При таких значениях параметров качество аппроксимации вида (12) очень высокое и оно характеризуется неравенством $|\Phi_\lambda^{-1}(u) - \Phi^{-1}(u)| < 0,001$, $\forall u \in [0; 1]$. Численные значения функции регрессии, вычисленные по формуле (13), приведены в табл. 1.

Т а б л и ц а 1

Значения функции регрессии $M(X_i | R_i = j) = M(X_{(j)})$ для нормального распределения

j	1	2	3	4	5
$n=19$	-1,844	-1,378	-1,099	-0,886	-0,707
j	6	7	8	9	10
$n=19$	-0,548	-0,402	-0,264	-0,131	0

Отметим, что в табл.1. значения $M(X_{(j)})$ при $n=19$ для $j=11, \dots, 19$ не приведены, так как стандартное нормальное распределение симметрично относительно нуля и, следовательно, $M(X_{(n-r+1)}) = -M(X_{(r)})$, $r=1, \dots, n$. Отметим также, что данные табл.1 с точностью до трех знаков после запятой совпадают с дан-

ными в таблице 13.1. работы [7], которая содержит математические ожидания порядковых статистик для нормального распределения. Численные значения функции регрессии, вычисленные по формуле (10), приведены в табл. 2.

Т а б л и ц а 2

Значения функции регрессии $M(R_i | X_i = x) = 1 + (n-1)\Phi(x)$ для нормального распределения

x	-3	-2	-1	0	1	2	3
$n=19$	1,018	1,414	3,862	10	16,138	18,586	18,982

Качественный анализ поведения функций регрессий $M(R_i | X_i = x)$, $x \in R^1$ и $M(X_i | R_i = j)$, $j=1, \dots, n$, для нормального распределения при различных объемах выборки n показывает, что эти функции пересекаются и их угол пересечения при увеличении объема выборки n неограниченно убывает, что говорит об усилении связи между случайными величинами X_i и R_i , $i=1, \dots, n$. Этот вывод сохраняется и для других распределений.

Для количественной оценки зависимости между случайными величинами X_i и R_i , $i=1, \dots, n$ получим выражение для коэффициента корреляции $\rho_{X_i R_i}$, $i=1, \dots, n$. Справедлива следующая теорема.

Теорема 3. Пусть $X_1, \dots, X_n$ – н.о.р. случайные величины с ф. р. F_X и плотностью $f_X(x)$, $x \in R^1$ и пусть $\vec{R} = (R_1, \dots, R_n)$ – ранговый вектор для выборки $\vec{X} = (X_1, \dots, X_n)$. Тогда коэффициент корреляции $\rho_{X_i R_i}$ случайных величин X_i и R_i , $i=1, \dots, n$, зависит от ф.р. F_X через отношение двух функционалов, характеризующих масштабный параметр с.в. X , и вычисляется $\forall i \in (1, \dots, n)$ по формуле

$$\rho_{X_i R_i}(F) = \frac{\sqrt{3}}{2} \left(\frac{n-1}{n+1} \right)^{1/2} \frac{\Delta(F)}{S(F)}, \quad (14)$$

где $\Delta(F)$ – средняя разность Джини, определяемая в виде

$$\Delta(F) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |x-y| dF_X(x) dF_X(y), \quad (15)$$

и $S(F)$ – стандартное отклонение, определяемое в виде

$$S(F) = \left(\frac{1}{2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-y)^2 dF_X(x) dF_X(y) \right)^{1/2}. \quad (16)$$

Доказательство. Согласно определению, коэффициент корреляции $\rho_{X_i R_i}$ случайных величин X_i и R_i , $i=1, \dots, n$, записывается в виде

$$\rho_{X_i R_i} = \frac{M(X_i, R_i) - M(X_i)M(R_i)}{\sqrt{D(X_i)D(R_i)}}. \quad (17)$$

Нетрудно убедиться, что $\forall i \in (1, \dots, n)$

$$M(R_i) = (n+1)/2, \quad D(R_i) = (n^2-1)/12. \quad (18)$$

Далее, используя (8), получаем

$$\begin{aligned}
 M(X_i R_i) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{X_i R_i}(x, y) dx dy = \\
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy n^{-1} \sum_{j=1}^n f_{X(j)}(x) \delta(y - j) dx dy = \\
 &= n^{-1} \sum_{j=1}^n \int_{-\infty}^{+\infty} xj f_{X(j)}(x) dx = \\
 &= \int_{-\infty}^{+\infty} x f_X(x) \left[\sum_{j=1}^n j C_{n-1}^{j-1} F_X^{j-1}(x) (1 - F_X(x))^{n-j} \right] dx = \\
 &= \int_{-\infty}^{+\infty} x f_X(x) [(n-1)F_X(x) + 1] dx = \\
 &= (n-1) \int_{-\infty}^{+\infty} x F_X(x) dF_X(x) + \int_{-\infty}^{+\infty} x dF_X(x).
 \end{aligned}$$

