ТАРТУСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ТРУДЫ

ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА

56

ТАРТУ
1988

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ

ТРУДЫ ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА

Выпуск 56

ТАРТУ 1988

Утверждено на заседании совета математического факультета ТГУ 2 декабря 1988 года.

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ.
Труды вычислительного центра. Выпуск 56.
На русском языке.
Резюме на эстонском и английском языках.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Ülikooli, 18.
Ответственный редактор Э. Эхасалу.
Подписано к печати 15.12.1988.
МВ 02923.
Формат 60x84/16.
Бумага ротаторная.
Машинопись. Ротапринт.
Условно-печатных листов 6,98.
Учётно-издательских листов 5,61. Печатных листов 7,5.
Тираж 350.
Заказ № 1033.
Цена 1 руб. 10 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Тийги, 78.

Труды ВЦ ТГУ (1988), 56, стр.3-16.

Оценивание коэффициентов полиномиального тренда временных рядов с применением разностного оператора

В. Иоала

Ключевые слова: временной ряд, полиномиальный тренд, оценивание коэффициентов, разностный оператор.

I. Введение.

На практике встречается большой класс временных рядов, имеющих кроме случайной составляющей некоторую общую тенденцию изменения-тренд . В ряде случаев оказывается возможным приблизить тренд полиномом достаточно низкой степени. Основную модель наблюдаемого процесса при этом можно представить следующим образом:

$$x(t)=a_0+\sum_{i=1}^{n}a_i t^i+u_t\ ,\ t=1,\dots,M\ , \qquad (I)$$

где n - степень полинома, а ненаблюдаемые одинаково распределенные случайные величины u_t некоррелированы, имеют нулевые средние значения и дисперсии $\sigma^2 (Eu_t=0; Eu_t^2=\sigma^2; Eu_t u_s=0, t\neq s)$.

Для оценивания коэффициентов $a_i (i=0,\dots,n)$ обычно применяется метод наименьших квадратов (МНК), так как оценки МНК

являются наилучшими среди линейных несмещенных оценок. Во многих прикладных задачах оценивание по МНК является на вычислительных машинах слишком медленным из-за большого количества вычислительных операций.

В работах [3] и [1] приведены методы оценивания коэффициентов полиномиального тренда, позволяющие при незначительной потере в точности существенно сократить время вычисления оценок. В первом случае оценки вычисляются с применением т.н. функций Уолша, а также модифицированных функций Уолша, которые получаются умножением прямоугольных периодических функций. Во втором случае оценки коэффициентов линейного тренда ($n=1$), вычисляются с применением разностного оператора.

В настоящей работе будет приведен класс линейных оценок, основанный на применении разностных операторов и заключающий в себе функций Уолша, а также модифицированные функции Уолша в качестве подклассов. Приводятся оптимальные оценки в среднеквадратическом смысле.

2. Разностный оператор.

Разностный оператор ρ-го порядка $\Delta_\rho(t)$, ($\rho=0,1,2,\ldots$) для процесса $x(t)$ определяется следующим образом:

$$\Delta_0(t) = x(t);$$

$$\Delta_1(t) = x(t+N_1) - x(t);$$

$$\Delta_2(t) = \Delta_1(t+N_2) - \Delta_1(t) =$$

$$= x(t+N_1+N_2) - x(t+N_2) - x(t+N_1) + x(t);$$

$$\vdots$$

$$\Delta_\rho(t) = \Delta_{\rho-1}(t+N_\rho) - \Delta_{\rho-1}(t),$$

где N_ρ - шаг разностного оператора.

В работе [2] доказана основная теорема исчисления разностей для случая, когда $N_1 = N_2 = \ldots = N_\rho = N$.

Теорема I. Для процесса определяемой полиномом степени n

$$x(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n \quad (a_n \neq 0),$$

n-я разность постоянна и равна $a_n n! N^n$, $(n+1)$-я разность равна нулю. Аналогично удалось доказать, что для любых шагов N_j $(j = 1, \ldots, n)$ при условии $n \leq \sum_{j=1}^{n} N_j < M$ n-я разность равна $a_n n! \prod_{j=1}^{n} N_j$.

Индексация разностей начинается с $t = 1$ и заканчивается на $t = M - \sum_{j=1}^{\rho} N_j$, причем $\rho \leq \sum_{j=1}^{\rho} N_j < M$. В остальных случаях применение разностей теряет смысл.

Необходимо обратить внимание на свойство, позволяющее выписать сумму разностей с помощью сумм элементов ряда. Покажем это для разностей второго порядка.

$$\sum_{t=1}^{M-N_1-N_2} \Delta_2(t) = \sum_{t=1}^{M-N_1-N_2} [x(t+N_1+N_2) - x(t+N_2) - x(t+N_1) + x(t)] =$$

$$= \sum_{t=N_1+N_2+1}^{M} x(t) - \sum_{t=N_2+1}^{M-N_1} x(t) - \sum_{t=N_1+1}^{M-N_2} x(t) + \sum_{t=1}^{M-N_1-N_2} x(t). \qquad (2)$$

Количество сумм здесь равно 2^n. Это свойство будет широко использовано в дальнейших рассуждениях.

3. Общая формула.

Предлагаемый метод оценивания коэффициентов полиномиального тренда временных рядов основывается на усреднении разностей этого ряда.

Теорема 2. Пусть шаги разностного оператора N_j $(j = 1, \ldots, n)$

выбраны так, что $n \leq \sum_{j=1}^{n} N_j < M$, тогда для модели (I) оценка

$$a_n = \frac{1}{n!\left(\prod_{j=1}^{n} N_j\right)\left(M - \sum_{j=1}^{n} N_j\right)} \sum_{t=1}^{M-\sum_{j=1}^{n} N_j} \Delta_n(t) \qquad (3)$$

является несмещенной и состоятельной оценкой коэффициента a_n. В случае $n=0$ необходимо принять $\prod_{j=1}^{n} N_j = 1$ и $\sum_{j=1}^{n} N_j = 0$.

Доказательство. Для доказательства перепишем оценку (3) в другом виде. Поставив модель (I) в формулу оценки (3) и применяя теорему I получим

$$\hat{a}_n = a_n + \frac{1}{n!\left(\prod_{j=1}^{n} N_j\right)\left(M - \sum_{j=1}^{n} N_j\right)} \sum_{t=1}^{M-\sum_{j=1}^{n} N_j} \Delta_n u_t \,, \qquad (4)$$

где $\Delta_n u_t$ – разность n-го порядка случайного процесса u_t $(t = 1, \ldots, M)$.

Применяя свойство (2) и учитывая, что случайные величины u_t имеют нулевые средние, получим

$$E(\hat{a}_n) = a_n \,,$$

что и доказывает несмещенность оценки (3).

Дисперсия оценки (3) выписывается аналогичным путем, учитывая свойство (2) и то, что случайные величины u_t некоррелированы и имеют дисперсию σ^2.

$$E(\hat{a}_n - a_n)^2 = E\left\{ a_n + \frac{1}{n!\left(\prod_{j=1}^{n} N_j\right)\left(M - \sum_{j=1}^{n} N_j\right)} \sum_{t=1}^{M-\sum_{j=1}^{n} N_j} \Delta_n u_t - a_n \right\}^2 =$$

$$= \frac{1}{\left[n!\left(\prod_{j=1}^{n} N_j\right)\left(M - \sum_{j=1}^{n} N_j\right)\right]^2} E\left\{ \sum_{t=1}^{M-\sum_{j=1}^{n} N_j} \Delta_n u_t \right\}^2 = \qquad (5)$$

$$= \frac{2^n}{\left[n!\left(\prod_{j=1}^{n} N_j\right)\right]^2 \left(M - \sum_{j=1}^{n} N_j\right)} \sigma^2 .$$

Так как при $M \to \infty$ дисперсия $E(\hat{a}_n - a_n)^2 \to 0$, то состоятельность оценки (3) доказан.

Другие коэффициенты полинома $a_i\ (i = 0, \ldots, n-1)$ оцениваются теми же формулами, причем предварительно необходимо вычесть составляющие тренда высших порядков. Поэтому всегда первым оценивается коэффициент a_n и последним a_0. Для оценивания коэффициентов $a_i\ (i = 0, \ldots, n-1)$ справедлива формула

$$\hat{a}_i = \frac{V(i,n)}{P_i} - \sum_{k=i+1}^{n} c_{ki} \hat{a}_k , \qquad (6)$$

где

$$V(i,n) = \sum_{t=1}^{M-\sum_{j=1}^{i} N_j} \Delta_i(t) , \qquad (7)$$

$$P_i = i!\left(\prod_{j=1}^{i} N_j\right)\left(M - \sum_{j=1}^{i} N_j\right) , \qquad (8)$$

$$c_{ki} = \frac{V^*(i,n)}{P_i} , \qquad (10)$$

причем $V^*(i,n)$ вычисляется по формуле (7), где $x(t)=t^k$.

4. Оптимальные оценки в среднеквадратическом смысле.

Обратим внимание на то , что формула (3) определяет целый класс оценок для коэффициента a_n - выбирая разные комплекты $N_j (j=1,\dots,n)$ получаем разные оценки. Называем этот класс классом разностных оценок.

Теорема 3. В классе разностных оценок минимальная дисперсия оценки достигается при $N_1 = N_2 = \dots = N_n = \frac{2M}{2n+1}$.

Доказательство. Будем искать минимума дисперсии (5) по $N_j (j=1,\dots,n)$. Будем искать максимума выражения $\left(\prod_{j=1}^{n} N_j\right)^2 \left(M - \sum_{j=1}^{n} N_j\right)$, так как в этом случае достигается минимум дисперсии

$$\frac{\partial \left(\prod_{j=1}^{n} N_j\right)^2 \left(M - \sum_{j=1}^{n} N_j\right)}{\partial N_i} = 0 .$$

Отсюда

$$\frac{2\left(\prod_{j=1}^{n} N_j\right)^2 \left(M - \sum_{j=1}^{n} N_j\right)}{N_i} - \left(\prod_{j=1}^{n} N_j\right)^2 = 0 ,$$

$$N_i = 2\left(M - \sum_{j=1}^{n} N_j\right) , \quad i = 1,\dots,n . \qquad (11)$$

Из этого следует, что минимум дисперсии оценки достигается при одинаковых шагах разностного оператора. Заменяя $Z = N_i$ и $\sum_{j=1}^{n} N_j = nZ$ получим, что

$$N_i = Z = \frac{2M}{2n+1} , \quad i = 1,\dots,n , \qquad (12)$$

что и требовалось доказать.

Оценку $\hat{a}_n$ можно в таком случае выписать в следующем виде

$$\hat{a}_n = \frac{(2n+1)^{n+1}}{n!\, 2^n\, M^{n+1}} V(n,n),$$

где $V(n,n)$ определяется по формуле (7).

Используя полученные результаты, можно функции $V(n,n)$ представить в явном виде при $n = 0, 1, 2, 3$.

$$V(0,0) = \sum_{t=1}^{M} x(t);$$

$$V(1,1) = \sum_{t=\frac{2M}{3}+1}^{M} x(t) - \sum_{t=1}^{\frac{M}{3}} x(t);$$

$$V(2,2) = \sum_{t=\frac{4M}{5}+1}^{M} x(t) - 2\sum_{t=\frac{2M}{5}+1}^{\frac{3M}{5}} x(t) + \sum_{t=1}^{\frac{M}{5}} x(t);$$

$$V(3,3) = \sum_{t=\frac{6M}{7}+1}^{M} x(t) - 3\sum_{t=\frac{4M}{7}+1}^{\frac{5M}{7}} x(t) + 3\sum_{t=\frac{2M}{7}+1}^{\frac{3M}{7}} x(t) - \sum_{t=1}^{\frac{M}{7}} x(t).$$

Как видно из формул функции $V(n,n)$, при зафиксированном n элементы ряда в суммах встречаются только один раз. Некоторые из элементов ряда вообще отсутствуют в $V(n,n)$. Для сохранения этого полезного свойства необходимо модифицировать суммы в $V(n,n)$ при выписании вычислительных формул.

5. Вычислительные формулы.

$n = 0 \quad x(t) = a_0 + u_t, \quad t = 1, \ldots, M$

2

$$\hat{a}_0 = \frac{1}{M}\sum_{t=1}^{M} x(t)$$

$n=1$ $\quad x(t) = a_0 + a_1 t + u_t \;,\; t = 1,\dots,M$

$$S_1 = \sum_{t=1}^{M/3} x(t)\;;\quad S_2 = \sum_{t=\frac{M}{3}+1}^{2M/3} x(t)\;;\quad S_3 = \sum_{t=\frac{2M}{3}+1}^{M} x(t)$$

$$V(0,1) = S_1 + S_2 + S_3$$

$$V(1,1) = S_3 - S_1$$

$$\hat{a}_1 = \frac{4,5}{M^2}\, V(1,1)$$

$$\hat{a}_0 = \frac{V(0,1)}{M} - \hat{a}_1 \frac{M+1}{2}$$

$n=2$ $\quad x(t) = a_0 + a_1 t + a_2 t^2 + u_t \;,\; t = 1,\dots,M$

$$S_1 = \sum_{t=1}^{M/5} x(t)\;;\quad S_2 = \sum_{t=\frac{M}{5}+1}^{M/3} x(t)\;,\quad S_3 = \sum_{t=\frac{M}{3}+1}^{2M/5} x(t)\;;\quad S_4 = \sum_{t=\frac{2M}{5}+1}^{3M/5} x(t)$$

$$S_5 = \sum_{t=\frac{3M}{5}+1}^{2M/3} x(t)\;;\quad S_6 = \sum_{t=\frac{2M}{3}+1}^{4M/5} x(t)\;;\quad S_7 = \sum_{t=\frac{4M}{5}+1}^{M} x(t)$$

$$V(0,2) = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7$$

$$V(1,2) = S_7 + S_6 - S_2 - S_1$$

$$V(2,2) = S_7 - 2\cdot S_4 + S_1$$

$$\hat{a}_2 = \frac{125}{8\cdot M^3}\, V(2,2)$$

$$\hat{a}_1 = \frac{4,5}{M^2}\, V(1,2) - \hat{a}_2 (M+1)$$

$$\hat{a}_0 = \frac{V(0,2)}{M} - \hat{a}_2 \frac{(M+1)(2M+1)}{6} - \hat{a}_1 \frac{M+1}{2}$$

$n=3$ $\quad x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + u_t \;, \; t = 1, \ldots, M$

$$S_1 = \sum_{t=1}^{M/7} x(t); \quad S_2 = \sum_{t=\frac{M}{7}+1}^{M/5} x(t); \quad S_3 = \sum_{t=\frac{M}{5}+1}^{2M/7} x(t); \quad S_4 = \sum_{t=\frac{2M}{7}+1}^{M/3} x(t)$$

$$S_5 = \sum_{t=\frac{M}{3}+1}^{2M/5} x(t); \quad S_6 = \sum_{t=\frac{2M}{5}+1}^{3M/7} x(t); \quad S_7 = \sum_{t=\frac{3M}{7}+1}^{4M/7} x(t); \quad S_8 = \sum_{t=\frac{4M}{7}+1}^{3M/5} x(t)$$

$$S_9 = \sum_{t=\frac{3M}{5}+1}^{2M/3} x(t); \quad S_{10} = \sum_{t=\frac{2M}{3}+1}^{5M/7} x(t); \quad S_{11} = \sum_{t=\frac{5M}{7}+1}^{4M/5} x(t); \quad S_{12} = \sum_{t=\frac{4M}{5}+1}^{6M/7} x(t)$$

$$S_{13} = \sum_{t=\frac{6M}{7}+1}^{M} x(t)$$

$$V(0,3) = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7 + S_8 + S_9 + S_{10} + S_{11} + S_{12} + S_{13}$$

$$V(1,3) = S_{13} + S_{12} + S_{11} + S_{10} - S_4 - S_3 - S_2 - S_1$$

$$V(2,3) = S_1 + S_2 - 2(S_6 + S_7 + S_8) + S_{12} + S_{13}$$

$$V(3,3) = S_{13} - S_1 + 3(S_6 + S_5 + S_4 - S_{10} - S_9 - S_8)$$

$$\hat{a}_3 = \frac{2401}{48 \cdot M^4} V(3,3)$$

$$\hat{a}_2 = \frac{125}{8 \cdot M^3} V(2,3) - \hat{a}_3 \frac{3(M+1)}{2}$$

$$\hat{a}_1 = \frac{4,5}{M^2} V(1,3) - \hat{a}_2 (M+1) - \hat{a}_3 \frac{16M^2 + 27M + 9}{18}$$

$$\hat{a}_0 = \frac{V(0,3)}{M} - \hat{a}_1 \frac{M+1}{2} - \hat{a}_2 \frac{(M+1)(2M+1)}{6} - \hat{a}_3 \frac{M(M+1)^2}{4}.$$

6. Сравнение с другими оценками.

Для сравнения качества оценивания используем дисперсии оценок МНК, а также оценки, приведенные в работе [3]. Из формулы (3) можно получить формулы оценивания, совпадающие с формулами оценивания при помощи функций Уолша, когда

$$N_i = \frac{M}{2^i}, \quad i = 1, \ldots, n. \tag{13}$$

Модифицированные функции Уолша можно получить, когда

$$N_i = \frac{M}{2^i}, \quad i = 1, \ldots, n-1; \tag{14}$$

$$N_n = \frac{M}{3 \cdot 2^{n-2}}.$$

В таблице I приведены дисперсии оценок (3) при разных шагах N_i, определенные выражениями (12), (13) и (14).

Таблица I.

Дисперсии оценок коэффициентов a_n.

n	(12)	(13)	(14)
0	$\frac{1}{M}\sigma^2$	$\frac{1}{M}\sigma^2$	$\frac{1}{M}\sigma^2$
1	$\frac{13,5}{M^3}\sigma^2$	$\frac{16}{M^3}\sigma^2$	$\frac{13,5}{M^3}\sigma^2$
2	$\frac{3125}{16M^5}\sigma^2$	$\frac{256}{M^5}\sigma^2$	$\frac{216}{M^5}\sigma^2$
3	$\frac{823543}{283M^7}\sigma^2$	$\frac{65536}{9M^7}\sigma^2$	$\frac{6144}{M^7}\sigma^2$

В таблице 2 приведены дисперсии оценок МНК и оценок, полученных с применением разностного оператора при $n=0,1,2,3$, выбирая $M=100$ и $\sigma^2=1$.

Таблица 2.

Сравнение дисперсий оценок.

n	МНК	(I2)	(I4)	(I3)
0	10^{-2}	10^{-2}	10^{-2}	10^{-2}
1	$1,21 \cdot 10^{-5}$	$1,35 \cdot 10^{-5}$	$1,35 \cdot 10^{-5}$	$1,60 \cdot 10^{-5}$
2	$1,80 \cdot 10^{-8}$	$1,95 \cdot 10^{-8}$	$2,16 \cdot 10^{-8}$	$2,56 \cdot 10^{-8}$
3	$2,30 \cdot 10^{-11}$	$2,86 \cdot 10^{-11}$	$6,14 \cdot 10^{-11}$	$7,28 \cdot 10^{-11}$

Как видно из таблицы 2, оптимальные оценки в срендеквадратическом смысле (I2) имеют дисперсии, близкие к оценкам МНК. При этом время вычисления оценок, как показано в работе [3], значительно уменьшается. При $n=3$, используя арифметику с плавающей запятой, выигрыш на ЭВМ семейства PDP-II был приблизительно восьмикратным.

Другой аспект, на который необходимо обратить внимание, это ограничения объема выборки M. Когда МНК не предъявляет никаких ограничений на объем выборки, такие ограничения имеют оптимальные оценки в среднеквадратическом смысле, при которых объем выборки должен соответствовать равенству

$$M = k \prod_{j=0}^{n} (2_j + 1), \quad k = 1,2,3,\ldots$$

Таким образом, для полинома третьего порядка ($n=3$) минимальный объем выборки равен $M=105$. Остальные оценки (I3), (I4) ограничены соответственно $M=2^7 k$ и $M=1,5\ 2^7 k$, ($k=1,2,3,\ldots$).

Не удовлетворяя этим условием, оценки коэффициентов будут смещенными.

7. Выводы.

В настоящей работе приведен класс линейных оценок коэффициентов полиномиального тренда временных рядов, который при незначительной потере в точности позволяет существенно уменьшить время вычислений оценок. Это достигается доведением до минимума количества операций умножения, а также тем, что каждый элемент ряда суммируется только один раз.

Хотя предлагаемый метод накладывает ограничения на объем выборки, его можно успешно использовать для повышения производительности технических систем, в которых изменение исследуемого процесса производится с применением аналого-цифровых преобразователей. Измеренные значения при этом получаются целочисленные, что позволяет для вычисления сумм S_j и $V(i,n)$ применять целочисленную арифметику. Это дает дополнительный выигрыш во времени при вычислении оценок по сравнению с МНК.

Полученные результаты могут быть использованы и при планировании эксперимента. Например, при оценивании коэффициента a модели $x(t)=at^m$ можно некоторые эксперименты (измерения) пропустить. Для оптимальных оценок в среднеквадратическом смысле при зафиксированном n доля пропущенных экспериментов составляет $nM/(2n+1)$. Когда эксперименты дорогие, можно таким образом существенно уменьшить стоимость исследований.

Литература

I. Йоала В., Ольман В. Метод оценивания коэффициентов линейной одномерной регрессии. Известия Академии Наук Эстонской ССР. Физика и Математика, т.36, 4, с. 422-424.

2. Хемминг Р.В. Численные методы для научных работников и инженеров. - М.: Наука, 1972. - 400 с.

3. Joala V. An Estimation Method for Coefficients of Polynomial Trend in Time Series. Acta et Commentationes Universitatis Tartuensis, 1988, 789, p. 68-75.

Поступило 6.10.1988

Resümee

Aegridade polünomiaalse trendi koefitsientide hindamine juurdekasvuoperaatorite abil

V.Joala

Antud töös on esitatud lineaarsete hinnangute klass aegridade polünomiaalse trendi koefitsientide hindamiseks, mis põhineb juurdekasvuoperaatori kasutamisel. On leitud optimaalsed hinnangud ruutkeskmise hälbe mõttes. Toodud on arvutusvalemid koefitsientide hindamiseks kuni kolmandat järku polünoomide jaoks. On esitatud hinnangute dispersioonide võrdlus vähimruutmeetodiga. Artiklis esitatud hindamismeetod võimaldab tunduvalt tõsta hinnangute arvutamise kiirust arvutitel, kaotades tühiselt hindamistäpsuses.

Summary

An Estimation Method for Coefficients of Polynomial Trend in Time Series Based on Using Differences

V.Joala

In this paper the author gives an estimation method of

coefficients of polynomial trend in time series based on using differences. The variance of estimated coefficients is minimized by this method and allows us to work much faster on computers than by using the least-square method. It has been shown that accuracy is approximately the same as in the case of the least-square method. Irrespective of the limitations as to the length of the time series the method can be easily applied to improve the productivity of computer programs in engineering and it is used in the planning of experiments.

Труды ВЦ ТГУ (1988), 58, стр.17-27.

Об алгоритмах вычисления некоторых основных параметров временных рядов

К. Кийранен

Ключевые слова: плотность распределения, частотная таблица, корреляционная и автокорреляционная функции

I. Постановка проблемы

Целью настоящей статьи является описание нескольких оригинальных программ, разработанных в лаборатории биофизики ТГУ для обработки информации, получаемой приборами, измеряющими физиологические сигналы. В лаборатории ведутся экспериментальные исследования по выяснению регуляции систем кровообращения и дыхания. Предлагается краткое знакомство с работами, выполняемыми при помощи специального комплекса приборов, позволяющего проводить одновременные измерения и регистрации целого ряда (обыкновенно от двух до восьми) физиологических параметров систем кровообращения и дыхания (среднее артериальное давление, частота сердечных сокращений, объемная скорость кровотока, частота дыхания и др.). Часть этих параметров (кровоток, объемная скорость дыхания и др.) являются непрерывными величинами. Другая часть названных параметров (частота сердца и др.) фикси-

руется за каждый цикл работы органа (например, сердца) или, как принято говорить, также непрерывно, хотя в действительности получается дискретный ряд наблюдений. Вся поступившая информация записывается в виде аналоговых (т.е. непрерывных) сигналов с помощью многоканального магнитографа на магнитную ленту, чем обеспечивается возможность повторной обработки данных. Обработка ведется или на АВМ, или ЦВМ. В последнем случае аналоговые сигналы пропускаются через аналог-код-преобразователи, где непрерывный сигнал квантуется. Нами разработано программное обеспечение для обработки и анализа информации, поступающей в виде многомерных временных рядов.

