

УДК 519.2 (311.2)

Ю.Г. Дмитриев, С.В. Курицина

ОБ ИСПОЛЬЗОВАНИИ ДОПОЛНИТЕЛЬНОЙ ИНФОРМАЦИИ В СТАТИСТИЧЕСКОМ ОЦЕНИВАНИИ ПАРАМЕТРОВ ДЕТЕРМИНАЦИОННОГО АНАЛИЗА

Рассматривается задача оценивания параметров детерминационного анализа в предположении, что имеется некоторая дополнительная информация о распределении признаков. Строятся различные оценки с использованием этой информации и изучаются их свойства. Анализируется влияние учета информации на качество оценивания, приводятся примеры применения оценок в анализе реальных данных.

Ключевые слова: *детерминационный анализ, интенсивность, емкость, дополнительная информация, статистические оценки, таблица сопряженности.*

Одним из методов анализа социально-экономических данных является детерминационный анализ (ДА), предложенный С.В. Чесноковым [1]. Этот метод служит для поиска и интерпретации взаимосвязей между переменными (признаками) или группами переменных на основе эмпирических материалов обследований и представляет собой вариант исчисления эмпирических условных частот (долей), которые содержатся в таблицах сопряженности. При этом устанавливаются ситуации, в которых по конкретным значениям одних индикаторов можно с известной определенностью предсказывать значения других. Важными характеристиками ДА являются интенсивность и емкость, вычисляемые по данным таблицы сопряженности выделенных признаков и представляющие собой оценки условных долей. Точность вычислений этих характеристик можно повысить, если наряду с исходными данными использовать имеющуюся дополнительную информацию об исследуемой генеральной совокупности. В данной работе дополнительная информация выступает в виде знания долей (или числа) объектов генеральной совокупности с заданными значениями, как изучаемых, так и других признаков. Предлагается подход к учету такого вида информации при построении статистических оценок долей, интенсивности и емкости, анализируется выигрыш в точности оценивания, приводятся примеры практического применения.

1. Оценки интенсивности и емкости с учетом дополнительной информации

Приведем основные понятия ДА. Пусть у объектов некоторой совокупности выделены два признаками X и Y . Среди множества значений признака X имеется значение a , а у признака Y – значение b . Согласно [1], правило как детерминация – это условное суждение вида: «Если a , то b » (или сокращенно $a \rightarrow b$) вместе с двумя своими характеристиками – точностью и полнотой, которые характеризуются соответственно интенсивностью и ёмкостью детерминации, являющимися условными долями:

$$I(a \rightarrow b) = \frac{N(a, b)}{N(a)}, \quad C(a \rightarrow b) = \frac{N(a, b)}{N(b)}. \quad (1)$$

Здесь $N(a, b)$ – количество объектов генеральной совокупности, у которых одновременно $X = a$ и $Y = b$, $N(a)$ – количество объектов с $X = a$, $N(b)$ – количество объектов с $Y = b$. Интенсивность выражает точность детерминации, а ёмкость – ее полноту. Интенсивность и ёмкость обычно подсчитываются по таблице сопряженности выделенных признаков, построенной на основании случайной выборки объема n .

Рассмотрим задачу статистического оценивания характеристик (1) в терминах событий $A = \{X = a\}$, $B = \{Y = b\}$, $AB = \{X = a, Y = b\}$ и соответствующих вероятностей $P(A)$, $P(B)$ и $P(AB)$. Тогда интенсивность и емкость детерминации $a \rightarrow b$, которую обозначим как $A \rightarrow B$, примут вид

$$I(A \rightarrow B) = P(B|A) = \frac{P(AB)}{P(A)}, \quad C(A \rightarrow B) = P(A|B) = \frac{P(AB)}{P(B)}. \quad (2)$$

Пусть случайной выборке объема n соответствуют эмпирические вероятности

$$P_n(A) = \frac{1}{n} \sum_{i=1}^n I_i(A), \quad P_n(B) = \frac{1}{n} \sum_{i=1}^n I_i(B), \quad P_n(AB) = \frac{1}{n} \sum_{i=1}^n I_i(AB),$$

где $I(\cdot)$ – индикаторная функция соответствующего события. Подставив эти эмпирические вероятности в (2), получим статистические оценки функции интенсивности и емкости в виде

$$J_n = \frac{P_n(AB)}{P_n(A)}, \quad G_n = \frac{P_n(AB)}{P_n(B)}. \quad (3)$$

В объектах, попавших в выборку, могут быть и другие наблюдаемые признаки. Пусть признак Z один из них. Рассмотрим событие $C = \{Z = c\}$, для которого известна вероятность $P(C)$. Другими словами, известна доля объектов генеральной совокупности, у которых Z имеет заданное значение c . Используем эту информацию для построения оценок вероятностей других событий и характеристик (2). Рассмотрим оценки, основанные на использовании формулы полной вероятности [2]:

$$\hat{P}_n(B) = P_n(B|C)P(C) + P_n(B|\bar{C})P(\bar{C}) = \frac{P_n(BC)}{P_n(C)}P(C) + \frac{P_n(B\bar{C})}{P_n(\bar{C})}P(\bar{C}), \quad (4)$$

где $\bar{C}$ – событие противоположное событию C . Применив формулу (4) для оценивания вероятности $P(A)$, методом подстановки получим оценки характеристик (2), с учетом имеющейся информации. Анализ свойств получающихся оценок и их сравнение с оценками (3) проведем на примере оценивания интенсивности детерминации. Далее предполагается, что объем генеральной совокупности достаточно большой по сравнению с объемом выборки, так что поправкой на бесповторность можно пренебречь, и асимптотические соотношения выполняются.