Подставив найденные выражения в (17), и учитывая, что $\Delta(F) = 4 \int_{-\infty}^{+\infty} (x F_X(x) - x/2) dF_X(x)$, получаем формулу (14). Доказательство завершено.

Замечание 3. Вычисление $\rho_{X_i R_i}(F) \quad \forall i \in (1, \dots, n)$ для каждой конкретной ф.р. F_X может быть проведено непосредственно. Например, в гауссовском случае, то есть при $F_X(x) = \Phi(x)$, имеем: $\Delta(\Phi) = 2/\sqrt{\pi}$, $S(\Phi) = 1$ и, следовательно, согласно (14) при больших объемах выборки n , получаем $\rho_{XR}(\Phi) = \sqrt{3/\pi} = 0,98$. Также нетрудно убедиться, что для логистического распределения коэффициент корреляции между наблюдением и его рангом равен 0,95, а для распределения Лапласа он равен 0,92. Более того, можно указать нижнюю границу, которая справедлива для любой

непрерывной ф.р. F_X с ограниченной сверху плотностью f_X . В самом деле, используя равенство

$$\Delta(F) = 2 \int_{-\infty}^{+\infty} F_X(x)(1 - F_X(x)) dx,$$

получаем следующее неравенство

$$\begin{aligned}
 \Delta(F) &= 2 \int_{-\infty}^{+\infty} \frac{F_X(x)(1 - F_X(x))}{f_X(x)} dF_X(x) \geq 2 [\text{Sup}_x \{f_X(x)\}]^{-1} \times \\
 &\quad \times \int_{-\infty}^{+\infty} (F_X(x) - F_X^2(x)) dF_X(x) = \\
 &= 2 [\text{Sup}_x \{f_X(x)\}]^{-1} [1/2 - 1/3] = (1/3) \text{Sup}_x f_X(x).
 \end{aligned}$$

Отсюда, с использованием (14), при больших объемах выборки n , получаем для любой непрерывной ф.р. F_X с ограниченной сверху плотностью f_X следующее неравенство

$$\rho_{XR}(F) \geq [2\sqrt{3} \cdot S(F) \cdot \text{Sup}_x \{f(x)\}]^{-1},$$

которое в гауссовском случае приобретает вид $\rho_{XR}(\Phi) \geq 0,72$. Отметим еще одно полезное неравенство. Поскольку $\Delta(F) > 0$ и $S(F) > 0$, то правая часть в (14) также положительная величина. Кроме того, $\rho_{XR}(F) \leq 1$ и, следовательно, $0 \leq \rho_{XR}(F) \leq 1$, то есть наблюдение и его ранг положительно коррелированы случайные величины. Далее, из неравенства

$$\sqrt{3}\Delta(F) \leq 2S(F),$$

следует неравенство

$$\Delta(F) \leq 2S(F)/\sqrt{3},$$

которое справедливо для любых абсолютно непрерывных функций распределения F_X .

ЛИТЕРАТУРА

1. Тарасенко Ф.П. Непараметрическая статистика. Томск: Изд-во Том. ун-та, 1976.
2. Шуленин В.П. Введение в робастную статистику. Томск: Изд-во Том. ун-та, 1993.
3. Гаек Я., Шидак З. Теория ранговых критериев / Пер. с англ. под ред. Л.Н. Большева. М.: Наука, 1971.
4. Уилкс С. Математическая статистика. М.: Наука, 1967.
5. Тарасенко Ф.П., Шуленин В.П. О статистической связи между наблюдением и его рангом // Труды СФТИ при Том. ун-те. 1971. Вып. 62. С. 220–228.
6. Ramberg J.S., Schmeiser B.W. An approximative method for generating symmetric random variables // Commun. ACM. 1972. V. 15. P. 987–990.
7. Введение в теорию порядковых статистик / Пер. с англ. под ред. А.Я. Боярского. М.: Статистика, 1970.

Статья представлена кафедрой теоретической кибернетики факультета прикладной математики и кибернетики Томского государственного университета, поступила в научную редакцию 30 апреля 2003 г.