В зависимости от характера анализируемых процессов предвидится возможность для выбора длины шага квантования. Ясно, что программы должны без больших изменений быть применимы во всех частных случаях (независимо, например, от числа временных рядов, шага квантования и т.д.). Приведенные программы используются главным образом для первичного анализа материала, притом широко используется визуализация.

На основании результатов этого этапа работы возможно судить о неизбежности применения некоторых других методов. Подходящими методами для первичного анализа многомерных временных рядов, описывающих физиологические процессы, являются([2] , [1]): 1) оценка одно- и многомерных распределений (таблиц частот), 2) оценка автокорреляционных и взаимных корреляционных функций.

Разумеется, приведенные программы пригодны при исследовании любых многомерных временных рядов ([3] , [4]).

2. Описание применяемой методики

Мы рассмотрим дискретные временные ряды, состоящие из определенной последовательности случайных величин ..., $w(t_1)$, $w(t_2)$, ..., $w(t_n)$, ..., где n пробегает множество целых чисел, а t_i означает некоторый определенный параметр, который чаще всего будет истолкован как время. Мы предположим, что элементы последовательности равностоящие (ряд квантован по равным интервалам), т.е. разность $t_i - t_{i-1} = c$ ($i = 2, \ldots, n$) одинакова при всех n.

В работе предлагаются алгоритмы для определения некоторых характеристик временных рядов - плотности распределения, взаимной и автокорреляционной функции. Программы их вычисления созданы на ЭВМ серий ЕС и "Искра", которые оформлены соответственно на языках "Фортран" и "Бейсик-2".

2.I. Плотность распределения.

При вычислении плотности временного ряда делается дополнительное предположение, что этот временной ряд является стационарным (случайным процессом), т.е. что его распределение не изменяется во времени. В данном случае можно рассмотреть наблюдения в различные моменты как элементы выборки и построить таблицу частот, которая является и оценкой функции плотности.

Пример I. Частотная таблица. В качестве примера временных рядов рассмотрим (см. Рис.I) частоту сердечных сокращений ϕ и среднее артериальное давление на пальцах $\bar{P}_a$, зарегистрированные через определенный промежуток времени (в данном случае через одну секунду). Ряды построены с соответствующими данными при t =I,...,70.

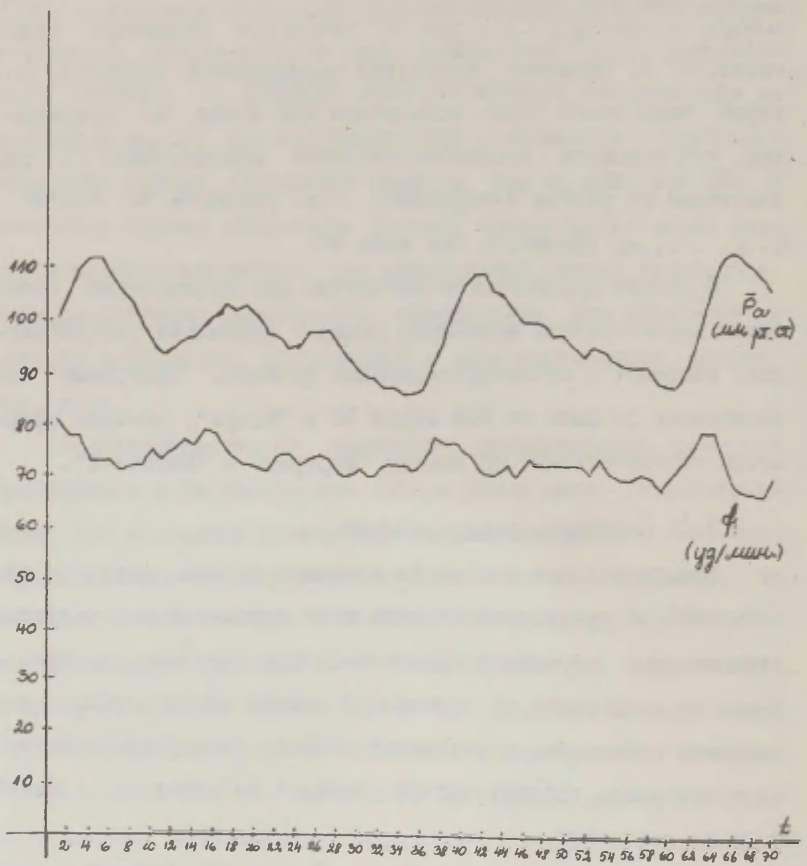
110
100
90
80
70
60
50
40
30
20
10
$\bar{P}_{ar}$
(мм. рт. ст)
ϕ
(уд/мин)
t
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70

Рис. 1.

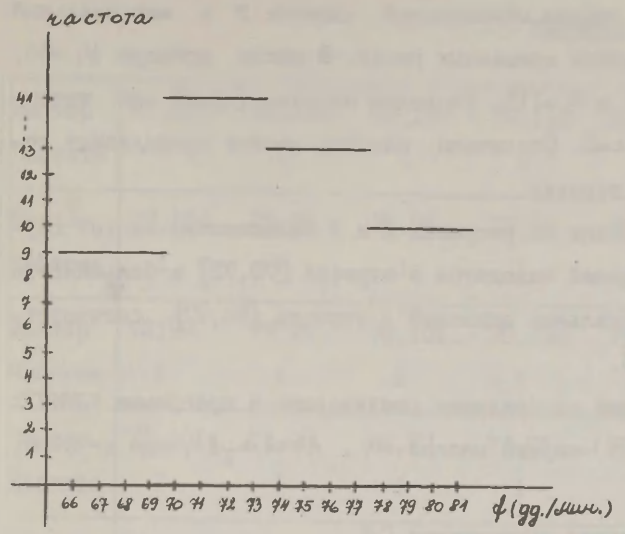
частота
41
13
12
11
10
9
8
7
6
5
4
3
2
1
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
ф (gg./мин.)

Рис. 2.

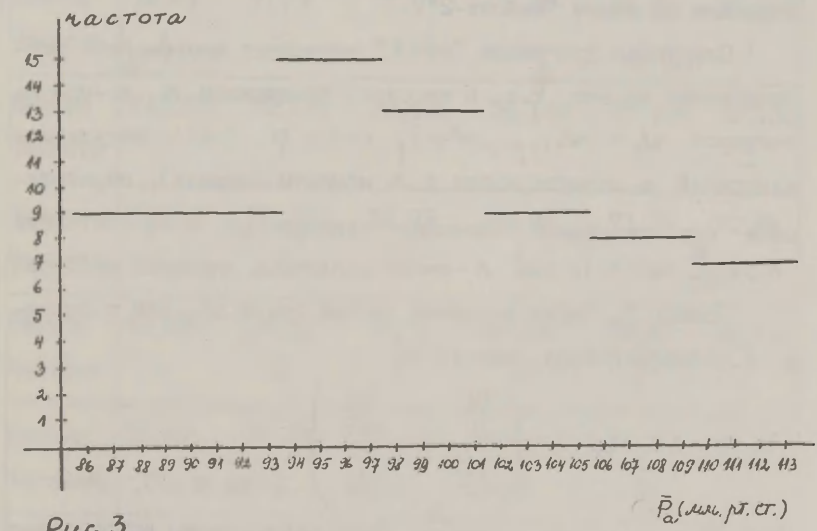
частота
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113
$\bar{P}_a$ (мм. рт. ст.)

Рис. 3.

Теперь найдем минимальный элемент y и максимальный элемент z данных временных рядов. В нашем примере y_1 =66, y_2 =86, z_1 =81 и z_2 =113. Разделим отрезок $[y,z]$ на интервалы (длины) c =3. Сосчитаем, сколько частот принадлежит тому или иному отрезку.

Как показано на рисунках 2 и 3 большинство частот сердечных сокращений находятся в отрезке [70,73] и большинство средних артериальных давлений в отрезке [94,97], соответственно 41 и 15.

Предыдущее рассуждение реализовано в программе "JAOT":
- задан $(n \times p)$-мерный массив $A4$, $A4 = \| a_{ij} \|$, где $i = 1, \dots n$;

- выбран интервал группировки $C1$.

Память ЭВМ "Искра" позволяет одновременно обрабатывать 800-элементные целочисленные массивы. Программа "JAOT" записана на языке "Бейсик-2".

Следующая программа "SAGLE" вычисляет многомерное распределение частот, т.е. в массиве, содержащем n p-мерных векторов $w_i = (w_{i1}, \dots, w_{ip})$, $i = 1, \dots n$, (это результаты измерений p характеристик в n моменты времени), определяются все различные значения вектора w_i и их кратности n_i ($\sum_{i=1}^{k} n_i = n$), где k -число различных значений вектора).

Пример 2. Пусть исходные данные такие же, как в примере I. Следовательно, массив W :

$$W = \begin{pmatrix} 81 & 101 \\ 78 & 105 \\ 78 & 110 \\ \vdots & \vdots \\ 69 & 106 \end{pmatrix}$$, где n =70, p =2

Требуется провести контроль того, как часто встречается

Таблица I.

Вектор	8I,I0I	78,I05	78,II0	73,II2	73,I09	7I,I06
Частота	I	I	I	2	I	I
Вектор	72,I04	72,99	75,96	73,94	75,95	77,97
Частота	I	2	I	I	I	I
Вектор	76,97	79,99	78,I0I	75,I03	73,I02	72,I03
Частота	I	I	I	I	I	I
Вектор	7I,I0I	7I,98	74,97	74,95	72,95	73,98
Частота	I	I	I	I	2	2
Вектор	74,98	70,95	70,92	7I,90	70,89	72,87
Частота	I	I	I	I	I	3
Вектор	72,86	7I,86	73,87	77,92	76,98	76,I03
Частота	I	I	I	I	I	I
Вектор	74,I07	72,I09	70,I09	70,I06	72,I05	70,I02
Частота	I	I	I	I	I	I
Вектор	73,I00	72,I00	72,97	70,93	73,95	70,94
Частота	I	I	2	I	I	I
Вектор	69,92	69,9I	70,9I	67,88	70,87	75,9I
Частота	I	2	I	I	I	I
Вектор	78,87	78,I04	73,II0	68,II3	67,II3	66,III
Частота	I	I	I	I	I	I
Вектор	66,I09	69,I06				
Частота	I	I				

каждый вектор ряда. Получим таблицу I.

Из таблицы I следует, что во время 70 секунд трижды встречается значение (72,87) вектора (f, $\bar{P}_\alpha$). Обычно время опыта длиннее, тогда и частоты в среднем больше.

Программа "SAGLE" записана на языках "Фортран" и "Бейсик-2".

2.2. Корреляционная и автокорреляционная функции.

Взаимная корреляционная функция двух временных рядов является характеристикой частоты взаимосвязи между этими рядами. Пусть задан $(n \times p)$-мерный массив W , где каждый столбец представляет собой один временной ряд:

$$W = \begin{pmatrix} w_{11} & w_{21} & w_{31} & \dots & w_{p1} \\ w_{12} & w_{22} & w_{32} & \dots & w_{p2} \\ w_{13} & w_{23} & w_{33} & \dots & w_{p3} \\ \dots & \dots & \dots & \dots & \dots \\ w_{1n} & w_{2n} & w_{3n} & \dots & w_{pn} \end{pmatrix}.$$

Обозначим корреляции между рядами через r_{kl} , где $k, l = 1, \dots, p$. Корреляция вычисляется при помощи следующей общеизвестной формулы:

$$r_{kl} = \frac{\sum_{i=1}^{n} (w_{ik} - \bar{w}_k)(w_{il} - \bar{w}_l)}{\left\{ \sum_{i=1}^{n} (w_{ik} - \bar{w}_k)^2 \sum_{i=1}^{n} (w_{il} - \bar{w}_l)^2 \right\}^{\frac{1}{2}}} \qquad (I)$$

где

$$\bar{w}_k = \frac{1}{n} \sum_{i=1}^{n} w_{ik} \quad \text{и} \quad \bar{w}_l = \frac{1}{n} \sum_{i=1}^{n} w_{il} .$$

Рассмотрим теперь элементы двух временных рядов, которые сдвинуты относительно друг друга на m шагов (моментов времени), $(m = 0, \dots, M; (M < n))$, тогда формула корреляции между рядами приобретет следующий вид:

$$r_{k\ell(m)} = \frac{\sum_{i=1}^{n-m} (w_{ik} - \bar{w}_k)(w_{i+m,\ell} - \bar{w}_\ell)}{\left\{ \sum_{i=1}^{n-m} (w_{ik} - \bar{w}_k)^2 \sum_{i=1}^{n-m} (w_{i+m,\ell} - \bar{w}_\ell)^2 \right\}^{\frac{1}{2}}} \qquad (2)$$

где

$$\bar{w}_k = \frac{1}{n-m} \sum_{i=1}^{n-m} w_{ik} \quad \text{и} \quad \bar{w}_\ell = \frac{1}{n-m} \sum_{i=1}^{n-m} w_{i+m,\ell} \; .$$

Формула (I) является частным случаем формулы (2) при сдвиге $m = 0$.

Взаимная корреляционная функция между временными рядами с индексами k и ℓ имеет $M+1$ значений $r_{k\ell}, r_{k\ell(1)}, \ldots, r_{k\ell(M)}$.

Если число рассматриваемых временных рядов есть p, то число различных взаимных корреляционных функций есть p^2.

Автокорреляция представляет собой корреляционную функцию временного ряда с самим собой. Автокорреляционная функция случайного процесса характеризует линейную зависимость значений процесса в определенный момент времени от его значений в другой момент.

Автокорреляционная функция вычисляется по тем же формулам (2), где только $k = \ell$. Вычисление корреляционных функций реализовано в программе "NKOR". Программа "NKOR" записана на языке "Бейсик-2" для ЭВМ типа "Искра 226".

Алгоритм программы "NKOR":

- задать $(n \times p)$-мерный массив W и указать желательный максимальный сдвиг M, который определяется по длине рассматриваемого ряда. Так как в реальных экспериментах длина ряда является всегда конечной, то M считается обыкновенно равным $\frac{1}{10} n$.

Программа вычисляет корреляции со сдвигом 0,...,M между всеми рядами.

Для программы "NKOR" число элементов временного ря-

ду не должно превышать 500.

Такой комплект программ работает с 1986 года, в связи с практическими нуждами его дополнили и сделали соответствующие выводы. В будущем хочется соединить его с комплексом программ спектральной плотности.

Литература

1. Бендат Дж., Пирсол А. Измерение и анализ случайных процессов. - М.: Мир, 1971.
2. Бендат Дж., Пирсол А. Применение корреляционного и спектрального анализа. - М.: Мир, 1983.
3. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976.
4. Кендэл М. Временные ряды. - М.: Финансы и статистика, 1981.

Поступило 14.07.1988

R e s ü m e e

Aegridade mõningate peamiste parameetrite arvutusalgoritmid

K.Kiiranen

Artiklis on kirjeldatud programme, mis on välja töötatud ja kasutatud TRÜ Biofüüsika laboratooriumis. Lühidalt tutvustatakse töid, mis võimaldavad töödelda üheaegselt mitmeid füsioloogilisi parameetreid (keskmine vererõhk, Südamelöögisagedus, hingamissagedus jt.). Programmid töötlevad ja analüüsivad informatsiooni, mis on esitatud mitmemõõtmeliste aegridade kujul. Arvutusprogrammid on kirjutatud arvutitele seeriast EC ja "Iskra" vastavalt keeltes "Fortran"ja "Basic".

S u m m a r y

Computation algorithms for some general parameters of time-series

K.Kiiranen

Some programs worked out and used in the Laboratory of Biophysics of Tartu State University have been described in this article. The programs which enable to process simultaneously several physiological parameters (mean blood pressure, heart rate, respiration rate,etc) have been presented by means of multidimensional time-series. The calculating programs are written for the EC and "Iskra" computers in Fortran and Basic respectively.

Труды ВЦ ТГУ (1988), 56, стр.28-35.

Верхние и нижние вероятности. Вычислительные результаты

Т. Кинкар

Ключевые слова: верхняя и нижняя вероятность.

Фундаментальную задачу для практической статистики Karl Pearson (1920) сформулировал так:

Во время $p+q=n$ испытаний, когда мы не имели априорной информации о частоте "событий" в генеральной совокупности, "событие" встречалось p раз. Каким является вероятность появления "события" r раз во время последующих $r+s=m$ испытаний?

Обычно при решении таких задач считается целесообразным вводить какие-то разумные предположения насчет распределения элементарных событий в генеральной совокупности. Иногда такие предположения имеют под собой реальную почву, но часто они не соответствуют реальности.

В связи с этим представляет определенный интерес изучение возможностей определения постериорных вероятностей событий опираясь только на реальные данные, полученные при первых (обучающих) испытаниях.

В своей статье A.P. Dempster (1967) исследует возможные пути решения именно такой задачи. Рассматривается сле-

цующая ситуация. Имеется конечная генеральная совокупность, которая состоит из N объектов и две выборки с размерами n и m, которые содержат $n+m$ разных элементов из генеральной совокупности. Предположим, что к моменту принятия решения мы имеем первую выборку с размером n. Решения принимаются насчет возможных параметров второй (следующей) выборки и генеральной совокупности. Объекты в генеральной совокупности разделяются в к взаимно непересекающиеся множества. Проведение наблюдений заключается в определении индекса множества, которому принадлежит соответствующий элемент. Таким образом, информацию о генеральной совокупности и о выборках можно передать в виде векторов:

$$\bar{N} = (N_1, N_2, \ldots, N_k),$$
$$\bar{n} = (n_1, n_2, \ldots, n_k),$$
$$\bar{m} = (m_1, m_2, \ldots, m_k).$$

где, например, n_2 указывает число элементов в первой выборке, принадлежащих второму множеству. Векторы $\bar{N}$, $\bar{n}$ и $\bar{m}$ должны удовлетворять некоторым очевидным требованиям: элементы должны быть неотрицательные числа и удовлетворять условиям

$$n_j + m_j \leqslant N_j \qquad j=1,2,\ldots,k \ ,$$

$$\sum_{j=1}^{k} N_j = N, \qquad \sum_{j=1}^{k} n_j = n \qquad \text{и} \qquad \sum_{j=1}^{k} m_j = m.$$

Предполагается, что n, m и N известны заранее.

Kupster в своей работе предлагает специальные формулы для вычисления верхних и нижних вероятностей $P^*(T)$ и $P_*(T)$, где

$$T = \left\{ \cdot \leqslant \sum_{j=a}^{b} m_j \leqslant \cdot \right\}$$

для $0 \le r < t$ и $1 < a < b \le k$. Понятия верхних и нижних вероятностей будут определены ниже.

При такой постановке задачи, предполагая, что элементы генеральной совокупности упорядочены и имеют индексы, мы можем представить себе, что у нас имеется пространство с элементами $[J,K]$, где J является вектором индексов элементов из первой выборки и K вектором индексов элементов из второй выборки. Элементы в генеральной совокупности расположены так, что элементы с индексами $(1, N_1)$ принадлежат первому множеству, с индексами (N_1+1, N_2) второму множеству итд.

Представим, что у нас имеется и пространство с элементами $[\bar{n},\bar{m},\bar{N}]$, где $\bar{n},\bar{m}$ и $\bar{N}$ - всевозможные векторы, описание и смысл которых представлены выше. Обозначим эти пространства соответственно X и S. Величины n,m,N и k известны.

Между этими пространствами можно построить отображение, которое является множественной в обоих направлениях. Фиксируем теперь некоторое значение $\bar{n}^*$. Этим в пространстве S определяется подпространство S^*, элементы которого содержат значение $\bar{n}^*$. После этого для каждого элемента из пространства X имеется три варианта:

- все его элементы попадут в пространство S^*
- только часть его отображений попадут в S^*
- ни одно из его отображений не попадет в S^*.

Верхние и нижние вероятности $P^*(T)$ и $P_*(T)$ для любого T из S определяются количеством элементов в множествах T^* и T_*, где T^* является множеством в пространстве X, отображения элементов которого хотя бы частично попали в T и T_*-множество в пространстве X, отображения элементов которого полностью попали в T.

Так как Dempster в своей работе предложил формулы для вычислений и привел лишь небольшой пример, возник интерес провести вычислительный эксперимент, чтобы более точно изучить свойства поведения верхних и нижних вероятностей при реальных данных. Эксперимент проводился на ПЭВМ Искра-1030.11. На языке Turbo-Pascal была написана соответствующая программа. Для генерирования данных с нормальным и равномерным распределением был использован генератор псевдо-случайных чисел из ППП "Статпак".

Все эксперименты проводились дважды, используя данные с равномерным и нормальным распределением. Результаты экспериментов при разных распределениях совпадали.

При такой постановке задачи полученные верхние и нижние вероятности можно сравнить только с вероятностью, вычисленной просто по полученным частотам в первой выборке.

Например, если $k=5$, $r=1$, $t=1$, $a=1$, $b=3$, $m=1$ и $\bar{n}=(4,7,3,2,4)$, $P_*(T)=0,6(6)$ и $P^*(T)=0,7143$. Обычная вероятность равняется $P(T)=((n_1+n_3+n_4)/n)=(14/20)=0,7$.

Оказалось, что во всех экспериментах обычная вероятность находилась между нижней и верхней вероятностью. Из этого видно, что между данными определениями вероятностей не существует противоречия. Обычная вероятность дает точечную оценку, которая почти всегда неверна. Верхняя и нижняя вероятности определяют отрезок, к которому относится исследуемая вероятность.

Основной задачей при проведении испытаний и было изучение динамики изменения отрезка между верхней и нижней вероятностью.

Испытания проводились следующим образом:

1: создание первой выборки (например $n=10$)

2: вычисление вероятности $P^*(T)$ и $P_*(T)$

3: если разница между вероятностями оказалась меньше величины Q (например, $Q=0.05$), эксперимент был закончен

4: в противном случае первой выборке добавили одно наблюдение и возвратились ко второму пункту.

Такие испытания проводились при разных значениях a, ℓ, r, t и m. Число множеств было всегда $k=5$.

Из вычислительных формул вероятностей вытекает, что исследуемая динамика не зависит от числа множеств k в генеральной совокупности. После проведения вычислений оказалось, что точность определения неизвестной вероятности (что определяется разницей между верхней и нижней вероятностью) зависит только от величин n и m.

Если вычислялись вероятности события, что следующее наблюдение попадет в некоторое множество или группу множеств, т.е. $m=1$, тогда во время многих длинных экспериментов были стабильно получены следующие результаты:

n	25	50	100	150	200	250
Q	0.08	0.04	0.02	0.012	0.01	0.008

Очевидно, что чем больше m, тем труднее принимать решения с заданной точностью. Чтобы получить разницу между верхней и нижней вероятностями меньше чем $Q=0.05$ при возрастающих значениях m, потребуется значительно больше наблюдений в первой выборке. При испытаниях были получены следующие результаты:

- при $m=1$ потребуется n приблизительно 40
- при $m=2$ потребуется n приблизительно 80

- при m=3 потребуется n приблизительно II0
- при m=4 потребуется n приблизительно I40
- при m=5 потребуется n приблизительно I60.

Если брать Q=0.I, тогда результаты будут такие:

- при m=I потребуется n приблизительно 20
- при m=2 потребуется n приблизительно 40
- при m=5 потребуется n приблизительно 75.

В итоге можно отметить, что такое определение вероятности может быть использовано и в решении практических задач. Разумеется, имеется и ряд трудностей перед широким внедрением верхних и нижних вероятностей в практическую статистику. Не изучены пути соединения информации,полученной из разных источников (например, априорной информации и информации, полученной из первой выборки). Но как показали испытания, не существует явного противоречия между частотным определением вероятности и верхними и нижними вероятностями,и результаты не являются чрезмерно расплывчатыми.

В поддержку верхних и нижних вероятностей говорит и то, что эти результаты являются более логическими чем результаты, полученные с использованием доверительных границ. Доверительные границы строятся симметрично вокруг точечной оценки без учета при этом расположения результатов наблюдений в первой выборке. При построении верхних и нижних вероятностей учитывается и эта информация и результаты получаются при этом более естественным образом.

Для описания вычислительных трудностей, встречаемых при нахождении одной пары верхних и нижних вероятностей, можно сказать, что для этого надо четыре раза вычислять значение функции распределения из гипергеометрического семейства и еще один уточняющий член. При разумной реализации это зай - мет не слишком много времени.

Литература

1. Dempster, A.P. Upper and lower probability inferences based on a sample from a finite univariate population. Biometrica, 1967, 54, 515-528.
2. Pearson, K. The fundamental problem of practical statistics. Biometrica, 1920, 13, 1-16.

Поступило 3 .IO. I988

R e s ü m e e

Ülemised ja alumised tõenäosused
Arvutuslikud tulemused

T.Kinkar

Artiklis vaadeldakse alumiste ja ülemiste tõenäosuste praktilise arvutamise probleemi, kasutades seejuures A.P.Dempsteri poolt 1967.a. pakutud valemeid.

On esitatud läbiviidud eksperimentide ja saadud tulemuste kirjeldused.

Tuginedes teostatud katsetele võib öelda, et ülemised ja alumised tõenäosused ei ole vastuolus tõenäosuse määratlusega sageduste kaudu. Nõutav arvutuste maht pole ülemäära suur. Pärast erinevatest allikatest pärineva info ühendamise probleemi lahendamist on alust oodata ülemiste ja alumiste tõenäosuste kasutamist praktilistes statistilistes arvutustes.

S u m m a r y

Upper and lower probabilities
Numerical results

T.Kinkar

In this paper problems of practical computing of upper and lower probabilities with the help of the formulas introduced by A.P.Dempster (1967) are considered.

In the first part realized computing experiments and achieved numerical results are described.

The paper concludes with some claims about practical usage of these formulas for upper and lower probabilities in statistical computing.

Труды ВЦ ТГУ (1988), 56, стр.36-49.

О распределении векторов коэффициентов главных компонент

Т. Колло
Э. Эхасалу

Ключевые слова: главные компоненты, асимптотическое распределение собственных векторов, эмпирическое распределение собственных векторов.

В статье исследуется распределение векторов коэффициентов главных компонент, найденных по выборочной корреляционной матрице. Сравниваются асимптотическое нормальное распределение координат этих векторов и их эмпирическое распределение, найденное методом Монте-Карло в случае нормально распределенной генеральной совокупности. Результаты моделирования представлены в виде таблиц.

I. Метод главных компонент, обозначения.

Пусть исследуемая совокупность характеризуется p-мерным случайным вектором X, с первыми моментами $EX=\mu$, $DX=\Sigma$ и с корреляционной матрицей P. Тогда p-вектор главных компонент $Z=(Z_1,\dots,Z_p)'$, найденных по корреляционной матрице, определяется равенством

$$X=\nu Z,$$

где матрица ν коэффициентов главных компонент определяется

равенствами

$$P\nu = \nu\Delta;$$
$$\nu'\nu = \Delta;$$

а Δ - диагональная матрица собственных значений $\delta_i (i = 1, \ldots, p)$ матрицы P. При этом $Z_i \perp Z_j$, если $i \neq j$. Итак, в матрице коэффициентов главных компонент ν i-ый столбец ν_i является собственным вектором матрицы P, соответствующим собственному значению δ_i и имеющим длину $\sqrt{\delta_i}$.

Пусть у нас имеется выборка объема n из нашей генеральной совокупности. Обозначим выборочную корреляционную матрицу через R, через D -диагональную матрицу ее собственных векторов, и собственный вектор с длиной $\sqrt{d_i}$, соответствующий собственному значению d_i, через W_i. Тогда выборочная матрица коэффициентов главных компонент $W = (W_1, \ldots, W_p)$ определяется из равенств

$$RW = WD;$$
$$W'W = D.$$

2. Асимптотическое распределение векторов коэффициентов главных компонент.

Пусть первые четыре центральных момента $M_i(X)$ исходного признак-вектора X конечны. Четвертый момент определяем с помощью прямого произведения:

$$M_4(X) = E[(X-\mu)\otimes(X-\mu)'\otimes(X-\mu)\otimes(X-\mu)'].$$

Если собственные значения δ_i матрицы P упорядочены по убыванию: $\delta_i > \delta_{i+1} (i = 1, \ldots, p-1)$, то для векторов коэффициентов главных компонент W_i имеет место сходимость по распределению к нормальному закону, если объем выборки $n \to \infty$

(см., например, [2]):

$$\sqrt{n}(W_i - \nu_i) \xrightarrow{\mathcal{D}} N(0, \Pi^i) \qquad (i = 1, \ldots, p), \tag{I}$$

где

$$\Pi^i = (\Psi^i)' \Gamma' \Phi \Gamma \Psi^i, \tag{2}$$

а

$$\Phi = \overline{M}_4(X) - \mathrm{vec}\,\Sigma (\mathrm{vec}\,\Sigma)'; \tag{3}$$

$$\Gamma = (\Sigma_d^{-\frac{1}{2}} \otimes \Sigma_d^{-\frac{1}{2}}) - \tfrac{1}{2} (I_{p,p})_d [(I_{p^2} \otimes \Sigma_d^{-1} P) + (\Sigma_d^{-1} P \otimes I_{p^2})]; \tag{4}$$

$$\Psi^i = (\nu A^i \Delta^{-1} \nu') \otimes \nu_i. \tag{5}$$

Здесь диагональная матрица A^i определена следующим равенством

$$(A^i)_{jj} = \begin{cases} (\delta_{ij} - \delta_{jj})^{-1} & i \neq j; \\ 0 & i = j, \end{cases} \tag{6}$$

Σ_d обозначает диагональную матрицу с элементами $\delta_{11}, \ldots, \delta_{pp}$ на главной диагонали, а $p^2 \times p^2$ - блок-матрица $I_{p,p}$ составлена из единиц и нулей таким образом, что в (i,j) -ом блоке равняется единице (j,i) -ый элемент, а все остальные элементы нулевые. О матричных понятиях и обозначениях смотри при необходимости в [3] например. В случае нормально распределенной генеральной совокупности матрица $\overline{M}_4(X)$ выражается через ковариационную матрицу Σ:

$$\overline{M}_4(X) = (I_{p^2} + I_{p,p})(\Sigma \otimes \Sigma) + \mathrm{vec}\,\Sigma (\mathrm{vec}\,\Sigma)'$$

и матрица Φ получит следующий вид:

$$\Phi = (I_{p^2} + I_{p,p})(\Sigma \otimes \Sigma). \tag{7}$$

3. Схема моделирования распределения W_i.

Оценивая матрицы P, Σ, Δ, ν и $\overline{M}_4(X)$ из конкретной выборки, получим оценки $\hat{P}, \hat{\Sigma}, \hat{\Delta}, \hat{\nu}$ и $\hat{\Pi}^i$. Считая, что при достаточно большом объеме выборки n вектор $\sqrt{n}(W_i - \nu_i)$ распре-

делен законом $N(0,\Pi^{i})$ приблизительно, имеем

$$W_i \approx N(\nu_i, \tfrac{1}{n}\Pi^{i}) \qquad (i=1,\dots,p).$$

С помощью этого распределения можем построить приближенный доверительный эллипсоид для вектора ν_i при уровне значимости α. Эмпирически изучены маргинальные распределения координат вектора W_i и доверительные границы вектора ν_i :

$$\underline{\nu}_{ij} = \hat{\nu}_{ij} - \sqrt{\tfrac{1}{n}\hat{\Pi}^{i}_{jj}}\,\Phi^{-1}(1-\tfrac{\alpha}{2}) \; ; \qquad (8)$$

$$\overline{\nu}_{ij} = \hat{\nu}_{ij} + \sqrt{\tfrac{1}{n}\hat{\Pi}^{i}_{jj}}\,\Phi^{-1}(1-\tfrac{\alpha}{2}) \; , \qquad (9)$$

где $\Phi^{-1}(x)$ - обратная функция функции распределения закона $N(0,1)$.

Для того, чтобы применять в практике анализа данных доверительные интервалы (8)-(9), необходимо знать, каким образом они зависят от корреляционной матрицы P, объема выборки n и размерности p исходного признак-вектора. Чтобы получить информацию о скорости сходимости (I) и зависимости распределения W_i от параметров, исследуем распределение W_i методом Монте-Карло. В задаче моделирования исходим из нормально распределенной генеральной совокупности с законом $N(\mu,\Sigma)$. Тогда ковариационная матрица Π^{i} задана равенствами (2), (4) -(7).

Процесс моделирования характеризуется следующими этапами.

I) фиксируем параметры μ и Σ;

2) вычисляем теоретическую корреляционную матрицу P, собственные ее векторы ν_i и предельные теоретические ковариационные матрицы $\Pi^{i}\,(i=1,\dots,p)$;

3) генерируем выборку объема n из распределения $N(\mu,\Sigma)$;

4) вычисляем выборочную корреляционную матрицу $\hat{R}$, ее собст-

венные векторы $\hat{\nu}_i$, оценки моментов I-, 2- и 4-го порядка и находим $\hat{\Pi}^{\nu}_{i}\ (i=1,\ldots,\rho)$;

5) повторяем этапы 3) и 4) к раз;

6) построим эмпирическое распределение для координат собственных векторов корреляционной матрицы;

7) проверяем с помощью χ^2-теста согласие между теоретическими асимптотически нормальными и эмпирическими распределениями;

8) вычисляем эмпирические доверительные границы для координат собственных векторов корреляционной матрицы;

9) варьируя параметры распределения и объем выборки n повторяем все предыдущие этапы.

В приведенных ниже примерах моделирования размерность исходного признак-вектора X и количество повторов κ были зафиксированы ($\rho=3$; $\kappa=100$). Размерность ρ не варьировалась из-за резкого увеличения объема рабсты, а относительно κ можно сказать, что предыдущие сравнительные задачи моделирования показали, что все основные тенденции исследуемого случайного явления при $\kappa=100$ вырисовывались в общих чертах достаточно хорошо.

Эмпирическое распределение координат собственных векторов ν_{ij} изучалось при различных объемах выборки от I00 до 5000. При построении доверительных границ использовалась вся информация, накопленная в ходе моделирования. В таблицах I, 2 и 4,5 приведены доверительные границы координат собственных векторов ν_{ij} при уровне значимости $\alpha=0{,}05$. При этом в столбцах под заголовком ТЕОР. представлены доверительные границы, полученные с помощью теоретического асимптотическо-

го нормального распределения, а в столбцах под заголовком ЭМП. представлены доверительные границы, полученные по равенствам

$$\underline{\nu}_{ij} = \tilde{\nu}_{ij} - \sqrt{\frac{1}{n}\tilde{\Pi}^{ii}_{jj}}\,\Phi^{-1}(1-\tfrac{\alpha}{2});$$

$$\overline{\nu}_{ij} = \tilde{\nu}_{ij} + \sqrt{\frac{1}{n}\tilde{\Pi}^{ii}_{jj}}\,\Phi^{-1}(1-\tfrac{\alpha}{2}),$$

где α - уровень значимости и оценки $\tilde{\nu}_{ij}$ и $\tilde{\Pi}^{ii}_{jj}$ получены усреднением оценок $(\hat{\nu}_{ij})_\ell$ и $(\hat{\Pi}^{ii}_{jj})_\ell$, полученных в ℓ-ом повторении цикла моделирования:

$$\tilde{\nu}_{ij} = \frac{1}{K}\sum_{\ell=1}^{K}(\hat{\nu}_{ij})_\ell;$$

$$\tilde{\Pi}^{ii}_{jj} = \frac{1}{K}\sum_{\ell=1}^{K}(\hat{\Pi}^{ii}_{jj})_\ell.$$

Сравнение примеров I и 2 позволяет проследить за зависимостью эмпирического распределения координат собственных векторов корреляционной матрицы от степени зависимости между компонентами исходного признак-вектора: в примере I рассматривается случай слабой зависимости, а в примере 2 случай сильной зависимости.

Пример I (маленькие корреляции).

$$P = \begin{pmatrix} 1 & 0{,}235 & -0{,}005 \\ 0{,}235 & 1 & -0{,}021 \\ -0{,}005 & -0{,}021 & 1 \end{pmatrix} \qquad (10)$$

Собственные значения матрицы P: $\lambda_1 = 1{,}23$; $\lambda_2 = 1{,}00$; $\lambda_3 = 0{,}76$.

Таблица I.

Доверительные границы координат собственных векторов корреляционной матрицы (объем выборки $n = 100$).

j / ν_{ij}	$j=1$		$j=2$		$j=3$	
	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.
$\underline{\nu}_{1j}$	0,707	0,032	0,742	-0,296	-1,090	-0,919
ν_{1j}	0,783		0,786		-0,086	
$\overline{\nu}_{1j}$	0,858	1,394	0,829	1,766	0,919	0,708
$\underline{\nu}_{2j}$	-0,734	-1,520	-0,794	-0,997	0,917	-0,159
ν_{2j}	0,089		0,020		0,995	
$\overline{\nu}_{2j}$	0,912	1,638	0,834	1,120	1,074	1,820
$\underline{\nu}_{3j}$	-0,658	-1,278	0,604	-0,763	-0,591	-0,569
ν_{3j}	-0,616		0,619		0,043	
$\overline{\nu}_{3j}$	-0,575	0,222	0,633	1,869	0,676	0,623

Таблица 2.

Доверительные границы координат собственных векторов корреляционной матрицы (объем выборки $n = 5000$)

j / ν_{ij}	$j=1$		$j=2$		$j=3$	
	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.
$\underline{\nu}_{1j}$	0,772	0,767	0,779	0,774	-0,228	-0,233
ν_{1j}	0,783		0,786		-0,086	
$\overline{\nu}_{1j}$	0,793	0,793	0,792	0,793	0,057	0,048
$\underline{\nu}_{2j}$	-0,028	-0,022	-0,095	-0,091	0,984	0,976
ν_{2j}	0,089		0,020		0,995	
$\overline{\nu}_{2j}$	0,205	0,210	0,135	0,138	1,007	1,006
$\underline{\nu}_{3j}$	-0,622	-0,623	0,617	0,614	-0,047	-0,047
ν_{3j}	-0,616		0,619		0,043	
$\overline{\nu}_{3j}$	-0,610	-0,608	0,621	0,623	0,132	0,134

Для сравнения эмпирического распределения координат собственных векторов корреляционной матрицы с ее асимптотически нормальным распределением разделим числовую прямую на 10 непересекающихся классов, используя стандартное отклонение асимптотического распределения рассматриваемого ν_{ij}. В таблицах 3 и 6 представлены теоретически ожидаемые частоты и полученные эмпирические частоты при разных объемах выборки для координат ν_{i1} первого собственного вектора ν_1.

Из таблицы 3 видно, что эмпирическое распределение координат собственных векторов очень медленно сходится к асимптотическому нормальному распределению.

Таблица 3.

Распределение координат первого собственного вектора корреляционной матрицы.

i	n	Теор.	0	6	9	I5	I9	I9	I5	9	6	0
	I00		I5	9	I3	II	I4	I5	9	II	3	0
I	I000	ЭМП.	5	I4	I0	I3	I9	20	II	6	2	0
	5000		7	II	II	I4	I5	I5	I6	8	3	0
	I00		26	I9	2	I0	I0	5	8	9	6	5
2	I000	ЭМП.	24	II	5	6	20	4	I4	5	9	2
	5000		II	I2	I2	I7	I3	I2	8	6	7	2
	I00		0	0	8	32	20	I2	I0	I7	I	0
3	I000	ЭМП.	0	5	I6	I4	I9	2I	9	I0	5	I
	5000		0	8	I0	20	I8	I8	9	I2	5	0

Пример 2. (большие корреляции).

$$P = \begin{pmatrix} I & 0{,}60I & 0{,}839 \\ 0{,}60I & I & 0{,}706 \\ 0{,}839 & 0{,}706 & I \end{pmatrix}. \qquad (II)$$

Собственные значения матрицы P: $\lambda_1=2{,}44$; $\lambda_2=0{,}42$; $\lambda_3=0{,}I5$.

Таблица 4.

Доверительные границы координат собственных векторов корреляционной матрицы (объем выборки n =100).

i \ j	$j=1$		$j=2$		$j=3$	
	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.
ν_{1j}	0,887	0,887	0,8II	0,8I3	0,922	0,922
ν_{1j}	0,908		0,846		0,947	
$\overline{\nu}_{1j}$	0,929	0,930	0,88I	0,884	0,97I	0,970
$\underline{\nu}_{2j}$	-0,425	-0,422	0,499	0,486	-0,2I9	-0,222
ν_{2j}	-0,350		0,528		-0,I36	
$\overline{\nu}_{2j}$	-0,275	-0,265	0,557	0,55I	-0,053	-0,046
$\underline{\nu}_{3j}$	-0,266	-0,270	-0,I35	-0,I4I	0,280	0,275
ν_{3j}	-0,232		-0,078		0,292	
$\overline{\nu}_{3j}$	-0,I97	-0,I90	-0,02I	-0,0I4	0,305	0,305

Таблица 5.

Доверительные границы координат собственных векторов корреляционной матрицы (объем выборки n =5000).

i \ j	$j=1$		$j=2$		$j=3$	
	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.	ТЕОР.	ЭМП.
$\underline{\nu}_{1j}$	0,905	0,905	0,84I	0,840	0,943	0,943
ν_{1j}	0,908		0,846		0,947	
$\overline{\nu}_{1j}$	0,9II	0,9II	0,85I	0,850	0,950	0,950
$\underline{\nu}_{2j}$	-0,36I	-0,360	0,524	0,524	-0,I48	-0,I49
ν_{2j}	-0,350		0,528		-0,I36	
$\overline{\nu}_{2j}$	-0,339	-0,339	0,532	0,533	-0,I24	-0,I25
$\underline{\nu}_{3j}$	-0,237	-0,237	-0,086	-0,086	0,290	0,29I
ν_{3j}	-0,232		-0,078		0,292	
$\overline{\nu}_{3j}$	-0,227	-0,227	-0,070	-0,070	0,294	0,294

Сравнение таблиц I, 2 и 4, 5 показывает, что в случае

больших корреляций эмпирические доверительные границы намного лучше приближают теоретических, чем в случае маленьких корреляций между координатами исходного признак-вектора.

Таблица 6.

Распределение координат первого собственного вектора корреляционной матрицы.

i	n	ТЕОР.	0	6	9	I5	I9	I9	I5	9	6	0
I	I00	ЭМП.	6	I3	4	I2	I5	8	II	I0	I4	7
	I000		5	I0	8	II	II	I3	II	9	I8	4
	5000		2	9	8	I5	I3	I2	23	8	6	4
2	I00	ЭМП.	8	8	9	9	7	I6	9	8	I5	II
	I000		4	I2	6	II	I3	I4	I2	I2	I4	2
	5000		5	I7	I0	II	9	I0	I3	7	I4	4
3	I00	ЭМП.	0	3	I0	I9	I8	22	20	8	0	0
	I000		0	3	9	I3	27	26	I3	7	2	0
	5000		0	4	7	I8	23	27	I5	5	I	0

Как в случае слабых корреляций (таблица 3), так и в данном примере тенденция сходимости эмпирического распределения координат собственного вектора к нормальному закону слабо заметна. Отсюда следует рекомендация не пользоваться предельным нормальным распределением собственных векторов для построения их доверительных областей.

Вероятностная характеристика собственных векторов матрицы R усложнена тем, что если исходный вектор X имеет коррелированные координаты ($P \neq I$), то собственные векторы матрицы R распределены асимптотически нормально. В то же время в случае некоррелированности координат вектора X ($P = I$) собственные векторы матрицы R не описываются нормальным распределением, так как не выполняется предположение о сходимости ко-

вариационной матрицы S к нормальному закону. Известно (см. [I], § I3.3), что в случае нормально распределенной генеральной совокупности собственные векторы выборочной ковариационной матрицы имеют условное инвариантное по Хаару распределение. Как ведут себя координаты собственных векторов выборочной корреляционной матрицы R, увидим на примере ее первого собственного вектора в следующей таблице 7.

Таблица 7

Эмпирическое распределение координат вектора ($P=I$).

ν_{i1}	классы / n	-0,8	-0,8 -0,6	-0,6 -0,4	-0,4 -0,2	-0,2 0	0 0,2	0,2 0,4	0,4 0,6	0,6 0,8	0,8
ν_{11}	I00	I	I9	6	I	4	I	6	I6	43	3
	I000	0	I4	2	3	5	4	4	23	45	0
	5000	0	I2	3	5	6	7	4	I5	48	0
ν_{21}	I00	0	I4	8	I	5	5	8	I3	42	4
	I000	I	I5	I3	6	4	7	9	I2	33	0
	5000	0	20	II	4	6	6	6	I9	28	0
ν_{31}	I00	I	2I	I4	6	5	2	5	I0	33	3
	I000	0	I5	I5	5	5	2	4	II	43	0
	5000	0	I4	I4	3	4	2	I0	I3	40	0

Из таблицы видно, что при теоретическом собственном векторе $\nu_1 = (1\ 0\ 0)'$ эмпирическое распределение координат собственного вектора выборочной корреляционной матрицы бимодально с экстремумами в $[-0,8;-0,6)$ и $[0,6;0,8)$. При этом глобальный максимум достигается в классе $[0,6;0,8)$. Такой ситуацией, видимо, объясняется медленная сходимость координат собственных векторов матрицы R к нормальному закону при малых значениях корреляции в матрице R (пример I). Пока остается необъяснимой причина медленной сходимости координат собственных векторов R, если корреляции в матрице R являются большими.

Литература

I. Андерсон Т. Введение в многомерный статистический анализ. – М., 1958.

2. Колло Т. О применении асимптотических распределений в модели главных компонент. II Всесоюзная научно-техническая конференция "Применение многомерного статистического анализа в экономике и оценке качества продукции". Тезисы докладов. Тарту, 1981, с.282-283.

3. Колло Т., Кинкар Т. Матричная производная с применением для блок-матриц. Труды ВЦ ТГУ, 51, Тарту, 1984, с. 98-107.

Поступило 09.08.1988

Resümee

Peakomponentide kordajavektorite jaotusest

T. Kollo, E. Ehasalu

Artiklis uuritakse peakomponentide kordajavektorite jaotust, kui peakomponentide meetodi aluseks on korrelatsioonimaatriks. Kui teoreetilisel korrelatsioonimaatriksil P on mittekordsed omaväärtused ja üldkogumi jaotusel eksisteerib neljandat järku tsentraalne moment $M_4(X_i) < \infty$, siis omavektorite $W_i\,(i = 1,2,\dots,p)$ jaoks kehtib koonduvus (1). Koonduvuse (1) kiirust uuriti empiiriliselt normaaljaotusega valimi korral: $X_i \sim N(\mu, \Sigma)$. Modelleerimine teostati juhul $p=3$ (X_i – p-vektorid) paralleelselt kahe erineva teoreetilise korrelatsioonimaatriksiga: maatriks (10) – väikesed korrelatsioonid, (11) – suured korrelatsioonid. Peakomponentide kordajavektorite koordinaatide empiirilist jaotust võrreldi teoreetilise asümptootilise normaaljaotusega. Võrdlustulemused on esitatud esimese kordajavektori γ_1 koordinaatide kohta tabelites 3 (väikesed korrelatsioonid) ja 6

(suured korrelatsioonid). Tabelitest 3 ja 6 nähtub, et koonduvus (1) on väga aeglane ning seega asümptootiline normaaljaotus pole kasutatav kordajavektorite v_i usalduspiiride(8)-(9) leidmisel. Tabelites 1, 2 (väikesed korrelatsioonid) ja 4, 5 (suured korrelatsioonid) on esitatud kordajavektorite usalduspiirid koordinaatide kaupa erinevate valimimahtude n jaoks (tabelites 1 ja 4 n =100, tabelites 2 ja 5 n =5000) Asümptootiline normaaljaotus peakomponentide kordajavektorite jaoks pole kasutatav juhul $P = I_p$. Tabelis 7 on esitatud esimese omavektori koordinaatide empiiriline jaotus p=3 korral.

S u m m a r y

On the distribution of vectors of principal component coefficients

T. Kollo, E. Ehasalu

The distribution of vectors of principal component coefficients is studied when the principal component model is based on the correlation matrix. When the correlation matrix P has nonmultiple roots, the convergence (1) takes place for the sample vectors $W_i (i = 1, \dots, p)$ of principal component coefficients, if there exists the fourth central moment $M_4(X) < \infty$ of the population distribution. The speed of the convergence (1) has been studied empirically in the normal distribution case: $X_i \sim N(\mu, \Sigma)$. The simulation experiment was carried out in the three-dimensional case (X_i - are p -vectors) parallelly with two different theoretical correlation matrices defined by equalities (10) (low correlation) and (11) (high correlation case). The empirical distribution of coordinates of vectors of principal component coefficients was compared with theoretical asymptotic normal distribution. Simulation results are presented in Tables 3 (low correlation case) and 6 (high correlation case) for the coordinates of the first principal component vector v_1 .

From Tables 3 and 6 it follows that the convergence (1) is very slow and it is not reasonable to use the asymptotics normal distribution for constructing the approximate confi-

dence limits (8)-(9). In the Tables 1, 2 (low correlation case) empirical confidence limits for coordinates of the principal component vectors γ_i are presented. For different values of the sample size n (in Tables 1, 4 $n=100$, in Tables 2, 5 $n=5000$). The asymptotic normal distribution is not valid for principal component vectors in the case $P=I_p$. For the three-dimensional sample ($p=3$) the empirical distribution of coordinates of the first principal component vector is presented in Table 7.

Труды ВЦ ТГУ (1988), 56, стр.50-58.

G-оценка расстояния Махаланобиса для случая произвольного непрерывного распределения наблюдаемых векторов

Т. Павленко

Ключевые слова: расстояние Махаланобиса, G - оценка, состоятельность, асимптотическая нормальность

Настоящая работа посвящена изучению свойств G- оценки расстояния Махаланобиса для случая двух многомерных генеральных совокупностей с произвольными непрерывными распределениями. Вопросы применения методов общего статистического анализа (G - анализа) к построению оценок некоторых статистик многомерного статистического анализа рассматривались в работах [I]-[4] .

В [4] построена G - оценка расстояния Махаланобиса для случая двух многомерных генеральных совокупностей и доказана ее асимптотическая нормальность. Однако условие нормальности наблюдений накладывает значительные ограничения на использование статистики Махаланобиса при решении практических задач. Как известно, на практике вид распределения наблюдаемых случайных векторов чаще всего неопределен. В данной статье показано, что при некоторых дополнительных условиях G - оценка расстояния Махаланобиса является состоятельной и асимпто-

тически нормальной в случае произвольного непрерывного распределения наблюдений.

Пусть ξ и η - независимые m-мерные случайные векторы с плотностями распределений $p_1(x)$ и $p_2(x)$ соответственно, $x_1, \dots, x_{n_1}$; $y_1, \dots, y_{n_2}$ - наблюдения над ξ и η. Под расстоянием Махаланобиса будем понимать выражение

$$d = (a_1 - a_2)' R^{-1} (a_1 - a_2),$$

где a_1 и a_2 - векторы математических ожиданий, R - ковариационная матрица. В качестве оценок векторов средних и ковариационной матрицы возьмем

$$\hat{a}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i \ , \quad \hat{a}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i \ ,$$

$$\hat{R} = \frac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \hat{a}_1)(x_i - \hat{a}_1)' + \sum_{j=1}^{n_2} (y_j - \hat{a}_2)(y_j - \hat{a}_2)' \right\}.$$

Будем предполагать, что выполняется следующее соотношение между размерностью наблюдаемых векторов и количеством наблюдений:

$$\lim_{n_i \to \infty} \frac{m}{n_i} = c_i \ , \quad 0 < c_i < 1 \ , \ i = 1, 2. \qquad (I)$$

Условие (I) называется G -условием. G -оценкой расстояния Махаланобиса называется выражение

$$G_d = (\hat{a}_1 - \hat{a}_2)' \hat{R}^{-1} (\hat{a}_1 - \hat{a}_2) \frac{n_1 + n_2 - m - 2}{n_1 + n_2 - 2} - \frac{m}{n_1} - \frac{m}{n_2}$$

Докажем следующую теорему.

Теорема I.

Пусть выполняется G -условие

$$\lim_{n_i \to \infty} \frac{m}{n_i} = c_i \ , \quad 0 < c_i < 1, \ i = 1, 2$$

и R - невырожденная матрица. Тогда существует такая константа c , что

$$\operatorname{plim}_{n_i \to \infty} \frac{1}{c}\left\{G_a - (a_1 - a_2)' R^{-1} (a_1 - a_2)\right\} = 0.$$

Доказательство.

Известно ([4], стр. 251), что

$$\hat{a}_1 \approx a_1 + \frac{1}{\sqrt{n_1}} \sqrt{R}\, \nu_1, \quad \hat{a}_2 \approx a_2 + \frac{1}{\sqrt{n_2}} \sqrt{R}\, \nu_2,$$

где ν_1 и ν_2 - независимые $N(0,I)$ распределенные случайные векторы. Знак $\approx$ означает совпадение распределений. Введем обозначения $x_{n_1+1} = y_1, \ldots, x_{n_1+n_2} = y_{n_2}$. Тогда эмпирическая ковариационная матрица R может быть представлена в виде

$$\hat{R} = \sqrt{R}\, \frac{\sum_{k=1}^{n_1+n_2-2} z_k z_k'}{n_1+n_2-2} \sqrt{R},$$

где

$$z_k = \sum_{i=1}^{n_1+n_2} h_{ik} \tilde{x}_i, \quad \tilde{x}_i = R^{-\frac{1}{2}}(x_i - M x_i),$$

h_{ik} - элементы вещественной ортогональной матрицы соответствующей размерности, причем $h_{i\, n_1+n_2} = (n_1+n_2)^{-\frac{1}{2}}$.

Пусть $R^{-\frac{1}{2}}(\hat{a}_1 - \hat{a}_2) = b$. Распределение матрицы $\left(\frac{\sum_{k=1}^{n_1+n_2-2} z_k z_k'}{n_1+n_2-2}\right)^{-1}$ инвариантно относительно ортогонального преобразования S. Выберем матрицу S таким образом, чтобы

$$Sb = \begin{pmatrix} \sqrt{\frac{(b,b)}{n_1+n_2-2}} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

тогда

$$(\hat{a}_1 - \hat{a}_2)' \hat{R}^{-1} (\hat{a}_1 - \hat{a}_2) \approx \left(\sum_{k=1}^{n_1+n_2-2} z_k z_k'\right)^{-1}_{11} (\hat{a}_1 - \hat{a}_2) R^{-1} (\hat{a}_1 - \hat{a}_2).$$

Далее будем также использовать формулу

$$\sum_{i=1}^{n} a_i \nu_i \approx \nu_j \sqrt{\sum_{i=1}^{n} a_i^2}, \qquad (2)$$

j может принимать любое значение в промежутке от I до n, $\nu_j \sim N(0,1)$. Отсюда следует, что индекс при ν в дальнейшем можно опустить. Теперь легко показать, что

$$(\hat{a}_1 - \hat{a}_2)' \hat{R}^{-1} (\hat{a}_1 - \hat{a}_2) \approx \Big(\sum_{k=1}^{n_1+n_2-2} z_k z_k' \Big)^{-1}_{11} \Big\{ (a_1 - a_2)' R^{-1} \times$$

$$\times (a_1 - a_2) + 2(a_1 - a_2)' R^{-\frac{1}{2}} \nu \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} + \Big(\frac{1}{n_1} + \frac{1}{n_2} \Big) (\nu, \nu) \Big\}.$$

Пусть T - ортогональная матрица собственных векторов матрицы K:

$$K = cd' + dc', \quad c = R^{-\frac{1}{2}} a, \quad d = R^{-\frac{1}{2}} b;$$

$$a = (S_{ip}, i = 1, \dots, m), \quad b = (S_{is}, i = 1, \dots, m); \quad p, s = 1, \dots, n_1 + n_2 - 2.$$

Рассмотрим ортогональное преобразование

$$B = Z'T, \quad Z = (z_1, \dots, z_{n_1+n_2-2}).$$

Тогда

$$(\hat{a}_1 - \hat{a}_2)' \hat{R}^{-1} (\hat{a}_1 - \hat{a}_2) \approx (BB')^{-1}_{11} \Big\{ (a_1 - a_2)' R^{-1} (a_1 - a_2) +$$

$$+ 2(a_1 - a_2)' R^{-\frac{1}{2}} \nu \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} + \Big(\frac{1}{n_1} + \frac{1}{n_2} \Big) (\nu, \nu) \Big\}.$$

Поскольку

$$p \lim_{n_i \to \infty} \frac{1}{\mu} (BB')^{-1}_{11} = 1, \tag{3}$$

где

$$\mu = M (BB')^{-1}_{11},$$

$$p \lim_{n_i \to \infty} 2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} (a_1 - a_2)' R^{-\frac{1}{2}} \nu = 0,$$

$$p \lim_{n_i \to \infty} (\nu, \nu) \Big(\frac{1}{n_1} + \frac{1}{n_2} \Big) = \frac{m}{n_1} + \frac{m}{n_2}.$$

Очевидно, что имеет место утверждение теоремы.

Покажем теперь, что оценка G_α является асимптотически нормальной.

Теорема 2.

Пусть выполняется G - условие

$$\lim_{n_i \to \infty} \frac{m}{n_i} = c_i, \quad 0 < c_i < 1, \quad i = 1, 2,$$

существует плотность распределения компонент векторов ξ и η,

$$M\tilde{x}_{ik}^4 = 3, \quad i = 1, \dots, m; \quad k = 1, \dots, n_1 + n_2,$$

для некоторого $\delta > 0$

$$\sup_{n_1+n_2} \sup_{i=1,\dots,m} M|\tilde{x}_{ik}| \leq \delta < \infty.$$

R - невырожденная матрица и

$$\lim_{n_i \to \infty} (a_1 - a_2)' R^{-1} (a_1 - a_2) \left[\frac{1}{n_1} + \frac{1}{n_2}\right] = 0.$$

Тогда

$$\lim_{n_i \to \infty} P\left\{\frac{G_\alpha - (a_1 - a_2)' R^{-1} (a_1 - a_2)}{\sqrt{D}} \sqrt{n_1 + n_2 - m - 2} < x\right\} =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy,$$

где

$$D = \alpha^2 + 2\mu^2 \frac{n_1 + n_2 - m - 2}{m} \left(\frac{m}{n_1} + \frac{m}{n_2}\right)^2 +$$

$$+ 4\mu^2 (a_1 - a_2)' R^{-1} (a_1 - a_2) \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 + n_2 - m - 2)^3}{(n_1 + n_2 - 2)^2}.$$

Доказательство.

Из теоремы I следует, что имеет место следующее представление:

$$(G_\alpha - \alpha)\sqrt{n_1+n_2-m-2} \approx \left\{\left(\frac{n_1+n_2-m-2}{n_1+n_2-2}(BB')^{-1}_{11} - 1\right)\alpha + \right.$$

$$+ \left[2(a_1-a_2)'R^{-\frac{1}{2}}\nu\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} + \nu'\nu\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right] \times \qquad (4)$$

$$\left. \times (BB')^{-1}_{11}\frac{n_1+n_2-m-2}{n_1+n_2-2} - \frac{m}{n_1} - \frac{m}{n_2}\right\}\sqrt{n_1+n_2-m-2}.$$

В[4], стр. 215 показано, что при выполнении условий теоремы 2 величина

$$\left[\frac{n_1+n_2-m-2}{n_1+n_2-2}(BB')^{-1}_{11} - 1\right]\sqrt{(n_1+n_2)\frac{n_1+n_2-m-2}{n_1+n_2-2}}$$

распределена по стандартному нормальному закону. Преобразовав первое слагаемое в выражении (4), получим

$$\sqrt{n_1+n_2-m-2}\left[\frac{n_1+n_2-m-2}{n_1+n_2-2}(BB')^{-1}_{11} - 1\right] \times$$

$$\times\sqrt{(n_1+n_2)\frac{n_1+n_2-m-2}{n_1+n_2-2}}\,\alpha\sqrt{\frac{1}{n_1+n_2}\,\frac{n_1+n_2-2}{n_1+n_2-m-2}} \approx \eta_1\alpha,$$

где $\eta_1 \sim N(0,1)$. Заметим также, что

$$\sqrt{n_1+n_2-m-2}\,(\nu,\nu)\left(\frac{1}{n_1}+\frac{1}{n_2}\right) = \frac{(\nu,\nu)}{m}\left(\frac{m}{n_1}+\frac{m}{n_2}\right)\sqrt{n_1+n_2-m-2} =$$

$$= \sum_{i=1}^{m}\frac{\nu_i^2-1}{\sqrt{m}}\sqrt{\frac{n_1+n_2-m-2}{m}}\left(\frac{m}{n_1}+\frac{m}{n_2}\right) +$$

$$+\left(\frac{m}{n_1}+\frac{m}{n_2}\right)\sqrt{n_1+n_2-m-2} \approx \eta_2\sqrt{2}\sqrt{\frac{n_1+n_2-m-2}{m}} \times$$

$$\times\left(\frac{m}{n_1}+\frac{m}{n_2}\right) + \left(\frac{m}{n_1}+\frac{m}{n_2}\right)\sqrt{n_1+n_2-m-2},$$

где ν_i - компоненты вектора ν, $\eta_2 \sim N(0,1)$ - распределенная случайная величина, не зависящая от η_1.

Преобразуем выражение (4), используя формулы (2) и (3):

$$(G_\alpha - \alpha)\sqrt{n_1+n_2-m-2} \approx \eta_1\alpha + \eta_2\sqrt{2}\,\mu\sqrt{\frac{n_1+n_2-m-2}{n_1+n_2-2}} \times$$

$$\times\left(\frac{m}{n_1}+\frac{m}{n_2}\right) + 2\mu\,(a_1-a_2)'R^{-\frac{1}{2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,\frac{n_1+n_2-m-2}{n_1+n_2-2} \times$$

$$\times\sqrt{n_1+n_2-m-2} \approx$$

$$\approx \eta_1\left[\alpha^2 + 2\mu^2\,\frac{n_1+n_2-m-2}{m}\left(\frac{m}{n_1}+\frac{m}{n_2}\right)^2 + \right.$$

$$\left. + 4\mu^2(a_1-a_2)'R^{-1}(a_1-a_2)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{(n_1+n_2-m-2)^3}{(n_1+n_2-2)^2}\right]^{\frac{1}{2}}$$

Таким образом,

$$\frac{(G_\alpha-\alpha)}{\sqrt{D}}\sqrt{n_1+n_2-m-2} \approx \eta_1,$$

что и доказывает утверждение теоремы 2.

Литература

I. Гирко В.Л. "Борьба с размерностью" в многомерном статистическом анализе. Тезисы третьей Всесоюзной научно-технической конференции "Применение многомерного статистического анализа в экономике и оценке качества продукции", Тарту, 1985, с. 43-52.

2. Гирко В.Л. G - анализ наблюдений большой размерности. Вычислительная и прикладная математика. 1986, вып.60, с.

I5-2I.

3. Гирко В.Л. Введение в общий статистический анализ. Теория вероятностей и ее применения. I987, 32, вып.2, с. 252-265.
4. Гирко В.Л. Многомерный статистический анализ. Киев. Выща школа. I988.
5. Фукунага К. Введение в статистическую теорию распознавания образов. М.: Наука. I979.
6. Андерсон Т. Введение в многомерный статистический анализ. М.: Физматгиз. I963.
7. Деев А.Д. Представление статистик дискриминантного анализа и асимптотические разложения при размерности пространства, сравнимой с объемом выборок. Докл. АН СССР. I970.№4, I95, с. 759-762.
8. Мешалкин Л.Д., Сердобольский В.И. Ошибки при классификации многомерных наблюдений. Теория вероятностей и ее применения. I978,23,вып.4, с.772-78I.

Поступило 29.06.I988

R e s ü m e e

Mahalanobise kauguse G-hinnang suvalise pideva lähtejaotuse korral

T. Pavlenko

Käesolevas artiklis uuritakse V.Girko poolt (vt. [1]-[ ]) kasutusele võetud G-hinnangut rakendatuna kahe mitmemõõtmelise üldkogumi vahelise Mahalanobise kauguse jaoks.

Hinnang on esitatud kujul

$$G=(\hat{a}_1-\hat{a}_2)'\hat{R}^{-1}(\hat{a}_1-\hat{a}_2)\frac{n_1+n_2-m-2}{n_1+n_2-2}-\frac{m}{n_1}-\frac{m}{n_2},$$

kus a_1 ja a_2 on valimikeskmised ja R - nende ühine, mittekõdunud kovariatsioonimaatriks. Eeldatakse, et kogumid on pide-

va jaotusega ja dimensioon m kasvab koos valimimahtude kasvuga nõnda, et oleks rahuldatud G-tingimus (1).

Neil eeldustel tõestatakse teoreemis 1, et G on mõjus hinnang.

Teise teoreemi tõestamisel tehakse mõningad lisaeeldused, nimelt nõutakse, et kõigi uuritavate tunnuste 4. momendid võrduksid 3-ga. Siis on kaugusehinnang G asümptootiliselt normaaljaotusega. See tulemus üldistab monograafias [4] esitatud tulemust, kus G asümptootiline normaalsus on tõestatud eeldusel, et lähtekogumid on mitmemõõtmelise normaaljaotusega.

S u m m a r y

G-estimation of Mahalanobis distance between two population having arbitrary continuous distribution

T.Pavlenko

In the paper the G-estimation defined by V. Girko [1] - [4], is used for Mahalanobis distance between two multivariate populations in the following form

$$G = (\hat{a}_1 - \hat{a}_2)' \hat{R}^{-1} (\hat{a}_1 - \hat{a}_2) \frac{n_1 + n_2 - m - 2}{n_1 + n_2 - 2} - \frac{m}{n_1} - \frac{m}{n_2},$$

where a_1 and a_2 are the estimations of means and R the non - degenerated covariance matrix, common for both populations.

The distributions of populations are continuous and the dimension m is increasing with sample sizes n_1 and n_2, providing the G-condition (1) remains fulfilled.

In the theorem 1 the consistency of the estimation G is proved.

In the second theorem some additional assumptions (the equality to 3 of all 4th moments) are made and the asymptotical normality of the distribution of G is proved. The result generalizes the theorem, proved in the monograph [4] for the special case when the initial distribution is multivariate normal.

Труды ВЦ ТГУ (1988), 56, стр.59-87.

Показатели зависимости двух признаков. Двумерные стандартные распределения для изучения и тестирования зависимостей

Э. Тийт
М. Унт

Ключевые слова: статистическая зависимость, коэффициент сопряженности, коэффициент контингентности, монотонная зависимость, корреляция, тестирование матобеспечения по статистике.

I. Постановка проблемы. Одним из наиболее часто применяемых методов (первичного) статистического анализа эмпирического материала является измерение степени зависимости между двумя признаками.

Существует большое количество различных показателей зависимости r_1, r_2, ..., r_s, ... Как правило, в каждом конкретном случае выбирается некоторое подмножество из них (учитывая цели исследования, особенности исходного материала, технические возможности, имеющееся математическое обеспечение), и, таким образом, для характеризации исследуемой пары признаков X,Y (двумерного эмпирического распределения P_{xy}) получается некое множество $\{r_1, \ldots, r_m\}$ коэффициентов зависимости. В общем случае отдельные элементы этого мно-

жества заметно отличаются друг от друга.

Для практической работы целесообразно выяснить:

1^0 Какие неравенства между коэффициентами функциональны, следовательно неравенства $\nu_i \le \nu_j \quad (i,j = 1,2,\dots)$ имеют место всегда (независимо от характера P_{xy}).

2^0 Какие неравенства $\nu_i \le \nu_j \quad (i,j = 1,2,\dots)$ имеют место при известном характере распределения P_{xy}; из таких неравенств возможно сделать содержательные выводы.

Учитывая лаконичность руководств программных средств, часто необходимо

3^0 выяснить точное определение (формулу) каждого коэффициента $\nu_i \ (i = 1,2,\dots)$.

4^0 тестировать точность и правильность вычисления коэффициентов $\nu_i \ (i = 1,2,\dots)$.

При решении всех этих задач целесообразно пользоваться некоторыми малопараметрическими семействами тестовых распределений, для которых значения наиболее часто применяемых показателей зависимости выражаются через параметры распределения. Тогда имеется возможность

1) Сравнивать значения коэффициентов аналитически; тем самым решая задачи 1^0 и 2^0;

2) Вычислить точные значения коэффициентов ν_i для многих комплектов параметров с тем, чтобы решать задачи 3^0 и 4^0.

В настоящей статье с этой целью строится семейство смесей дискретных равномерных распределений, описываемое в пункте 6; для них в пункте 7 выводятся выражения всех показателей зависимости, описываемых в п 4.

2. Типы зависимости между двумя признаками.

Предположим, что X и Y – два признака, имеющие сов-

местное эмпирическое распределение P_{xy}. Проблемы, связанные с исходным теоретическим распределением, в данной статье не рассматриваются.

Распределения X и Y – это маргинальные распределения P_x и P_y. Объем рассматриваемой совокупности (выборки)–n, числа разных значений (классов) признаков X и Y равняются соответственно k и h, их значения соответсвенно $a_1, \ldots, a_k$ и $b_1, \ldots, b_h$; частоты соответствующих значений обозначаются соответственно через n_{ij}, $n_{i\cdot}$ и $n_{\cdot j}$ ($i=1, \ldots,k$; $j=1,\ldots,h$). Множество всевозможных дискретных признаков обозначаем через $\mathfrak{D}$; если множество значений имеет содержательную упорядоченность, то признак называется порядковым; множество порядковых признаков обозначается через $\mathfrak{B}$ и множество количественных (числовых) признаков через $\mathfrak{A}$; (заметим, что $\mathfrak{A} \subset \mathfrak{B} \subset \mathfrak{D}$).

Мы рассмотрим следующие типы зависимости:

I. Статистическая зависимость для пары X,Y из $\mathfrak{D}\times\mathfrak{D}$.

II. Монотонная зависимость для пары X,Y из $\mathfrak{B}\times\mathfrak{B}$.

III. Регрессионная зависимость для пары X,Y из $\mathfrak{A}\times\mathfrak{A}$.

IV. Корреляционная зависимость для пары X, Y из $\mathfrak{A}\times\mathfrak{A}$.

Если тип зависимости не определен, то мы говорим о H-зависимости.

А. Направленность зависимости.

При изучении зависимости между признаками в основу взята прогнозируемость одного признака через другой; учитывая это следует, как правило, прежде всего зафиксировать направление H-зависимости.

Если исследуется прогнозирование признака Y по признаку X (в смысле H-зависимости, символически $X \Rightarrow Y$), то соответствующие коэффициенты обозначаются через $\nu(Y/X)$ (перед на-

клонной чертой - прогнозируемый, за ней - аргумент-признак).

Ненаправленную Н-зависимость определяют симметричные относительно признаков X и Y коэффициенты, обозначаемые через $\nu(X,Y)$, $\nu(X,Y)=\nu(Y,X)$.

Для рассуждений, имеющих место как для направленных, так и для ненаправленных зависимостей, пользуемся символом $\nu(X,Y)=\nu(\cdot)$.

Б. Теснота (сила) зависимости.

Для каждого типа зависимости Н фиксируем следующие экстремальные ситуации:

1^0 Полная Н-зависимость (признака Y от признака X), которое имеет место в случае, когда признак Y полностью прогнозируется по признаку X (или наоборот).

2^0 Полная Н-независимость (признака Y от признака X), при которой знание конкретных значений признака X не дает никакой дополнительной информации для признака Y.

Измерение Н-зависимости признака Y от признака X при помощи некоторого коэффициента $\nu(Y/X)$ состоит в следующем:

Фиксируем значение ν_* коэффициента $\nu(Y/X)$, характеризующее экстремальную ситуацию 1^0, а также и ν_0, характеризующее ситуацию 2^0.

Для данной конкретной пары признаков X, Y вычисляется значение $\nu(X,Y)$, которое должно удовлетворять условиям:

$$\nu_0 \leq \nu(X,Y) \leq \nu_* . \qquad (I)$$

Таким образом для всевозможных пар признаков (X,Y) определена упорядоченность, притом, если

$$\nu(X,Y) < \nu(W,Z),$$

то говорят, что Н-зависимость между признаками W и Z силь-

нее, чем между признаками X и Y.

В. Знак зависимости.

Если $X, Y \in \mathfrak{B} \times \mathfrak{B}$, то следует говорить о знаке (направления) зависимости. Уточняем понятие полной H-зависимости 1^0 в таком случае:

3^0 Если между признаками X и Y имеется полная H-зависимость, притом они имеют одностороннюю упорядоченность, то зависимость называется положительной (возрастающей); соответствующую ситуацию обозначает r_+.

4^0 Если между признаками X и Y имеется полная H-зависимость, притом они упорядочены противоположно (с "увеличением" одного признака другой, как правило, "уменьшается"), то зависимость называется отрицательной (убывающей), соответствующую ситуацию обозначает r_-.

В таком случае вместо соотношения (I) имеет место следующее:

$$r_- \leq r(X,Y) \leq r_+, \tag{2}$$

притом, если

$$r_- \leq r(X,Y) < r_0, \tag{3}$$

H-зависимость называется отрицательной, и если

$$r_0 < r(X,Y) \leq r_+, \tag{4}$$

H-зависимость - положительна.

3. Определение основных типов зависимости на основании эмпирического распределения.

I. Статистическая зависимость.

Признак Y зависит полностью от признака X (в смысле статистической зависимости), если для каждого индекса i $(i = 1, \dots, k)$ найдется индекс $j = j(i)$ $(j \in [1, h])$ так, что

$$n_{ij} = 0, \text{ если } j \neq j(i), \quad j = 1, \dots, h; \; i = 1, \dots, k. \tag{5}$$

Из таблицы (см. рис. I) видно, что для этого необходимо, чтобы выполнялось условие $h \leq k$.

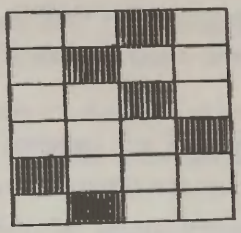

Рис. I.

Таблица распределения. Полная статистическая зависимость Y от X

$\square$ ----- $n_{ij} = 0$,

$\boxplus$ ----- $n_{ij} \neq 0$.

Соответствие между признаками X и Y в таком случае однозначное, и это соответствует, например, получению признака (Y) путем дискретизации или сжатия шкалы из исходного признака (X).

Если наряду с условием (5) выполняется и условие (5'): для каждого индекса j найдется индекс $i(j)$ так, что

$$n_{ij} = 0 \quad , \quad i \neq i(j) \, , \quad i = 1, 2, \ldots k \, ; \quad j = 1, \ldots, h \, , \qquad (5')$$

то между признаками X и Y полная взаимная статистическая зависимость, они во взаимно однозначном соответствии (в статистическом смысле эквивалентны).

Признаки X и Y в статистическом смысле независимы (это свойство всегда взаимное), если имеет место равенство

$$n_{ij} = \frac{n_{i\cdot} \, n_{\cdot j}}{n} \, , \quad i = 1, \ldots, k \, ; \quad j = 1, \ldots, h \, . \qquad (6)$$

II. Полная положительная монотонная зависимость признака Y от признака X имеет место в таком случае, когда имеет мес-

то соотношение (7):

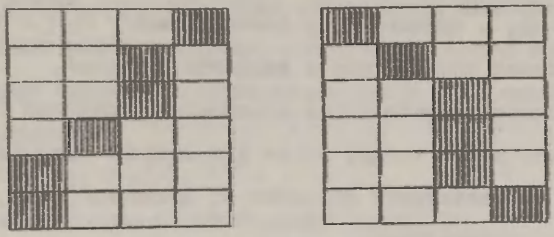

Рис. 2.

Полная монотонная зависимость Y от X.

$$y_i < y_j \Rightarrow x_i < x_j \quad (i,j = 1, \ldots, n), \tag{7}$$

(см. рис. 2, слева). Если же имеет место соотношение

$$y_i < y_j \Rightarrow x_i > x_j \quad (i,j = 1, \ldots, n), \tag{7'}$$

то зависимость монотонная отрицательная (см. рис. 2, справа).

Заметим, что в случае полной монотонной зависимости вариационные ряды обоих признаков совпадают, но у прогнозируемого признака может быть больше повторяющихся значений.

Если наряду с соотношением (6) имеет место еще соотношение

$$x_i < x_j \Rightarrow y_i < y_j, \quad (i,j = 1, \ldots, n), \tag{7''}$$

то имеет место взаимная полная монотонная положительная зависимость; пользуясь неравенством (7') легко определяется и взаимная полная монотонная отрицательная зависимость.

Взаимно монотонные признаки имеют совпадающие вариационные ряды; они эквивалентны как в статистическом смысле, так и по упорядоченности.

Признак Y вполне монотонно независимый от признака X, если имеет место равенство

$$P((x_j - x_i) \wedge (y_j < y_i)) = P((x_j < x_i) \wedge (y_j > y_i)), \tag{8}$$

$i,j = 1, \ldots, n$

и вероятность определяется по совместному эмпирическому распределению f_{xy}, описываемому соотношением $P_{xy}(a_i, b_j) = \frac{n_{ij}}{n}$.

Монотонная независимость является взаимной.

III. Полная регрессионная зависимость признака Y от признака X имеет место тогда, когда для каждого значения a_i признака X соответствует значение b_j признака Y так, чтобы имело место соответствие:

$$X = a_i \Rightarrow Y = b_j \qquad (i = 1, \dots, k;\ j = 1, \dots, h), \tag{9}$$

притом учитывается, что $h \geq 2$. и $n_{j\cdot} \neq 0$, $j = 1, \dots, h$.

Заметим, что формально полная регрессионная зависимость совпадает с полной статистической зависимостью.

Если, кроме того, имеет место и соотношение

$$Y = b_j \Rightarrow X = a_i \quad (j = 1, \dots, h;\ i = 1, \dots, k), \tag{9'}$$

то между признаками X и Y имеет место взаимная полная регрессионная зависимость.

Прогноз в смысле регрессионной зависимости определяется через равенство

$$\tilde{y}_j := E(Y/X = x_j), \qquad j = 1, \dots, n, \tag{10}$$

где $\tilde{y}_j$ – прогноз признака Y при j-ом наблюдений, а $E(Y/X=a)$ – условное математическое ожидание признака Y при условии, что признак X имеет значение a.

Признак Y является независимым от признака X в смысле регрессионной зависимости, если имеет место равенство

$$E(Y/X = a_i) = \text{const} \qquad (i = 1, \dots, k). \tag{11}$$

Следовательно, из статистической независимости вытекает регрессионная независимость, но противоположное следствие, в общем, не имеет место.

Регрессионная независимость, в общем, не является взаимной.

IУ Полная корреляционная зависимость признака Y от признака X имеет место в случае, когда имеет место соотношение (9), притом существуют постоянные α и β так, чтобы имеет место равенство

$$b_j = \alpha a_i + \beta,$$

или, другими словами, имеет место равенство

$$y_j = \alpha x_j + \beta \qquad (j = 1, \ldots, n). \qquad (I2)$$

Из формулы (I2) вытекает, что полная корреляционная зависимость взаимная, из определения следует, что полная корреляционная зависимость влечет за собой и полную регрессионную зависимость, и, следовательно, полную статистическую зависимость.

Кроме того, из (I2) следует, что если $\alpha > 0$, то корреляционная зависимость положительная и из нее вытекает положительная монотонная зависимость; если $\alpha < 0$, то корреляционная зависимость отрицательная и из нее вытекает отрицательная монотонная зависимость.

Полная корреляционная независимость (некоррелированность) признака Y от признака X имеет место тогда, когда вместе с условием (II) имеет место и

$$E(X/Y = b_j) = const \qquad (j = 1, \ldots, h), \qquad (II')$$

значит, некоррелированность совпадает с взаимной независимостью в регрессионном смысле. Следовательно, некоррелированность взаимная.

4. Коэффициенты, измеряющие зависимость.

В настоящем пункте описывается комплект различных коэффициентов зависимости, наиболее часто реализуемых в программных средствах по статистике.

I.Статистическая зависимость.

Статистическая зависимость между двумя признаками является инвариантным относительно взаимно однозначных преобразований (перекодирований) их значений, а также и относительно их перестановок. Поэтому все коэффициенты статистической зависимости зависят только от таблицы частот (n_{ij}, $n_{i\cdot}$, $n_{\cdot j}$, $i = 1, \dots, k$; $j = 1, \dots, h$).

Наиболее известной характеристикой статистической зависимости является χ^2-статистика, см. [I], стр. 26-29, [2], стр. II6-I2I, [3], [5], стр. 232-235,

$$\chi^2 = \frac{1}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{h} (n_{ij} n - n_{i\cdot} n_{\cdot j})^2 / n_{i\cdot} n_{\cdot j}, \qquad (I3)$$

или ее логарифмический аналог, см. [I], стр. 26,

$$y^2 = 2 \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{n_{ij}}{n} \ln \left(\frac{n n_{ij}}{n_{i\cdot} n_{\cdot j}} \right). \qquad (I4)$$

Наиболее часто применяются следующие коэффициенты статистической зависимости, вычисляемые из χ^2-статистики* (I3):

I° Коэффициент Крамера (Cramer) V, см. [I], стр.35, [2], стр. I2I, [5], стр. 236,

$$V = \sqrt{\frac{\chi^2}{n(\min(k-1),(h-1))}}, \qquad (I5)$$

2° Коэффициент Чупрова T, см. [I], стр. 35, [3], [5], стр. 236-235,

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(k-1)(h-1)}}}, \qquad (I6)$$

3° Среднеквадратический коэффициент сопряженности φ, см. [I], стр. 25

* Весьма редко при их вычислении вместо χ^2 пользуются y^2.

$$\Psi = \sqrt{\frac{\chi^2}{n}}, \qquad (17)$$

4^0 Коэффициент Пирсона или коэффициент контингентности C, [1], стр. 25; [2], стр. 122, [4], стр. 54:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}. \qquad (18)$$

5^0 Модифицированная (направленная) χ^2-статистика используются в τ-коэффициентах, выведенных Гудманем и Краскалем (см. [1], стр. 36),

$$\tau_a = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}\left(n n_{ij} - n_{i\cdot} n_{\cdot j}\right)^2 / n_{i\cdot}}{n\left(n^2 - \sum_{i=1}^{k} n_{i\cdot}^2\right)}, \qquad (19)$$

$$\tau_b = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}\left(n n_{ij} - n_{i\cdot} n_{\cdot j}\right)^2 / n_{\cdot j}}{n\left(n^2 - \sum_{i=1}^{k} n_{i\cdot}^2\right)}, \qquad (19')$$

$$\tau_c = \frac{\tau_a\left(n^2 - \sum_{j=1}^{h} n_{\cdot j}^2\right) + \tau_b\left(n^2 - \sum_{i=1}^{k} n_{i\cdot}^2\right)}{n\left[2n^2 - \sum_{i=1}^{k} n_{i\cdot}^2 - \sum_{j=1}^{h} n_{\cdot j}^2\right]}. \qquad (19'')$$

Все эти коэффициенты построены исходя из факта, что в случае статистической независимости $\chi^2 = 0$, и следовательно, они также равняются нулю.

6^0 Второй класс статистик, измеряющих статистическую зависимость признаков, построен исходя из максимальных частот n_{i*} i-ой строки, n_{*j} j-го столбца и соответственных максимальных частот. Это так называемые λ-коэффициенты, выведенные Гутманом (см. [1], стр. 32-35, [2], стр. 126).

$$\lambda_a = \frac{\sum_{i=1}^{k} n_{*j} - n_{*\cdot}}{n - n_{*\cdot}}, \qquad (20)$$

$$\lambda_b = \frac{\sum_{i=1}^{k} n_{i*} - n_{.*}}{n - n_{.*}}, \qquad (20')$$

$$\lambda = \frac{(n - n_{.*})\lambda_b + (n - n_{*.})\lambda_a}{2n - n_{*.} - n_{.*}}. \qquad (20'')$$

II. Монотонная зависимость.

Для измерения монотонной зависимости имеется два подхода, исходя из определений полной зависимости и полной независимости, см. [5], стр. 237-240.

I° Полная монотонная зависимость имеет место при совпадении вариационных рядов $x'_1, \ldots, x'_n$ и $y'_1, \ldots, y'_n$ рассматриваемых признаков. Предположим, что все наблюдения различные, т.е.

$$k = h = n, \qquad (21)$$

и обозначим ранги признаков X и Y соответственно через q_j и t_j, т.е.

$$x_j = x'_{q_j}, \quad y_j = y'_{t_j}, \qquad (22)$$

и определим статистику D, см. [5], стр. 238-239.

$$D = \sum_{j=1}^{n} (q_j - t_j)^2. \qquad (23)$$

Через статистику D определяется известный коэффициент ранговой корреляции Спирмена, см. [2], [4], стр. 63-67, [5], стр. 240-245.

где
$$\rho = 1 - D/2S, \qquad (24)$$

$$S = n(n^2 - 1)/3. \qquad (25)$$

2° Если условия (22) не выполнены, следует вместо рангов пользоваться усредненными рангами, вычисляемыми при помощи маргинальных распределений:

$$\begin{cases} Q_i = n_{i-1,\cdot} + \dfrac{1 + n_{i\cdot}}{2}, & i = 1, \ldots, k, \\ T_j = n_{\cdot j-1} + \dfrac{1 + n_{\cdot j}}{2}, & j = 1, \ldots, h, \end{cases} \tag{22'}$$

$(n_{0\cdot} = n_{\cdot 0} = 0)$, тогда

$$D^* = \sum_{i=1}^{k} \sum_{j=1}^{h} n_{ij} (Q_i - T_j)^2, \tag{23'}$$

$$\rho^* = 1 - D^*/2S^*, \tag{24'}$$

$$S^* = S - \sum_{i=1}^{k} \frac{n_{i\cdot}(n_{i\cdot}^2 - 1)}{6} - \sum_{j=1}^{h} \frac{n_{\cdot j}(n_{\cdot j}^2 - 1)}{6}. \tag{25'}$$

Легко видно, что в частном случае (22) формулы (23)-(25) непосредственно вытекают из формул (23') - (25').

3^0 Исходя из определения полной монотонной независимости в качестве статистик используются числа пар индексов, которым соответствуют односторонние и противоположные изменения признаков, см. [1], стр. 37, [4], стр. 69-70.

$$\begin{cases} S_0 = \varkappa\{(x_g < x_f) \cap (y_g < y_f),\ g, f = 1, \ldots, n\}, \\ D_0 = \varkappa\{(x_g < x_f) \cap (y_g > y_f),\ g, f = 1, \ldots, n\}. \end{cases} \tag{26}$$

Если не выполнено (22), то следует подсчитать еще пары, соответствующие равным значениям признаков:

$$\begin{aligned} T_a &= \varkappa\{(x_g = x_f) \cap (y_g \neq y_f);\ g, f = 1, \ldots, n,\ g < f\}, \\ T_b &= \varkappa\{(x_g \neq x_f) \cap (y_g = y_f);\ g, f = 1, \ldots, n,\ g < f\}, \\ T_c &= \varkappa\{(x_g = x_f) \cap (y_g = y_f),\ g, f = 1, \ldots, n,\ g < f\}, \end{aligned} \tag{26'}$$

в общем случае имеет место равенство

$$S + D + T_a + T_b + T_c = n^2.$$

При помощи статистик (26), (26') определены следующие коэффициенты монотонной зависимости:

γ (Гудман-Краскал); см. [1], стр. 37, [2], 123,

$$\gamma = \frac{S_0 - D_0}{S_0 + D_0}, \tag{27}$$

τ_k (Кендала), см. [1], стр. 38,

$$\tau_k = \frac{2(S_0 - D_0)}{\sqrt{(S_0 + D_0 + T_a)(S_0 + D_0 + T_b)}}, \tag{28}$$

и коэффициенты Сомерса, см. [1], стр. 39-40, [2], 123,

$$\begin{cases} d_{ba} = \dfrac{S_0 - D_0}{S_0 + D_0 + T_b}, \\ d_{ab} = \dfrac{S_0 - D_0}{S_0 + D_0 + T_a}, \\ d = \dfrac{S_0 - D_0}{S_0 + D_0 + 0.5(T_a + T_b)}. \end{cases} \tag{29}$$

4° Для случая $k=2$, $h=2$ хорошо известно еще коэффициент Юла (см. [1], стр. 22):

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}, \tag{30}$$

который по существу измеряет монотонную зависимость.

Заметим, что все статистики, измеряющие монотонную зависимость содержат кроме частот n_{ij} еще и ранги (или только ранги), следовательно, они, в отличие от статистик, измеряющих статистическую зависимость, существенно изменяются в результате перестановки значений признаков, но являются инвариантными относительно их монотонных преобразований.

III. Регрессионная зависимость.

Прогнозом (в смысле регрессионной зависимости) является условное математическое ожидание и степень прогнозирования измеряется по отношению к компонентам дисперсии, см. [5], стр. 94-104; следовательно, статистиками для измерения регрессионной зависимости являются моменты и условные моменты:

$$\begin{cases} M_{1\cdot} = \frac{1}{n}\sum_{i=1}^{k} a_i n_{i\cdot}; & M_{\cdot 1} = \frac{1}{n}\sum_{j=1}^{h} b_j n_{\cdot j}; \\ \overline{M}_{2\cdot} = \frac{1}{n}\sum_{i=1}^{k} n_{i\cdot}(a_i - M_{1\cdot})^2; & \overline{M}_{\cdot 2} = \frac{1}{n}\sum_{j=1}^{h} n_{\cdot j}(b_j - M_{\cdot 1})^2; \\ \overline{M}_{11} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij}(a_i - M_{1\cdot})(b_j - M_{\cdot 1}) \end{cases} \tag{31}$$

$$\begin{cases} M_{i1} = \frac{1}{n_{i\cdot}}\sum_{j=1}^{h} b_j n_{ij}, & M_{1j} = \frac{1}{n_{\cdot j}}\sum_{i=1}^{k} a_i n_{ij}, \\ \overline{M}_{i2} = \frac{1}{n_{i\cdot}}\sum_{j=1}^{h} n_{ij}(b_j - M_{i1})^2; & \overline{M}_{2j} = \frac{1}{n_{\cdot j}}\sum_{i=1}^{k} n_{ij}(a_i - M_{1j})^2; \\ i = 1, \ldots, k; & j = 1, \ldots, h. \end{cases} \tag{32}$$

На основании приведенных статистик определен коэффициент регрессионной зависимости $\eta(Y/X)$, известный также под названием "корреляционное" (или регрессионное) отношение, см. [2], [5], стр. 100, [4], 180-186, [3], стр. 128,

$$\eta(Y/X) = \sqrt{1 - \frac{\sum_{i=1}^{k} n_{i\cdot}\overline{M}_{i2}}{n\,\overline{M}_{\cdot 2}}}, \tag{33}$$

$$\eta(X/Y) = \sqrt{1 - \frac{\sum_{j=1}^{h} n_{\cdot j}\overline{M}_{2j}}{n\,\overline{M}_{2\cdot}}}. \tag{33'}$$

Так как в состав формулы для вычисления $\eta(Y/X)$ входят значения признака Y, то этот коэффициент существенно зависит от всевозможных преобразований признака Y (в том числе и перестановки их значений). В то же время, в формулу (33) не входят ни значения признака X, ни их индексы, и поэтому $\eta(Y/X)$ является инвариантным относительно любых взаимно однозначных преобразований признака X.

Для описания "взаимной" регрессионной зависимости иногда используется коэффициент

$$\eta(X,Y) = \max\left(\eta(X/Y), \eta(Y/X)\right). \tag{33''}$$

IУ. Корреляционная зависимость.

Корреляционную зависимость измеряет хорошо известный коэффициент (линейной) корреляции , вычисляемый по статистикам (3I), см. [5], стр. II0, [2], [4], 94-99.

$$r = \frac{\bar{M}_{11}}{\sqrt{\bar{M}_{2\cdot}\ \bar{M}_{\cdot 2}}} \,. \qquad (34)$$

Так как в состав статистик входят значения обоих признаков, то r существенно зависит от всевозможных преобразований признаков. Единственным различием является линейное преобразование, относительно которого коэффициент корреляции является инвариантным:

если

$$Z = \alpha X + \beta, \qquad W = \gamma Y + \delta,$$

то

$$r(Z,W) = \operatorname{sgn}(\alpha\gamma)\, r(X,Y).$$

5. Свойства коэффициентов зависимости.

Коэффициентами зависимости считаются функционалы распределения, рангов или значений рассматриваемых признаков: $r(n_{ij})$, $r(n_{ij}, q_i, t_j)$ или $r(n_{ij}, a_i, t_j)$, для которых разработаны некоторые (традиционные) свойства, облегчающие их интерпретацию. Перечислим некоторые из этих свойств:

I° Симметричность (относительно направления, только для взаимных коэффициентов зависимости):

$$r(X,Y) = r(Y,X).$$

2° Центрированность:

$r(X,Y) = 0$, если X и Y H-независимы;

$r(Y/X) = 0$, если Y H-независим от X .

3° Симметричность (относительно знака; применима для монотонной и корреляционной зависимости):

$$r(X,Y) = -r(-X,Y) = -r(X,-Y);$$
$$r(Y/X) = -r(-Y/X) = -r(Y/-X),$$

где $(-\mathcal{Z})$ – признак, полученный из $\mathcal{Z}$:

1) умножением на – 1 (если $\mathcal{Y}$ –числовый признак),

2) заменой порядка на противоположное $z_g := z_{n+1-g}$, $g=1,\dots,n$ (если $\mathcal{X}$ –порядковый признак).

4° Нормированность.

Пусть $\mathcal{K}_r$ – множество всевозможных значений коэффициента r (при эмпирических распределениях с конечным объемом выборки n, $n \in \mathcal{N}$). Обозначаем еще через $\mathcal{K}_r(k,h)$ множество всевозможных значений коэффициента r при зафиксированных значениях k и h $(k,h \in \mathcal{N})$.

Вводим следующие обозначения:

$$\underline{r} = \inf_{r \in \mathcal{K}_r} r \quad , \quad \overline{r} = \sup_{r \in \mathcal{K}_r} r \quad , \tag{35}$$

$$\underline{r}(k,h) = \inf_{r \in \mathcal{K}_r(k,h)} r \quad , \quad \overline{r}(k,h) = \sup_{r \in \mathcal{K}_r(k,h)} r \quad . \tag{35'}$$

Коэффициент r называется

нормированным, если

$$\overline{r} = 1, \tag{36}$$

(k,h)-нормированным, если

$$\overline{r}(k,h) = 1. \tag{36'}$$

При монотонной и корреляционной зависимости к требованиям (36) и (36') прибавляется их аналог для $\underline{r}$ и $\underline{r}(k,h)$.

Рис. 3.

Таблица I

Название коэффициента	Символ	Тип зависимости	Направленность	Центрированность	Симметричность	Нормированность $\bar{\tau}$	Нормированность $\bar{\tau}(k,h)$
хи-квадрат	χ^2	C	$X \Rightarrow Y$	+		∞	$n\cdot(h-1)$
Крамера	V	C	$X \Leftrightarrow Y$	+		1	1
Чупрова	T	C	$X \Leftrightarrow Y$	+		$\sqrt{\frac{h-1}{k-1}}$	$\sqrt{\frac{h-1}{k-1}}$
коэфф. сопряж.	φ	C	$X \Leftrightarrow Y$	+		$\sqrt{h-1}$	$\sqrt{h-1}$
Пирсона	C	C	$X \Leftrightarrow Y$	+		$\sqrt{\frac{h-1}{h}}$	$\sqrt{\frac{h-1}{h}}$
Гудман-Краскала	τ_a	C	$X \Rightarrow Y$	+		1	1
	τ_b	C	$Y \Rightarrow X$	+		1	$\sqrt{(h-1)/(k-1)}$
	τ	C	$X \Leftrightarrow Y$	+		1	$(h-1)/(k-1)$
λ-коэффициент	λ_a	C	$X \Rightarrow Y$	+		1	1
	λ_b	C	$Y \Rightarrow X$	+		1	$(h-1)/(k-1)$
	λ	C	$X \Leftrightarrow Y$	+		1	$\frac{h-1}{k-1} \le \bar{\tau} < 1$
Спирмена	ϱ	M	$X \Leftrightarrow Y$	+	+	1	1
Кендалла	τ_K	M	$X \Leftrightarrow Y$	+	+	2	< 2, если $h \le n$
Гамма	γ	M	$X \Rightarrow Y$	+	+	1	1
Сомерса	d_a	M	$X \Rightarrow Y$	+	+	1	1, если $h \le n$
	d_b	M	$Y \Rightarrow X$	+	+	1	
	d	M	$X \Leftrightarrow Y$	+	+	1	≤ 1, если $k \le n$
	$\eta(Y/X)$	R	$X \Rightarrow Y$	+		1	1,
Регрессионное (корреляционное отношение)	$\eta(X/Y)$	R	$Y \Rightarrow X$	+		1	< 1
	$\eta(X,Y)$	R	$X \Leftrightarrow Y$	$\pm$	+	1	1
Коэффициент корреляции	r	K	$X \Leftrightarrow Y$	+	+		1, если $h < k$

Везде предполагается, что $2 \le h \le k \le n$.

6. Семейство $\mathcal{U}$ дискретных равномерных распределений для исследования и тестирования поведения коэффициентов.

Для исследования свойств неких статистических процедур, как правило, вводятся стандартные распределения, при которых эти процедуры имеют известные статистические свойства.

Чаще всего за такие распределения выбрано (одно- или многомерное) нормальное распределение.

Нормальное распределение не применимо с целью исследования поведения эмпирического распределения, так как каждое эмпирическое распределение является лишь приближением для нормального распределения.

Более подходящим стандартным распределением для исследования эмпирического распределения является дискретное распределение, определяемое соотношениями

$$\begin{cases} P(a_i) = 1/k, & i = 1, \dots, k; \\ P(b_j) = 1/h, & j = 1, \dots, h. \end{cases}$$

В пользу применения равномерного распределения в качестве стандартного распределения говорит и тот факт, что оно инвариантно относительно перестановки значений признаков, что хорошо согласуется со свойствами т.п. номинальных признаков, $\chi (\chi \in \mathfrak{B} \setminus \mathfrak{R})$, а также и то, что равномерное распределение максимизирует энтропию H (при заданном k), и, следовательно, соответствует целесообразно выбранной шкале.

В двумерном случае необходимо определить и распределение, имеющее заданную степень зависимости. Для этого определим вполне независимое двумерное распределение при помощи соотношения

$$P(a_i, b_j) = 1/kh \quad (i = 1, \dots, k, j = 1, \dots, h),$$

притом соответствующее эмпирическое распределение существует при $n = ckh$, $c \in \mathbb{N}$.

Вполне зависимое двумерное распределение (для всех типов зависимости) можно определить как обобщенно-диагональное распределение, следующим образом:

пусть $k = eh$, $e \in \mathbb{N}$;

$$p_{ij} = \begin{cases} 1/k, & \text{если } i = e(j-1)+1, \dots, ej;\ j = 1, \dots, h, \\ 0 & \text{иначе,} \end{cases}$$

(см. рис. 3).

При таком распределении признак Y является вполне H-зависимым от признака X в смысле статистической и регрессионной зависимости; для того, чтобы зависимость была полной монотонной, необходимо, чтобы $e=1$ (т.е. $k = h$), т.е., чтобы распределение было диагональное. Для того, чтобы имела место полная корреляционная зависимость, необходимо, чтобы значения a_i и b_j $(i=1,\dots,k)$ были пропорциональными (например, $a_i = b_i = i$). Заметим, что при $e > 1$ и $a_i = i$, $b_j = j$ вышеописанное диагональное распределение максимизирует коэффициент корреляции r при заданных маргинальных распределениях.

Пусть фиксированы числа h и $k = eh$, и имеется два двумерных эмпирических распределения P_{xy} соответсвеннно с частотами n_{ij} $(i = 1, \dots, k;\ j = 1, \dots, h)$ и Q_{xy} с частотами m_{ij} $(i = 1, \dots, k;\ j = 1, \dots, h)$; пусть их объемы соответственно n и m. Тогда эмпирическое распределение W_{xy}, определенное через частоты ν_{ij} $(i = 1, \dots, k,\ j = 1, \dots, h)$,

$$\nu_{ij} = \alpha n_{ij} + \beta m_{ij} \quad (\alpha, \beta \in \mathbb{N}) \tag{37}$$

является смесью распределений P_{xy}, Q_{xy}

$$W_{xy} = \gamma P_{xy} + (1-\gamma) Q_{xy} \tag{38}$$

Все параметры распределения W_{xy} легко вычисляемы по соответствующим параметрам исходных распределений P_{xy} и Q_{xy}, см. [6].

Выбирая (при фиксированных h и $k=\ell-h$) распределения $P_{xy}(n_{ij})$ и $Q_{xy}(m_{ij})$ следующим образом:

$$n_{ij}=1 \quad (i=1,\dots,k;\ j=1,\dots,h)$$

и

$$m_{ij}=\begin{cases}1, & \text{если } i=\ell(j-1)+1,\dots,\ell j,\\ 0, & \text{иначе } j=1,\dots,h,\end{cases}$$

то по заданному рациональному числу γ ($\gamma=p/q$, $|p|,|q|\in N$, $q\neq 0$) легко вычислить параметры a и b так, чтобы получить смесь (38).

Пусть ℓ – наибольший общий делитель чисел $|p|$ и h, $p=p_1\ell$, $h=h_1\ell$. Тогда определяя

$$a=p_1,\quad b=(q-p)h_1, \tag{39}$$

мы получим следующую двумерную таблицу

$$\left.\begin{array}{cccccc}
a & a & \dots & \dots & a+b & \Big\}\ell\\
\dots & \dots & \dots & \dots & \dots & \\
a & a & \dots & \dots & a+b & \\
\dots & \dots & \dots & \dots & \dots & \\
a+b & a & \dots & \dots & a & \\
\dots & \dots & \dots & \dots & \dots & \\
a+b & a & \dots & \dots & a &
\end{array}\right\}\ell\cdot h, \tag{40}$$

$\underbrace{\qquad\qquad\qquad\qquad}_{h}$

При $p<0$ изменяется направление главного диагонала. Объем распределения (40) равняется $k(ha+b)$.

Таким образом, мы получим 3-параметрическое семейство распределений $\{\ell, n\}$, где $\gamma=\dfrac{b}{ha+b}$.

Чтобы исследовать влияния n, можно в качестве дополни-

тельного параметра включить сомножитель c ($n := c(tn+um)$, $c \in N$), а для изучения влияния преобразования значений признаков к регрессионной и корреляционной зависимости, ввести преобразования: $a_i := \varphi(a_i)$ $(i=1,\dots,k)$, $b_j := \psi(b_j)$ $(j=1,\dots,h)$. Для простоты мы в дальнейшем считаем $\ell := 1$, $a_i := i$ $(i=1,\dots,k)$, $b_j := j$ $(j=1,\dots,h)$.

7. Формулы для вычисления значений коэффициентов зависимости для таблиц из семейства $\mathcal{O}_f(\gamma; \ell, h)$

В дальнейшем выпишем точные выражения коэффициентов заданных в таблице I по параметрам h, ℓ, γ. Для простоты выражений в некоторых формулах пользуют и значения a, b и k.

I. Коэффициенты статистической зависимости.

Следующим шагом является вычисление статистик (I9)-(34) для введенного распределения.

Непосредственно из (I3) вытекает, что

$$\chi^2 = \frac{\ell^2 k(h-1)}{ha+b} = \gamma^2 (h-1) n,$$

откуда получаем

$$V = \frac{\ell}{ha+b} = \gamma,$$

$$T = \frac{\ell}{ha+b}\sqrt{\frac{h-1}{k-1}} = \gamma\sqrt{\frac{h-1}{k-1}},$$

$$\varphi = \frac{\ell}{ha+b}\sqrt{h-1} = \gamma\sqrt{h-1},$$

$$C = \frac{\ell\sqrt{h-1}}{\sqrt{\ell^2(h-1)+(ha+b)^2}}.$$

Аналогично выводятся и τ-коэффициенты:

$$\tau_a = \frac{b^2}{(ha+b)^2} \cdot \frac{h-1}{k-1} = \gamma^2 \frac{h-1}{k-1},$$

$$\tau_b = \frac{b^2}{(ha+b)^2} = \gamma^2,$$

$$\tau = \frac{b^2}{(ha+b)^2} \frac{(h-1)(l-1)}{2k-(l+1)}.$$

Учитывая факт, что модальная частота в каждой строке и в каждом столбце равняется $l(a+b)$, легко вычислить и λ-коэффициенты:

$$\lambda_a = \frac{b}{ha+b} \frac{h-1}{k-1},$$

$$\lambda_b = \frac{b}{ha+b} = \gamma,$$

$$\lambda = \frac{b}{ha+b} \frac{(h-1)(1+l)}{2k\,(l+1)}.$$

Из приведенных формул следует содержательная близость почти всех приведенных коэффициентов: все они измеряют в некотором смысле удельный вес $\gamma = \frac{b}{ha+b}$ диагонального компонента смеси; все направленные коэффициенты учитывают и различные числа значений признаков X и Y : при прогнозе Y (имеющего меньше значений) коэффициент имеет максимальное значение, при прогнозе X (имеющего больше значений) коэффициент умножается на сомножитель $\frac{h-1}{k-1}$ (или $\sqrt{\frac{h-1}{k-1}}$), который меньше 1.

Коэффициенты V и T не направленные, один из них (V) измеряет большую из направленных зависимостей; другой (T) - меньшую; ненаправленные τ и λ получены из направленных путем (взвешенного) усреднения.

От других коэффициентов (которые получаются друг из дру-

га в результате весьма простых преобразований), отличаются ненормированный в общем φ, совпадающий с V и T в случае 2x2 таблиц, и ненормированный C, который приближается к I лишь при $a=0$, $h\to\infty$.

Для практической работы кроме χ^2 (на основе которой проверяется значимость зависимости) можно пользоваться не более чем двумя-тремя коэффициентами статистической зависимости. Такими комплектами могут быть, например: V и T; или τ_a, τ_b, τ или λ_a, λ_b, λ. Параллельное применение коэффициентов V, T и λ только вызывает недоразумений, но не прибавляет информации.

II Коэффициенты монотонной зависимости.

I° Для вычисления коэффициентов (28)-(30) необходимо найти для данного распределения значения статистик (26) и (26'). Непосредственное вычисление дает

$$T_a = ak(h-1)(ha+2b),$$
$$T_b = k(a^2h(k-1)+2ab(k-1)+b^2(b-1)),$$
$$T_c = k((a+b)^2+(h-1)a^2).$$

Для вычисления D воспользуются вспомогательными статистиками

$$K = b[b(a+b)^2+a^2(h-2)(k-2b+1)/2+a(a+b)(2k-3b+1)],$$
$$L = 2/3\ abck(h-1)(h-2),$$

отсюда

$$D = h(h-1)K - L - T_a$$

и из соотношения (27) получим:

$$S = k^2(ha+b)^2 - D - T_c - T_a - T_b.$$

Отсюда легко получить точные значения для коэффициентов (28)-(30).

Заметим, что коэффициент γ (27) не учитывает повторных наблюдений в вариационных рядах, и поэтому в общем случае

$$\gamma^2 \geq \tau_k \; ; \; \gamma^2 \geq d_a \, , \; d_\ell \, , \; d \, ,$$

притом $\gamma^2 = 1 \iff D = 0$.

Коэффициенты τ и d выражают полную монотонную зависимость, если вариационные ряды не содержали повторяющихся наблюдений, т. е. $k = n$ и $k = h$, соответственно.

2^0 Для вычисления коэффициента ранговой корреляции Спирмена ϱ необходимо вычислить выражения статистик D (23) и S (25), для которых имеются следующие формулы.

$$D = \frac{q^2}{12}\left(a(2k^2 - \ell^2 - 1)h + b(\ell^2 - 1)\right),$$

$$S = \frac{1}{3}\left(n(n^2-1) - k\,\frac{q(q^2-1)}{2} - h\,\frac{\ell q(\ell^2 q^2 - 1)}{2}\right),$$

где $q = ha + b$, $n = kq$, $k = \ell h$.

Как видно из формул, ϱ достигает максимального значения при $\ell = 1$, т.е. при $k = h$, притом не требуется, чтобы $k = h = n$; следовательно, в отличие от τ , ϱ может быть равным I и в случае, когда вариационные ряды имеют повторяющиеся значения.

Следовательно, для полного описания монотонной зависимости можно пользоваться, например, коэффициентами ϱ, d_a, d_ℓ и τ (различие $\varrho - \tau$ характеризуют степень повторений, различие $d_a - d_\ell$ направленность монотонной зависимости).

3^0 Коэффициент Юла, применяемый лишь при $k = h = 2$, имеет значение

$$Q = \frac{(a+b)^2 - a^2}{(a+b)^2 + a^2} \, .$$

Применение этого коэффициента оправдывает преимущест-

венно, традициями.

III Коэффициенты регрессионной и корреляционной зависимостей.

Пользуясь выражениями для математических ожиданий и дисперсий:

$$EX = \frac{k+1}{2}, \quad EY = \frac{h+1}{2}, \quad DX = \frac{k^2-1}{12}, \quad DY = \frac{h^2-1}{12},$$

а также и для специального момента:

$$EXY = \frac{l(h^2-1)}{3} + \frac{(k-1)(k+1)}{4},$$

мы получим для обобщенно-диагонального распределения

$$r = \frac{l(h^2-1)}{\sqrt{(k^2-1)(l^2h^2-1)}} = \sqrt{\frac{l^2h^2-l^2}{l^2h^2-1}},$$

а, следовательно, для смеси (38), пользуясь результатом из (6),

$$r = \gamma \sqrt{\frac{l^2h^2-l^2}{l^2h^2-1}}.$$

Учитывая тот факт, что в данном случае регрессионная зависимость максимально аппроксимируется линейной, мы имеем:

$$\eta(Y|X) = \gamma,$$
$$\eta(X|Y) = r,$$
$$\max \eta = \gamma.$$

В качестве примера в таблице 2 все приведенные коэффициенты вычислены для двух распределений:

Q_1: $\gamma = 0.5$, $k = 3$, $l = 1$

$$\begin{pmatrix} 1 & 1 & 4 \\ 1 & 4 & 1 \\ 4 & 1 & 1 \end{pmatrix}$$

Q_2: $\gamma = 0.5$, $k = 3$, $l = 2$

$$\begin{pmatrix} 1 & 1 & 4 \\ 1 & 1 & 4 \\ 1 & 4 & 1 \\ 1 & 4 & 1 \\ 4 & 1 & 1 \\ 4 & 1 & 1 \end{pmatrix}$$

Следовательно, для распределения Q_1 имеется $n = 12$, $k = 3$, $m = 3$ выборка следующая:

(1,1),(1,1),(1,1),(1,1),(1,2),(1,3),(2,1),(2,2),(2,2),(2,2),(2,2),(2,3),(3,1),(3,2),(3,3),(3,3),(3,3),(3,3)

Таблица 2.

Коэффициент	Q_1	Q_2
χ^2	9	I8
V	0,5	0,5
T	0,5	0,3I622777
φ	0,707I0678	0,707I0678
c	0,57735027	0,57735027
τ_a	0,25	0,I
τ_b	0,25	0,25
τ	0,25	0,I6666667
λ_a	0,5	0,2
λ_b	0,5	0,5
λ	0,5	0,333333333
ϱ	0,5	0,4776II94
τ_k	0,4722222	0,42236840
γ	0,62963	0,53968254
d_a	0,47222222	0,37777778
d_b	0,47222222	0,0,47222222
d	0,47222222	0,4I975309
$\eta(X\|X)$	0,5	0,5
$\eta(X\|Y)$	0,5	0,47809I44
$\max\eta$	0,5	0,5
r	0,5	0,47809I44

Здесь для вычисления τ_k применилась формула

$$\tau_k = \tau_k$$

где τ_k вычислено по (28), см. также [2], стр. I23.

Литература

1. Аптон Г. Анализ таблиц сопряженностей. М., 1972.
2. Парринг А.-М., Тийт Э.-М. Методическое руководство для пользователя пакета САИСИ. Тарту, 1986.
3. Kollo T., Tooding L.-M. Kahemõõtmeline statistiline analüüs. Programmeerimiskõigile IX, Tartu, 1975.
4. Tiit E.-M. Andmeanalüüs II. Tartu, 1982.
5. Tiit E.-M., Parring A.-M. Tõenäosusteooria ja matemaatiline statistika. Tln., 1977.
6. Tiit E. Definition of mixtures with given moments. TRÜ Toimetised, 1986, N° 733, 3-13.

Поступило 19.10.1988

Resümee

Kahe tunnuse vahelise seose kordajad. Kahemõõtmeline standardjaotus seosekordajate uurimiseks ja testimiseks

E.Tiit, M.Unt

Käesoleva artikli eesmärgiks on töötada välja metoodika mitmesuguste empiiriliste seosekordajate vahekordade selgitamiseks ja interpreteerimiseks ning nende arvutusliku õigsuse testimiseks.

Sel eesmärgil vaadeldakse diskreetset (empiirilist) jaotust iseloomustavaid seosekordajaid ja esitatakse seosekordajate teatav süstemaatika, mille aluseks prognoosi suund (ühepoolne või vastastikune), prognoosi tüüp, mis oluliselt sõltub vaadeldavate tuunuste tüübist (statistiline, monotoonne, regressiooniline või korrelatiivne), samuti ka seose sümmeetrilisus.

Nimetatud tunnuste alusel käsitletakse 21 seosekordajat, mis moodustavad põhiosa levinumates pakettides ja programmides arvutatavatest ning praktilistes rakendustes kasutatavatest seosekordajatest. Kõigi jaoks esitatakse ka maksimumväärtused (sõltuvalt tabeli dimensioonidest), vt. tab. 1.

Seosekordajate testimiseks konstrueeritakse lihtne kahemõõtmeline diskreetsete jaotuste pere, mille marginaaljaotused on ühtlased ja seos genereeritud sõltumatu ja diagonaalse komponendi seguna. Pere parameetriteks on tabeli mõõtmed k ja h ning segu diagonaalkomponendi kaal, mis eeldatakse olevat ratsionaalne. Kokkuleppeliselt $k = \ell \cdot h$, kus $\ell \in N$.

Selle kolmeparameetrilise pere jaoks tuletatakse kõigi varem esitatud seosekordajate täpsed avaldised parameetrite k ja h funktsioonidena. Tabelis on antud seosekordajate väärtused kahe konkreetse parameetrite komplekti puhuks.

S u m m a r y

Coefficients of dependence between two variables. A standard distribution for testing and studying the coefficients of dependence.

E.Tiit, M.Unt

The aim of the paper is to elaborate the method for the investigation of the relations between several empirical coefficients, measuring different kinds of statistical dependence and testing the correctness of their computation.

21 different coefficients, used in most popular packages and practical researches, have been systematized by several items: the direction (mutual or directed), the type of dependence (statistical, monotonic, regressional, correlative), and the symmetry. For all coefficients the maximal value (depending on the size of the table) is given.

A family of two-dimensional discrete distributions with uniform marginals and dependencies, generated with the help of mixture of diagonal and independent components is constructed. The family depends on three parameters - the table sizes k and h, and the weight of the diagonal component of the mixture (assumed to be rational). For all the coefficient their exact expression by the family parameters are given.

For two concrete sets of parameters the values of the coefficients are given in Table 2.

Труды ВЦ ТГУ (1986), 56, стр.88-101.

Изучение оценок среднего значения на базе реального потока задач

Л.М. Тоодинг

Ключевые слова: робастная оценка, форма кривой эмпирического распределения, оценка среднего значения

I. В связи с оживлением в последние десятилетия исследовательских работ по изучению свойств робастности статистических оценок в некоторой мере расширились и возможности практического применения этих оценок в обработке данных. В большинство широко известных стандартных статистических пакетов включены элементы "неклассического" подхода: ранговые критерии и оценки, вычислительные процедуры для уменьшения значения выбросов и т.п.Тем самым появилась возможность эмпирического анализа статистических оценок, базирующего на реальном статистическом материале и отличающего от часто применяемого при исследовании проблем робастности подхода с моделированием, построенным в соответствии с определенными теоретическими моделями.

В основу настоящей статьи положено исследование реального потока задач статистической обработки данных, накоп-

ленных в архиве статистических данных в Лаборатории прикладной математики ТГУ. Были рассмотрены признаки (статистические выборки переменных) из 26 случайно выбранных наборов данных. Целью эмпирического исследования была

I) характеристика формы кривой эмпирического распределения признака,

2) сопоставительное сравнение некоторых наиболее часто применяемых оценок среднего значения, в том числе робастных. Эмпирическая характеристика исходных данных прикладной статистики способствует выбору и обоснованию теоретических предположений при моделирующем подходе к изучению свойств робастности оценок, а эмпирическое сопоставление самих оценок - обоснованному выбору из них, например, при ограниченных вычислительных ресурсах.

Выбранные для настоящего исследования из архива наборы данных принадлежат к различным областям эмпирического исследования: к медицине, социологии, экономике, психологии и др. В исследование включены все записанные в архиве признаки каждого набора, причем их количество в наборах варьируется в пределах от 8 до 370. Объем выборки n исследуемых признаков-выборок также изменяется в большом диапазоне: $50<n<3100$.

Изучаемые данные были получены при помощи пакетов статистической обработки данных STELLA, SAS и BMDP и собраны в два информационных массива. По первому массиву (данные о 2500 признаках-выборках) изучалась степень согласия эмпирического распределения с заданным теоретическим, причем в зависимости от объема выборки и количества градаций (различных значений) признака. По второму массиву (данные о 700

признаках) также изучалась форма кривой эмпирического распределения, но и свойства различных оценок среднего значения. О каждом признаке в массиве были накоплены следующие данные: размах значений признака, коэффициенты асимметрии и эксцесса, арифметическое среднее, медиана, мода, три робастные оценки среднего значения, стандартное отклонение, полуразность квартилей и некоторые другие показатели вида эмпирического распределения.

Оба массива данных были обработаны на ЭВМ ЕС при помощи пакетов статистической обработки данных STELLA и SAS .

2. При изучении вида кривой распределения признака была проверена гипотеза

$H_0: G = F,$

где G - теоретическое распределение признака, а F - предполагаемое распределение, в качестве которого были использованы равномерное и нормальное распределения (соответственно условия наибольшей информативности в смысле энтропии и центральное предположение классической статистики). Были использованы хи-квадрат критерий и критерий Колмогорова - Смирнова (при тестировании нормального распределения).

Как и предполагалось, распределение признаков не ограничивается этими двумя классами распределения. А именно , выяснилось, что нулевую гипотезу на уровне значимости 0.05 относительно равномерного распределения можно было отвергать в 99% случаев, а относительно нормального распределения в 92% случаев. Отметим, что результаты применения хи- квадрат критерия и критерия Колмогорова-Смирнова в общем совпадали,

причем в случае $n < 500$ нулевую гипотезу по хи-квадрат критерию пришлось отвергать чаще, чем по критерию Колмогорова-Смирнова.

Для характеристики симметричности распределения был использован коэффициент асимметрии относительно среднего значения:

$$a = m_3 / m_2^{3/2},$$

где m_k - выборочный к-ый центральный момент, для характеристики крутизны - коэффициент эксцесса:

$$e = m_4 / m_2^2 - 3 .$$

Стандартные отклонения этих коэффициентов при предположении нормального распределения оценивались соответственно по формулам $s_a = (6/n)^{1/2}$ и $s_e = (24/n)^{1/2}$ (см.[I], с. 298).

Анализ показал, что в среднем наблюдается положительная асимметрия, причем отрицательные значения коэффициента а имеются у 30% из распределений. Судя по коэффициенту эксцесса, приблизительно 60% из распределений более плоские, чем кривая нормального распределения. На рис. I и 2 изображены эмпирические распределения коэффициентов а и е.

Вид распределения признака зависит от типа использованной шкалы измерения. Приводим некоторые средние характеристики эмпирического распределения в разрезе статистической типологии признаков, используемой в пакете STELLA (см. [2]), где различаются количественные показатели (тип R), целочисленные результаты подсчета (тип A), качественные упорядоченные признаки (тип B) и номинальные признаки (тип C).

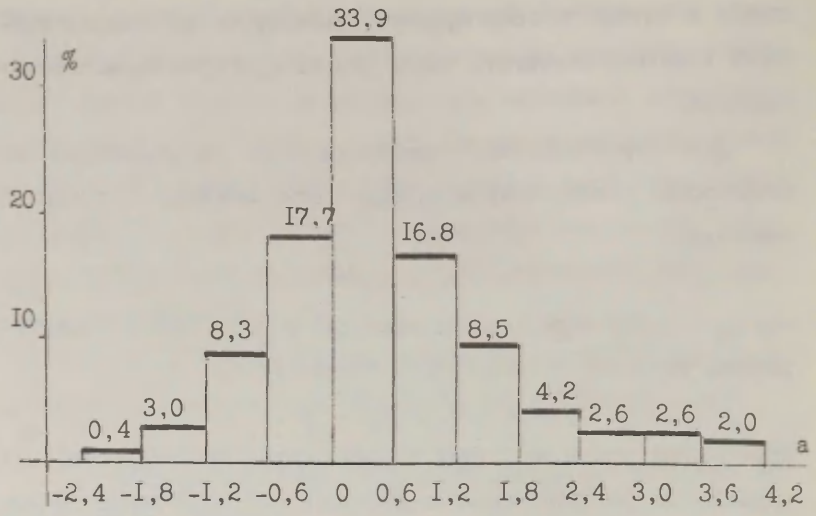

Рис. I. Распределение коэффициента асимметрии а.

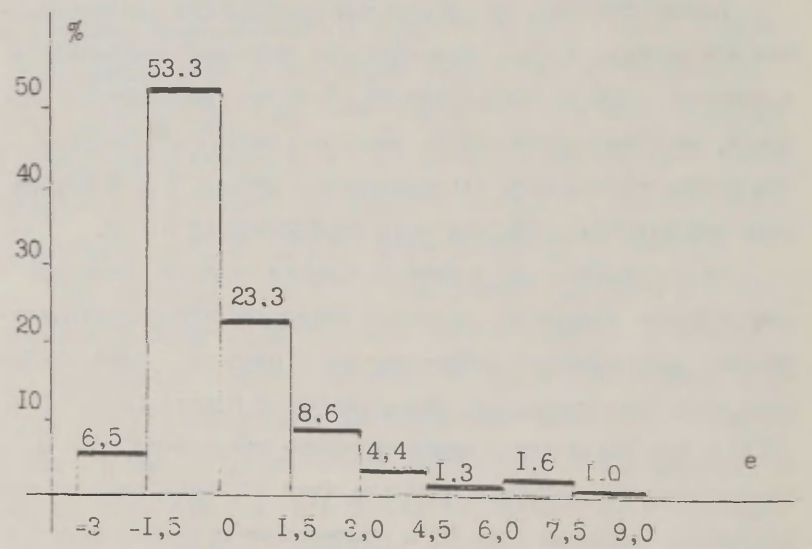

Рис. 2. Распределение коэффициента эксцесса е.

Номинальные признаки в данном случае не рассматриваются, так как у них изучаемые свойства не определены. Распределение признаков по типу в используемой выборке соответствует распределению, установленному при изучении реального потока задач статистического анализа данных (см. [2]). В качестве отдельной группы признаков рассматривались прогнозы по факторной модели.

В таблице I в разрезе типа признака приведены средние оценки различных характеристик распределения, причем указываются и 95%-ые доверительные границы среднего значения. Во всех группах наблюдается положительная асимметрия, причем распределения признаков типа A имеют наиболее вытянутый правый "хвост". Порядковые качественные шкалы (тип B) в среднем обеспечивают наиболее симметричное распределение ответов, имеющее более плоский вид, чем при остальных типах. Данная группа имеет в среднем близкие к нулю коэффициенты симметрии и эксцесса. Признаки типа R и прогнозы по факторной модели по степени симметричности и крутизны не различаются.

В таблице I приведены также данные о вариативности признаков. Тип A характеризуется наибольшим размахом значений - ширина шкалы в среднем 6.8 стандартных отклонений. Причем разбросанность по хвостам больше чем в других группах - расстояние между квартилями наименьшая и равняется 0.9 стандартных отклонений.

Приведенные выводы, в частности, свидетельствуют об обоснованности методики выделения при управлении обработкой данных различных статистических типов признаков, внедренной в пакете STELLA.

Таблица I

Характеристика распределения признаков различного статистического типа

Арифметическое среднее	Тип R	Факторы	Тип A	Тип B
a	0,99±0,3I	0,8I±0,40	2,44±0,73	0,I7±0,II
a/s_a	5,I ± I,9	3,7 ± I,9	I4,I±4,I	0,8 ± 0,8
e	4,78±2,97	4,I5±3,50	I7,73±8,33	0,II±0,52
e/s_e	I4,2±I0,9	9,5 ± 8,8	49,3±22,7	-0,0I±I,28
Размах (в единицах станд. откл.)	5,I6±0,3I	5,29±0,37	6,76±I,86	3,44±0,I0
Полуразность квартилей (в единицах станд. откл.)	0,59±0,04	0,62±0,04	0,44±0,07	0,68±0,03
Доля типа в данной выборке по [2]	0,I8	0,II 0,33	0,I2 0,08	0,52 0,50

3. Выводы предыдущего раздела о распределении реальных данных и различие этого распределения от нормального распределения в преобладающем большинстве случаев приводят к необходимости применения "неклассических" оценок среднего значения. Рассмотрим в данном разделе как при анализе реальных данных согласуются между собой следующие оценки среднего значения: арифметическое среднее $\bar{x}$, мода M_o, медиана M_e, и вычисляемые в ППП BMDP три робастные оценки, при которых проводится взвешивание значений признака. Подчеркиваем, что сравнение оценок в данном случае не может дать ответа, какая из них точнее, так как подход на базе реальных данных оставляет истинное среднее значение неизвестным. Выводы касаются различий значений оценок, а, тем самым, и различий смещений исследуемых оценок.

Пусть $x_1, x_2, \ldots, x_n$ - вариационный ряд, соответствующий выборке признака, распределение которого предполагается симметричным.

Робастные средние, рассматриваемые ниже, вычисляются по формуле

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} g(x_i)\, x_i \quad ,$$

где веса $g(x_i)$ при первой оценке - усеченным среднем $\bar{x}_T$ (the trimmed mean) определяются по правилу:

$$g(x_i) = \begin{cases} 1, & \text{если } 0.15\, n < i < 0.85\, n, \\ 0 & \text{в остальных случаях.} \end{cases}$$

Оценка Гампеля (the Hampel's mean) $\bar{x}_H$ и взвешенная оценка $\bar{x}_B$ (the biweight mean) вычисляются итеративным способом (см. [3]), причем вес значения x_i на к-том шагу для вы-

числения оценки $\bar{x}^{(k)}$ зависит от величины $(x - \bar{x}^{(k-1)})/m_{\Delta}$, где m_{Δ} - медиана абсолютных отклонений значений признака от медианы признака. При оценке Гампеля, в отличие от оценки $\bar{x}_B$, веса в окрестности среднего значения равны между собой. "Хвосты" распределения в обоих случаях взвешиваются меньшими коэффициентами, чем близкие к среднему значению элементы выборки.

Арифметические средние рассматриваемых оценок по всему набору исследуемых признаков значимо различаются (см. табл. 2). Как и ожидалось, робастные оценки дают меньшие по значению оценки, чем оценка $\bar{x}$. Если сравнивать различные оценки попарно (в таблице приведены 95%-ые доверительные границы среднего значения), то между робастными оценками разли-

Таблица 2.

Эмпирические средние оценок

	Арифметическое среднее оценки	Медиана оценки
$\bar{x}$	17,3 ± 6,8	2,1
M_e	14,3 ± 5,4	2,0
M_o	9,4 ± 4,4	2,0
$\bar{x}_H$	8,1 ± 3,3	2,0
$\bar{x}_T$	8,5 ± 3,6	2,0
$\bar{x}_B$	8,1 ± 3,3	2,0

чий нет, причем и мода дает в среднем ту же оценку. Отметим, что эмпирические медианы распределения рассматриваемых оценок практически не различаются.

В качестве показателей близости оценок были изучены их соотношения с арифметическим средним. Все оценки меньше $\bar{x}$, причем в среднем соотношения выравниваются: 0.856 ± 0.025 для медианы, $\bar{x}_H$ и $\bar{x}_B$, 0.899 ± 0.022 для $\bar{x}_T$ и 0.745 ± 0.035 для моды.

Для характеристики расстояния между оценками были вычислены значения коэффициента ранговой корреляции r_{ij} между оценками и вычислены коррелятивные расстояния $d_{ij} = \sqrt{1 - r_{ij}^2}$, $(i,j = 1,2,\ldots,6)$. Близость рассматриваемых оценок, а тем самым и их смещений, можно охарактеризовать следующими средними расстояниями (см. табл. 3): робастные оценки между собой - 0.14, медиана и робастные оценки - 0.17, арифмэтиче-

Таблица 3

Коррелятивные расстояния между оценками среднего значения

$\bar{x}$	0					
M_e	0,283	0				
M_o	0,574	0,500	0			
$\bar{x}_T$	0,266	0,141	0,497	0		
$\bar{x}_H$	0,172	0,209	0,512	0,178	0	
$\bar{x}_B$	0,270	0,148	0,495	0,045	0,184	0
	$\bar{x}$	M_e	M_o	$\bar{x}_T$	$\bar{x}_H$	$\bar{x}_B$

ское среднее и робастные оценки - 0.24, медиана и $\bar{x}$ - 0.28, мода и все остальные - 0.52, мода и робастные оценки -0.50. Таким образом, медиана в данном смысле ближе к робастным оценкам, чем мода и арифметическое среднее.

В предыдущем разделе было установлено разнообразие формы кривых распределения в зависимости от типа шкалы обрабатываемых признаков. Рассмотрим, как в разрезе типа признака варьируются значения различных оценок среднего значения. В таблице 4 приведены арифметические средние и медианы рассматриваемых оценок для групп различного типа.

Таблица 4

Эмпирические средние оценок в разрезе статистического типа признака

	Тип R		Факторы		Тип A		Тип B	
	Арифм. средн.	Медиана	Арифм. средн.	Медиана	Арифм. средн.	Медиана	Арифм. средн.	Медиана
$\bar{x}$	82±35	8.0	0 ± 0	0,00	4,3±I,4	2,8	2,3±0,I	2,2
M_e	66±28	7,0	-0,09±0,04	-0,I3	4,I±I,4	2,5	2,2±0,I	2,0
M_o	52±29	5,0	-0,39±0,23	-0,66	3,7±I,5	I,5	2,0±0,I	2,0
$\bar{x}_H$	36±I9	7,5	-0,06±0,03	-0,05	4,I±I,4	2,7	2,2±0,I	2,0
$\bar{x}_T$	38±2I	7,9	-0,06±0,02	-0,06	4,2±I,4	2,6	2,2±0,I	2,I
$\bar{x}_B$	36±I9	7,6	-0,06±0,03	-0,05	4,I±I,4	2,6	2,2±0,I	2,0

В случае типа R робастные оценки по значению меньше медианы и оценки $\bar{x}$, которые между собой в среднем не различаются. Причем оказалось, что соотношения оценок с $\bar{x}$ при робастных оценках и при медиане изменяются в пределах от 0.82 до 0.90, а при моде в пределах 0.59 ... 0.79.

Анализ эмпирических средних при созданных по факторной модели признаков показал, что робастные оценки находятся в пределах от - 0.09 до - 0.03, а мода дает оценку в пределах - 0.6 ... -0.2.

В случае типа A, как было указано выше, мы имеем дело с сильно асимметричными распределениями. Поэтому данные робастные процедуры, предполагающие по существу симметрию, не имеют "эффекта" и все оценки (кроме моды) в среднем совпадают. Несколько различаются медианы.

В группе признаков типа B средняя мода статистически значимо меньше среднего других оценок, а медиана и $\bar{x}$ близки к робастным оценкам, что объясняется установленными выше симметричностью и "правильной" формой кривой распределения в случае упорядоченной шкалы (хвосты незначительные). Медианы также практически совпадают.

В итоге можно заключить, что применение различных оценок при различных типах признаков имеет различный эффект, причем применение робастных оценок дает при типе R результаты более явно различающиеся от $\bar{x}$, чем при типе B.

Кроме типа признака расстояния различных оценок среднего значения были изучены и в разрезе уровня симметричности и уровня крутизны распределения признака. Оказалось, что имеется статистически значимая коррелятивная связь между

изученными нами соотношениями оценок с арифметическим средним и коэффициентами асимметрии(корреляция порядка -0.5 ... -0.6) и эксцесса (-0.3 ... -0.4). Различный уровень близости оценок на противоположных полюсах асимметрии и крутизны вызван, скорее всего, несимметричным распределением коэффициентов а и е в реальном потоке задач.

Литература

I. Закс Л. Статистическое оценивание. - М., Статистика, 1976, 598 с.

2. Тоодинг Л.М. Описание структуры реального потока задач анализа данных. - Труды ВЦ ТГУ, 48, 1981, с. 50-66.

3. Andrews, D.F., Bickel, P.J. et al. Robust Estimates of Location: Survey and Advances. Princeton, 1972.

Поступило 4.10.1988

Resümee

Keskväärtuse hinnangute uurimine
reaalse ülesannete voo põhjal

L.M. Tooding

Viimaste aastakümnete vältel on süvenenud mitteklassikaliste statistikameetodite teoreetiline käsitlus, mis on laiendanud ka nende meetodite praktilise kasutuse võimalusi. Artiklis vaadeldakse kuut praktikas enamlevinud keskväärtuse hinnangut ($\bar{x}$, mediaan, mood, 3 robastset hinnangut), toetudes seejuures reaalsele andmevoole, mis pärineb TRÜ Rakendusma-

temaatika labori statistilise andmetöötluse arhiivist. Eri hinnanguid kõrvutatakse empiiriliste keskväärtuste ja hinnangute korrelatiivse kauguse alusel, sealjuures ka tunnuse statistilise tüübi lõikes.

On esitatud tunnuse jaotusfunktsiooni empiiriline iseloomustus. Valdava enamuse uuritud tunnuste korral (92%) on võimalik tõestada empiirilise jaotuse erinevust normaaljaotusest, jaotused on keskmiselt positiivse asümmeetria ja ekstsessiga.

S u m m a r y

The investigation of estimates of the mean on the basis of the real tasks' stream

L.M. Tooding

During the last decade the theoretical approach to non-classical statistical methods has become more popular and in this connexion their application in practice is more wide spread. In the paper six estimates of the mean, most often used in practice, are considered - $\bar{x}$, the median, the mode, three robust estimates. The investigation is based on the real data stream derived from the statistical data processing archives at the Laboratory of Applied Mathematics of Tartu State University. The different estimates are compared on the basis of correlative distances between the estimates, including the analysis by the statistical type of the variable.

An empirical characterization of the distribution function of variables is presented. In the majority of cases (92%) the zero-hypothesis about the normality of the underlined distribution may be rejected. The asymmetric coefficient and curtosis are positive in average.

Труды ВЦ ТГУ (1988), 56, стр.102-119.

Об оценке распределения статистики Стьюдента

И. Траат

Ключевые слова: исходное распределение, распределение статистики Стьюдента, равномерное расстояние, третий абсолютный момент.

I. Введение.

Пусть $x_1, \ldots, x_n$ независимые одинаково распределенные случайные величины с функцией распределения (ф.р.) $F(x)$, так что

$$E x_i = 0, \quad E x_i^2 = 1.$$

Рассматривается статистика Стьюдента

$$T_n = \frac{\sqrt{n}\,\bar{x}}{s}, \tag{I.1}$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Известно, что статистика T_n имеет распределение Стьюдента при $F(x) = \Phi(x)$, где $\Phi(x)$ ф.р. нормального закона. При $F(x) \neq \Phi(x)$ известны лишь некоторые оценки распределения T_n и асимптотические результаты. Например, Славова (1985) доказывает неравенство Берри-Эссеена при условии $\nu_3 = E|x_i|^3 < \infty$:

$$\sup_x |P(T_n < x) - \Phi(x)| \leq \frac{c(\nu_3)}{\sqrt{n}}, \quad \forall n \geq 2, \tag{I.2}$$

где $c(\nu_3)$ - константа, зависящая от ν_3. Hall (1987) находит, что распределение статистики T_n равномерно разложимо в ряд Эджворта с точностью $o(n^{-k/2})$ при несингулярной $F(x)$ с конечными абсолютными моментами $E|x_i|^{k+2} < \infty$. В частном случае $k=1$ результат имеет вид:

$$P(T_n < x) = \Phi(x) + \frac{\tau}{6\sqrt{n}}(2x^2+1)\Phi'(x) + o(n^{-1/2}), \quad (I.3)$$

где $\tau = E x_i^3$.

В этих результатах распределение статистики T_n сравнивается с нормальным распределением, которое в данном случае является предельным распределением этой статистики. Предельное распределение статистики T_n при функции $F(x)$, принадлежащей области притяжения устойчивого закона, выведено в работе Золотарева (1985).

Целью настоящей статьи является сравнение распределения статистики T_n не с ее предельным распределением, а с распределением Стьюдента со степенями свободы $n-1$.

2. Постановка задачи.

Рассматривается равномерное расстояние

$$\rho(T,G) = \sup_x |T(x) - G(x)|, \quad (2.I)$$

где

$G(x)$ - ф.р. статистики T_n (I.I),

$T(x)$ - ф.р. Стьюдента со степенями свободы $n-1$.

Мы пытались ответить на следующие вопросы:

I) Каким является $\rho(T,G)$ по величине?

2) Как зависит $\rho(T,G)$ от исходного распределения $F(x)$?

3) Какие из характеристик исходного распределения лучше характеризуют $\rho(T,G)$: третий абсолютный момент ν_3 ис-

ходного распределения, существующий обычно в неравенствах типа Берри-Эссеена. или равномерное расстояние исходного распределения от нормального — $\rho(F,\Phi)$?

4) Существует ли для $\rho(T,G)$ неравенство вида

$$\rho(T,G) \le \frac{c(\cdot)}{\sqrt{n}}, \qquad (2.2)$$

где $c(\cdot)$ является функцией или от ν_3 или от $\rho(F,\Phi)$?

Изучение не имеет общетеоретического характера. Оно проводится в некоторых классах исходных распределений, где функция распределения $G(x)$ получается в случае $n=2$ аналитически, в случае $n>2$ с помощью статистического моделирования. Рассматриваются симметричные и несимметричные исходные распределения.

3. Симметричные исходные распределения.

Статистика T_n имеет простейший вид в случае $n=2$:

$$T_2 = \frac{x_1 + x_2}{x_1 - x_2}. \qquad (3.1)$$

Выведем функцию распределения T_2 в случае равномерного исходного распределения и в случае смеси нормальных распределений с одинаковыми средними значениями Применяется известная формула между функциями плотностей (ф.п.):

$$g(u,v) = f(x_1(u,v), x_2(u,v))\,|J|, \qquad (3.2)$$

где в данном случае

$$u = x_1 + x_2,$$

$$v = x_1 - x_2,$$

$$|J| = 1/2,$$

$$f(x_1,x_2) = f(x_1)f(x_2), \qquad (3.3)$$

$f(x_i)$ - ф.п. случайной величины x_i. Из совместной ф.п. $g(u,v)$ получается ф.п. соотношения $z = u/v$:

$$g(z)=\int_{-\infty}^{\infty} g(zv,v)|v|\,dv, \qquad (3.4)$$

которая и есть ф.п. статистики T_2 .

3.I. Равномерное распределение.

Пусть $x_i \sim U(-\sqrt{3},\sqrt{3})$, т.е. $f(x_i)=\frac{1}{2\sqrt{3}}$ и

$$f(x_1,x_2)=\begin{cases} 1/12 & \text{при } x_1,x_2 \in [-\sqrt{3},\sqrt{3}\,], \\ 0 & \text{при остальных } x_1,x_2. \end{cases}$$

Следовательно,

$$g(u,v)=\frac{1}{24} \quad \text{в области} \begin{cases} -2\sqrt{3}+u \le v \le 2\sqrt{3}-u, & u \ge 0, \\ -2\sqrt{3}-u \le v \le 2\sqrt{3}+u, & u<0. \end{cases} \qquad (3.5)$$

При вычислении интеграла (3.4) нам достаточно рассматривать область изменения v только в случае $u \ge 0$. В противном случае результат будет аналогичным. Заменой переменной $u=zv$ из (3.5) следует :

$$-2\sqrt{3}+zv \le v \le 2\sqrt{3}-zv, \quad zv \ge 0,$$

откуда получается область изменения v через z :

$$-\frac{2\sqrt{3}}{1-z} \le v \le \frac{2\sqrt{3}}{1+z}, \quad zv \ge 0, \qquad (3.6)$$

Чтобы избавиться от знака абсолютного значения, интеграл (3.4) разделяется на две части:

$$g(z)=\int_{v\ge 0} \frac{v}{24}\,dv - \int_{v<0} \frac{v}{24}\,dv. \qquad (3.7)$$

Соответственно неравенствам (3.6) первый интеграл интегрируется в интервале $[0, \frac{2\sqrt{3}}{1+z}\,]$, где предполагается $z \ge 0$, а второй в интервале $[-\frac{2\sqrt{3}}{1-z}, 0\,]$, где $z<0$, т.е. в обоих интервалах в знаменателе стоит $1+|z|$. В результате получается

$$g(z)=\frac{v^2}{48}\Bigg|_{0}^{\frac{2\sqrt{3}}{1+|z|}} - \frac{v^2}{48}\Bigg|_{\frac{-2\sqrt{3}}{1-|z|}}^{0}$$

откуда следует окончательный вид функции плотности статистики T_2 в случае равномерного исходного распределения:

$$g(z) = \frac{1}{2(1+|z|)^2}. \qquad (3.8)$$

Функция распределения статистики T_2 представляется по формуле

$$G(x) = \int_{-\infty}^{x} \frac{dz}{2(1+|z|)^2} = \begin{cases} \frac{1}{2} + \frac{x}{2(1+x)}, & x > 0, \\ \frac{1}{2(1-x)}, & x \le 0, \end{cases}$$

которая при помощи знака абсолютного значения приобретает вид:

$$G(x) = \frac{1}{2} + \frac{x}{2(1+|x|)}. \qquad (3.9)$$

3.2. Смесь нормальных распределений

Пусть функция плотности случайной величины x_i, $i=1,2$ представляется в виде смеси:

$$f(x) = \gamma\Phi_1(x) + (1-\gamma)\Phi_2(x), \qquad (3.10)$$

где

$$0 < \gamma < 1,$$

$$\Phi_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{x^2}{2\sigma_1^2}\right), \qquad (3.11)$$

$$\Phi_2(x) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{x^2}{2\sigma_2^2}\right). \qquad (3.12)$$

Совместная ф.п. случайных величин x_1, x_2 получается умножением выражения (3.10) на себя, откуда при помощи (3.2) следует совместная ф.п. случайных величин $u = x_1 + x_2$, $v = x_1 - x_2$:

$$g(u,v) = \frac{1}{2}\left[\gamma^2 a_1 + \gamma(1-\gamma)a_2 + \gamma(1-\gamma)a_3 + (1-\gamma)^2 a_4\right], \qquad (3.13)$$

где

$$a_1 = \Phi_1\left(\frac{u+v}{2}\right)\Phi_1\left(\frac{u-v}{2}\right), \qquad (3.14)$$

$$a_2 = \Phi_1\left(\frac{u+v}{2}\right)\Phi_2\left(\frac{u-v}{2}\right), \qquad (3.15)$$

$$a_3 = \Phi_2\left(\frac{u+v}{2}\right)\Phi_1\left(\frac{u-v}{2}\right), \qquad (3.16)$$

$$a_4 = \Phi_2\left(\frac{u+v}{2}\right)\Phi_2\left(\frac{u-v}{2}\right). \qquad (3.17)$$

Осуществляя замену переменной $u = zv$, из (3.13) получается:

$$g(zv, v) = \frac{1}{2}\left[\beta^2 a_1^* + \beta(1-\beta)a_2^* + \beta(1-\beta)a_3^* + (1-\beta)^2 a_4^*\right], \qquad (3.18)$$

где слагаемые $a_1^*, \ldots, a_4^*$ после преобразований с использованием выражений (3.11), (3.12) представляются в виде:

$$a_1^* = \frac{1}{2\pi\sigma_1^2}\exp(-c_1 v^2), \text{ где } c_1 = \frac{z^2+1}{4\sigma_1^2}, \qquad (3.19)$$

$$a_2^* = \frac{1}{2\pi\sigma_1\sigma_2}\exp(-c_2 v^2), \text{ где } c_2 = \frac{1}{8}\left[\frac{(z+1)^2}{\sigma_1^2} + \frac{(z-1)^2}{\sigma_2^2}\right], \qquad (3.20)$$

$$a_3^* = \frac{1}{2\pi\sigma_1\sigma_2}\exp(-c_3 v^2), \text{ где } c_3 = \frac{1}{8}\left[\frac{(z+1)^2}{\sigma_2^2} + \frac{(z-1)^2}{\sigma_1^2}\right], \qquad (3.21)$$

$$a_4^* = \frac{1}{2\pi\sigma_2^2}\exp(-c_4 v^2), \text{ где } c_4 = \frac{z^2+1}{4\sigma_2^2}. \qquad (3.22)$$

Вычисляя интеграл

$$g(z) = \int_{-\infty}^{\infty} g(zv, v)|v|\,dv = \int_{-\infty}^{0} g(zv, v)(-v)\,dv + \int_{0}^{\infty} g(zv, v)v\,dv,$$

получим

$$g(z) = \frac{1}{2}\left[\beta^2 \frac{1}{2\pi\sigma_1^2 c_1} + \beta(1-\beta)\frac{1}{2\pi\sigma_1\sigma_2 c_2} + \beta(1-\beta)\frac{1}{2\pi\sigma_1\sigma_2 c_3} + (1-\beta)^2\frac{1}{2\pi\sigma_2^2 c_4}\right].$$

Отсюда, после подставлений $c_1, \ldots, c_4$, следует ф.п. статистики r_2 в случае исходного распределения (3.10):

$$g(z) = \frac{1}{\pi}\left[(2\beta^2 - 2\beta + 1)\frac{1}{z^2+1} + 2\beta(1-\beta)\sigma_1\sigma_2\left(\frac{1}{az^2+2bz+a} + \frac{1}{az^2-2bz+a}\right)\right], \qquad (3.23)$$

где

$$a = \sigma_1^2 + \sigma_2^2, \qquad (3.24)$$

$$b = \sigma_2^2 - \sigma_1^2. \qquad (3.25)$$

Интегрированием

$$G(x) = \int_{-\infty}^{x} g(z)\,dz$$

получается соответственная функция распределения:

$$G(x) = \frac{1}{\pi}\left[(2\beta^2 - 2\beta + 1)\operatorname{arctg} x + \beta(1-\beta)\left(\operatorname{arctg}\frac{ax+b}{2\sigma_1\sigma_2} + \operatorname{arctg}\frac{ax-b}{2\sigma_1\sigma_2}\right)\right] + \frac{1}{2}. \qquad (3.26)$$

3.3. Расстояние от распределения Стьюдента

Рассматривается статистика T_2 с функцией распределения $G(x)$ и вычисляется равномерное расстояние

$$\rho(T, G) = \sup_x |T(x) - G(x)|, \qquad (3.27)$$

где $T(x)$ функция распределения Коши (закон Стьюдента со степенями свободы I):

$$T(x) = \frac{1}{\pi} \operatorname{arctg} x + \frac{1}{2}. \qquad (3.28)$$

В качестве $G(x)$ принимаются выведенные выше законы (3.9) и (3.26).

В случае равномерного исходного распределения, т.е. в случае $G(x)$ в виде (3.9), расстояние представляется в виде

$$\rho(T, G) = \sup_x |h(x)|,$$

где

$$h(x) = \frac{\operatorname{arctg} x}{\pi} - \frac{x}{2(1+|x|)}. \qquad (3.29)$$

Для нахождения максимума $h(x)$ дифференцируем (3.29) по x и приравниваем к нулю. Получим уравнение

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

которое из-за симметричности $T(x)$ и $G(x)$ достаточно решить в области $x \geq 0$. Получим

$$(\pi - 2)x^2 - 4x + \pi - 2 = 0,$$

откуда

$$x_{1,2} = \frac{2 \pm \sqrt{\pi(4-\pi)}}{\pi - 2}$$

Таким образом, максимальное различие между $T(x)$ и $G(x)$ достигается в точках

$$x_1 = 0.3134,$$

$$x_2 = 3.1904$$

и оно равно

$\rho(T,G) = 0.0226$.

В случае исходного распределения в виде смеси нормальных распределений, $G(x)$ выражается формулой (3.26), и соответственное расстояние вычисляется из формулы

$$\rho(T,G) = \frac{\gamma(1-\gamma)}{\sqrt{\pi}} \sup_x |h(x)|, \qquad (3.30)$$

где

$$h(x) = 2\operatorname{arctg} x - \operatorname{arctg}\frac{ax+b}{2\sigma_1\sigma_2} - \operatorname{arctg}\frac{ax-b}{2\sigma_1\sigma_2}. \qquad (3.31)$$

Коэффициенты a и b являются функциями (3.24), (3.25) от σ_1 и σ_2. Дифференцируя $h(x)$ получим уравнение

$$\frac{1}{x^2+1} - \sigma_1\sigma_2\left[\frac{1}{ax^2+2bx+a} + \frac{1}{ax^2-2bx+a}\right] = 0,$$

которое после избавления от знаменателя представляется в виде:

$$rx^4 + 2(r-2b^2)x^2 + r = 0, \qquad (3.32)$$

где

$$r = a^2 - 2\sigma_1\sigma_2 a.$$

Уравнение (3.32) имеет решение

$$x^2_{1,2} = \frac{2b^2 - r \pm 2b\sqrt{b^2-r}}{r},$$

которое после подставления a, b и r приобретает вид:

$$x^2_{1,2} = \frac{\sigma_1^2 + 4\sigma_1\sigma_2 + \sigma_2^2 \pm 2(\sigma_1+\sigma_2)\sqrt{2\sigma_1\sigma_2}}{\sigma_1^2+\sigma_2^2},$$

откуда

$$x^2_{1,2} = \frac{\left(\sigma_1+\sigma_2 \pm \sqrt{2\sigma_1\sigma_2}\right)^2}{\sigma_1^2+\sigma_2^2}.$$

Таким образом, максимальное различие между функциями распределений $T(x)$, $G(x)$ достигается в точках:

$$x_1 = \frac{\sigma_1+\sigma_2+\sqrt{2\sigma_1\sigma_2}}{\sqrt{\sigma_1^2+\sigma_2^2}}, \qquad (3.33)$$

$$x_2 = \frac{\sigma_1+\sigma_2-\sqrt{2\sigma_1\sigma_2}}{\sqrt{\sigma_1^2+\sigma_2^2}}. \qquad (3.34)$$

Умножая x_1 и x_2, заметим, что

$$x_1 = x_2^{-1}$$

Расстояние $\rho(T,G)$ прямо зависит от параметров исходного распределения - $\gamma, \sigma_1, \sigma_2$. Нас интересуют значения этих параметров, причиняющие максимальные расстояния $\rho(T,G)$. Предположим, что $\rho(T,G)$ тем больше, чем больше отличается исходное распределение $F(x)$ от $\Phi(x)$. Поскольку первые моменты исходного распределения зафиксированы — $E x_i = 0$, $E x_i^2 = 1$, $E x_i^3 = 0$, то различие от $\Phi(x)$ определяется четвертым моментом.

Требуется, чтобы четвертый момент $F(x)$ κ раз превысил четвертый момент распределения $\Phi(x)$:

$$E x_i^4 = 3\kappa.$$

Для определения такого исходного распределения $F(x)$, из (3.10) получим систему уравнений:

$$\begin{cases} \gamma \sigma_1^2 + (1-\gamma)\sigma_2^2 = 1, \\ \gamma \sigma_1^4 + (1-\gamma)\sigma_2^4 = \kappa, \end{cases}$$

откуда

$$\sigma_1^2 = 1 - \sqrt{(\kappa-1)\frac{1-\gamma}{\gamma}},$$

$$\sigma_2^2 = 1 + \sqrt{(\kappa-1)\frac{\gamma}{1-\gamma}}.$$

Заметим, что κ может иметь большие значения в случае $\gamma \to 1$. В случае $\gamma = \frac{1}{2}$ для κ имеется условие $1 < \kappa < 2$, т.е. в этом случае мы не можем получить исходного распределения $F(x)$, отличающегося от $\Phi(x)$ в смысле четвертого момента более 2 раз. Исходное распределение $F(x)$ с параметром $\gamma = \frac{1}{2}$ представляет особый интерес из-за того, что в этом случае по (3.30) расстояние $\rho(T,G)$ имеет максимальный коэффициент $\gamma(1-\gamma)$.

3.4. Результаты вычислений в случае симметричных распределений

В качестве исходных распределений рассматривались равномерное распределение $U(-\sqrt{3},\sqrt{3})$ и шесть распределений с законом (3.10) и с параметрами, приведенными в таблице 3.1. В таблицу включены и точки x_1, x_2, вычисленные по формулам (3.33), (3.34), в которых разница между функциями $T(x)$ и $G(x)$ является максимальной. Напомним, что κ выражает соотношение четвертых моментов распределений $F(x)$ и $\Phi(x)$.

Таблица 3.1

Параметры исходного распределения $F(x)$

κ	γ	σ_1	σ_2	x_1	x_2
1.1	0.5	0.827	1.147	2.370	0.422
1.5	"	0.541	1.307	2.148	0.466
1.9	"	0.227	1.396	1.710	0.585
2.0	0.6	0.428	1.492	1.966	0.509
5.0	0.9	0.577	2.646	1.836	0.545
10.0	0.95	0.558	3.752	1.676	0.597

При рассмотренных исходных распределениях вычисляются расстояния $\rho(T,G)$. Результаты приводятся в таблице 3.2, куда занесены и третьи абсолютные моменты ν_3 и расстояния $\rho(F,\Phi)$ исходных распределений. Известно, что при нормальном распределении $\nu_3 = 1.596$.

Таблица 3.2

Расстояние распределения статистики T_2
от распределения Отьюдента

κ	ν_3	$\rho(F,\Phi)$	$\rho(T,G)$
$F(x)$- функция распределения закона $U(-\sqrt{3},\sqrt{3})$			
0.6	I.299	0.057	0.023
$F(x)$- функция распределения смеси (3.I0)			
I.I	I.656	0.007	0.002
I.5	I.906	0.046	0.0I4
I.9	2.I80	0.I32	0.044
2.0	2.I94	0.089	0.024
5.0	3.232	0.I0I	0.0I2
I0.0	4.478	0.I2I	0.009

Прежде всего заметим, что расстояние $\rho(T,G)$ при всех случаях довольно малое, не превышающее значения 0.044, хотя исходное распределение по параметрам $\kappa, \nu_3, \rho(F,\Phi)$ иногда сильно отличается от нормального. При этом расстояние $\rho(T,G)$ всегда меньше расстояния между исходными распределениями (два, три раза и даже больше). Последнее может быть основой при приблизительном оценивании значения $\rho(T,G)$.

При $F(x)$, близкой к нормальному распределению (κ=I.I, ν_3 = I.656, $\rho(F,\Phi)$ = 0.007), $\rho(T,G)$ является близким к нулю (0.002). Интересно отметить, что $\rho(T,G)$ может быть близким к нулю и в случае, когда $F(x)$ в смысле κ, ν_3, ρ сильно отличается от нормального распределения (случай κ = I0). Поскольку в этом случае $F(x)$ является смесью нормальных рас-

пределений с коэффициентом смешивания $\gamma = 0.95$, то отсюда вытекает полезное замечание для практической статистики, а именно, что редкие большие ошибки в нормальных данных почти не влияют на распределение статистики T_2. Самое большое расстояние $\varrho(T,G)=0.044$ достигается в случае, когда коэффициент смешивания $\gamma = \frac{1}{2}$, а дисперсия одного распределения малая, другого по возможности большая ($\sigma_1 = 0.227, \sigma_2 = 1.396$).

Еще видим из таблицы 3.2, что в один раз большому расстоянию $\varrho(F,\Phi)=0.132$ соответствует самое большое расстояние $\varrho(T,G)=0.044$, а в другой раз почти такому же по величине расстоянию $\varrho(F,\Phi)=0.121$ соответствует одно из малейших значений $\varrho(T,G)=0.009$. Следовательно, расстояние между исходными распределениями не может быть той характеристикой, с которой полностью определяется поведение $\varrho(T,G)$. Можно предположить, что расстояние $\varrho(T,G)$ зависит от нескольких характеристик исходного распределения $F(x)$.

4. Несимметричные исходные распределения

Разложение (1.3) показывает, что при симметричных исходных распределениях F, функция распределения статистики T_n стремится к $\Phi(x)$ со скоростью $o(n^{-1/2})$, т.е. быстрее, чем в случае несимметричных F. В данном пункте изучается поведение интересующего нас расстояния $\varrho(T,G)$ в классе несимметричных исходных распределений. В качестве F рассматривается смесь смещенных нормальных распределений:

$$F(x) = \gamma\Phi\left(\frac{x-\mu_1}{\sigma}\right) + (1-\gamma)\Phi\left(\frac{x-\mu_2}{\sigma}\right), \qquad (4.1)$$

где параметры $\gamma, \mu_1, \mu_2, \sigma$ определяются так, чтобы $F(x)$ являлась центрированной, нормированной, с третьим моментом τ,

отличающим от нуля. Для этого решается система уравнений, приведенная в работе Траат (1984). Значение γ фиксируется

$$\gamma = \frac{3+\sqrt{3}}{6} \approx 0.789,$$

гарантирующее равнение четвертого кумулянта $F(x)$ нулю. Таким образом, распределения данного класса совпадают со стандартным нормальным распределением по первому, второму и четвертому моментам, и отличаются по третьему моменту (показатель симметрий).

Рассматриваются четыре различных $F(x)$, параметры которых занесены в таблицу (4.I).

Таблица 4.I

Параметры исходных распределений $F(x)$

№	μ_1	μ_2	σ	τ
I	-0.507	I.893	0.200	I.330
2	-0.423	I.577	0.577	0.770
3	-0.338	I.262	0.757	0.394
4	-0.254	0.946	0.872	0.I66

Соответствующая ф.р. $G(x)$ статистики T_n оценивается при помощи метода Монте-Карло. Поскольку проводится I000 повторений, полученные расстояния являются довольно хорошими аппроксимациями для $\varrho(T,G)$. Результаты вычислений занесены в таблицу 4.2, где в каждой строке находятся значения $\varrho(T,G)$ при конкретном исходном распределении и разных значениях n. В начале таблицы приводятся характеристики рассматриваемого исходного распределения. Целочисленными значениями в первой строке таблицы 4.2 обозначаются значе-

ния n .

Таблица 4.2

Расстояния распределения статистики T_n и распределения Стьюдента - $\rho(T,G)$

τ	ν_3	$\rho(F,\Phi)$	3	4	5	6	7	8	9
I.330	I.446	0.320	0.77I	0.3I4	0.279	0.209	0.I74	0.I55	0.II9
0.770	I.600	0.I07	0.II6	0.084	0.07I	0.070	0.063	0.067	0.053
0.394	I.60I	0.04I	0.045	0.043	0.028	0.038	0.059	0.053	0.025
0.I66	I.598	0.0I4	0.033	0.032	0.020	0.032	0.023	0.027	0.038

Продолжение таблицы 4.2

I0	II	I2	I3	I4	I5	I6	I7	I8
0.089	0.076	0.08I	0.087	0.083	0.078	0.057	0.067	0.064
0.05I	0.043	0.062	0.04I	0.034	0.060	0.037	0.038	0.03I
0.055	0.0I9	0.04I	0.024	0.032	0.024	0.03I	0.033	0.024
0.028	0.025	0.026	0.027	0.022	0.033	0.022	0.0I5	0.03I

Поскольку в данном случае мы имеем дело с расстояниями между эмпирической ф.р. и ф.р. Стьюдента, то с помощью теста Колмогорова-Смирнова можем проверить гипотезу

$$H_0: \rho(T,G) = 0.$$

Критическое значение теста вычисляется по асимптотической формуле

$$\lambda_{\alpha,K} \approx \sqrt{-\frac{\ln(\alpha/2)}{2K}},$$

где

K =I000 в нашем эксперименте,

α - уровень значимости.

На уровне значимости $\alpha = 0.05$, имеем $\lambda = 0.043$ и, следовательно, гипотеза H_0 отвергается при всех рассматриваемых значениях n, если исходное распределение имеет коэффициент асимметрии $\gamma = 1.330$. В случае $\gamma = 0.77$, гипотезу нельзя отвергать начиная с $n = 16$; в случае $\gamma = 0.394$, начиная с $n = 10$. Если $\gamma = 0.166$, то гипотеза H_0 принимается при всех значениях n. Видно, что асимметричность исходного распределения сильно влияет на величину $\rho(T, G)$. Расстояние $\rho(T, G)$ (как и было видно в предыдущем пункте) имеет тенденцию быть меньше расстояния между исходными распределениями. Основным исключением является случай $n = 3$. Напомним, что в случае $\rho(F, \Phi) = 0.014$ расстояние $\rho(T, G) = 0$ по гипотезе H_0.

Утверждением известных фактов является стремление $\rho(T, G)$ к нулю при возрастании n.

Самым интересным является вытекающий из таблицы факт, что исходным распределениям с одинаковыми третьими абсолютными моментами соответствуют отличающиеся по величине расстояния $\rho(T, G)$. Следовательно, ν_3 не может быть той характеристикой исходного распределения, с помощью которой возможно точно оценить расстояния $\rho(T, G)$.

5. Заключение

На основе полученных результатов можно сказать, что расстояние между ф.р. статистики T_n и ф.р. Стьюдента - $\rho(T, G)$ всегда меньше, чем расстояние между исходными распределениями - $\rho(F, \Phi)$. Исключением являются случаи сильной асимметрии при малом значении n. Хотя $\rho(F, \Phi)$ можно считать первичной оценкой для $\rho(T, G)$, оно все же является довольно грубой. Большим значениям $\rho(F, \Phi)$ могут отвечать очень

маленькие значения $\varrho(T,G)$. При изучении неравенства

$$\varrho(T,G) \le \frac{c(\cdot)}{\sqrt{n}},$$

мы нашли, что рассматриваемые характеристики исходного распределения, т.е. расстояние $\varrho(F,\Phi)$ и третий абсолютный момент γ_3 не определяют однозначно расстояния $\varrho(T,G)$, и, таким образом, функция $c(\cdot)$ от этих характеристик не может обеспечивать минимальную верхнюю границу для $\varrho(T,G)$.

Литература

1. Золотарев В.М. Предельное распределение статистики Стьюдента в случае негауссовской генеральной совокупности. - Проблемы устойчивости стохастических моделей. Труды семинара. М.: ВНИИСИ, 1985, с. 57-63.
2. Траат И.К. Представление неизвестных распределений статистик с помощью смеси нормальных распределений. - Тр. Вычисл. центра Тартуск. ун-та, 1984, вып. 51, с.126-134.
3. Hall P. Edgeworth expansion for Student's t statistic under minimal moment conditions. - Ann. Prob., 1987, vol. 15, N° 3, 920-931.
4. Slavova V.V. On the Berry-Esseen bound for Student's statistic. - Lect. Notes Math., 1985, B.1155, 355-390.

Поступило 5.10 . 1988

Resümee

Student'i statistiku jaotuse hindamisest

I.Traat

Student'i statistiku T_n (1.1) jaotusfunktsiooni $G(x)$ võrreldakse Student'i jaotusfunktsiooniga $T(x)$ vabadusastmete arvuga $n-1$, kus n on valimimaht. Selleks kasutatakse ühtlast kaugust $\varrho(T,G) = \sup_x |T(x) - G(x)|$. Üldkogumi jaotusena F vaadeldakse ühtlast jaotust ja normaaljaotuste risustustüüpi ja nihke tüüpi segusid. Esimese kahe jaotuse jaoks tuletatakse $n=2$ korral jaotusfunktsiooni $G(x)$ avaldis. Nihke tüüpi segu korral hinnatakse $G(x)$ Monte-Carlo meetodil. Uuritakse kauguse $\varrho(T,G)$ sõltuvust jaotuse F niisugustest parameetritest, nagu 3. absoluutne moment ν_3 ja ühtlane kaugus normaaljaotusest $\varrho(F,\Phi)$.

Tulemusena saadi, et üldjuhul $\varrho(T,G)$ on väiksem kui $\varrho(F,\Phi)$, sümmeetrilise F korral isegi mitu korda väiksem. Seetõttu võimaldab jaotuste F ja Φ vaheline kaugus esialgselt prognoosida statistiku T_n jaotuse erinevust Student'i jaotusest. Osutus aga, et ei ν_3 ega $\varrho(F,\Phi)$ määra üheselt kaugust $\varrho(T,G)$, st. erinevad üldkogumijaotused sama ν_3 või sama $\varrho(F,\Phi)$-ga põhjustavad erinevaid kaugusi $\varrho(T,G)$. Seega ν_3 ja $\varrho(F,\Phi)$ ei ole ka sobivad argumendid funktsioonile $c(\cdot)$ minimaalse ülemise tõkke määramisel võrratuses $\varrho(T,G) \leq c(\cdot)/\sqrt{n}$.

Summary

On estimation of the distribution of Student's statistic

I.Traat

The distribution function (d.f.) $G(x)$ of Student's statistic T_n (1.1) is compared with the Student's d.f. $T(x)$ with degrees of freedom $n-1$, where n is sample size. For that the uniform distance $\varrho(T,G)=\sup|T(x)-G(x)|$ is used. The population distribution is considered to be uniform, contamination and shifted type normal mixture in this paper. In the special case $n=2$ for the first two distributions the expression of $G(x)$

is derived. For the shifted type mixture $G(x)$ is estimated with the help of statistical simulation. The influence of the population distribution to the distance $\varrho(T,G)$ is investigated. The population d.f. is characterized by its third absolute moment ν_3 and by the uniform distance from normal distribution $\varrho(F,\Phi)$.

As a result it was obtained that in general $\varrho(T,G)$ is smaller than $\varrho(F,\Phi)$, for symmetric $F(x)$ even many times smaller. This fact may be used for preliminary estimation of $\varrho(T,G)$. Unfortunately $\varrho(F,\Phi)$ and also ν_3 are not suitable for more exact estimation of $\varrho(T,G)$, because they do not determine it uniquely, i.e., the population distributions with equal moments ν_3 or with equal distances $\varrho(F,\Phi)$ cause the different values of $\varrho(T,G)$. Thus neither ν_3 nor $\varrho(F,\Phi)$ can be suitable argumets of the function $c(\cdot)$ for determing the minimal upper bound in the inequality $\varrho(T,G) \leq c(\cdot)/\sqrt{n}$.

Содержание

1 руб. 10 коп.