Метод подстановки приводит к следующей оценке интенсивности с учетом дополнительной информации:

$$\hat{J}_n = J(\hat{P}_n(AB), \hat{P}_n(A)) = \frac{\hat{P}_n(AB)}{\hat{P}_n(A)} = \frac{\frac{P_n(ABC)}{P_n(C)}P(C) + \frac{P_n(AB\bar{C})}{P_n(\bar{C})}P(\bar{C})}{\frac{P_n(AC)}{P_n(C)}P(C) + \frac{P_n(A\bar{C})}{P_n(\bar{C})}P(\bar{C})}. \quad (5)$$

Исследуем асимптотическое поведение данной оценки. Поскольку при увеличении n оценки вероятностей сходятся по вероятности к истинным значениям, то можно провести разложение $\hat{J}_n$ в окрестности истинных вероятностей по формуле Тейлора с остатком в форме Лагранжа. Имеем

$$\begin{aligned} \hat{J}_n = & \frac{P(AB)}{P(A)} + \frac{1}{P(A)} \left(P_n(ABC) - P(ABC) + P_n(AB\bar{C}) - P(AB\bar{C}) \right) - \frac{P(AB)}{P^2(A)} (P_n(AC) - \\ & - P(AC) + P_n(A\bar{C}) - P(A\bar{C})) + \frac{1}{P(A)} \left[-\frac{P(ABC) - P(AB)P(C)}{P(C)(1 - P(C))} \right] (P_n(C) - P(C)) + \\ & + \frac{P(AB)}{P^2(A)} \left[\frac{P(AC) - P(A)P(C)}{P(C)(1 - P(C))} \right] (P_n(C) - P(C)) + R_n. \end{aligned} \quad (6)$$

Обозначим главную часть в (6) через $\hat{J}_g$. Анализ показывает, что математическое ожидание и дисперсия главной части удовлетворяют соотношениям

$$\begin{aligned} M\{\hat{J}_g\} &= J, \quad nD\{\hat{J}_g\} = \sigma^2 = \\ &= \frac{P(AB)}{P^2(A)} \left(1 - \frac{P(AB)}{P(A)} \right) - \frac{1}{P(C)(1 - P(C))} \left[\frac{P(ABC) - P(AB)P(C)}{P(A)} - \right. \\ & \quad \left. - \frac{P(AB)}{P^2(A)} (P(AC) - P(A)P(C)) \right]^2. \end{aligned} \quad (7)$$

Рассмотрим выражение

$$\sqrt{n}(\hat{J}_n - J) = \sqrt{n}(\hat{J}_g - J) + \sqrt{n}R_n.$$

В силу центральной предельной теоремы последовательность

$$\sqrt{n}(\hat{J}_g - J) \sim N(0, \sigma_2^2),$$

т.е. имеет асимптотически нормальное распределение с нулевым математическим ожиданием и дисперсией, определяемыми формулой (7).

Последовательность $\sqrt{n} \cdot R_n$ слабо сходится к нулю в силу теоремы непрерывности (см. [3, гл. 6]). Следовательно, последовательность $\sqrt{n}(\hat{J}_n - J)$ имеет асимптотически нормальное распределение с нулевым математическим ожиданием и дисперсией σ_2^2 . Аналогично показывается, что оценка интенсивности без учета информации (см. (3)) является асимптотически нормальной, т.е. $\sqrt{n} \cdot (J_n - J) \sim N(0, \sigma_1^2)$ с дисперсией

$$\sigma_1^2 = \frac{P(AB)}{P^2(A)} \left(1 - \frac{P(AB)}{P(A)} \right). \quad (8)$$

Сравнение двух выражений (7) и (8) показывает, что учет дополнительной информации в оценивании приводит к уменьшению асимптотической дисперсии оценки на величину

$$\delta = \frac{1}{P(C) \cdot (1 - P(C))} \left[\frac{P(ABC) - P(AB)P(C)}{P(A)} - \frac{P(AB)}{P^2(A)} (P(AC) - P(A)P(C)) \right]^2. \quad (9)$$

Из (9) следует, что при независимости событий A , B и C уменьшения асимптотической дисперсии нет, так как $\delta = 0$. Выигрыш в точности можно характеризовать отношением асимптотических дисперсий, т.е. величиной

$$W_1 = \frac{\sigma_1^2}{\sigma_2^2} = 1 - \frac{[P(A)(P(ABC) - P(AB)P(C)) - P(AB)(P(AC) - P(A)P(C))]^2}{P(A)P(C)P(\bar{C})P(AB)(P(A) - P(AB))}. \quad (10)$$

Формула (10) показывает, что чем W_1 ближе к нулю, тем сильнее влияние учета дополнительной информации на точность оценивания интенсивности, и чем ближе к единице, тем слабее это влияние. При независимости событий A и C величина $W_1 = 1$ и выигрыша в точности оценивания нет.

Из формул (7) – (10), в частности, следуют выражения для асимптотической дисперсии и характеристик выигрышей в точности для оценки (4). Для этого нужно предположить, что событие A всегда наступает (является достоверным событием). В этом случае, полагая $P(A) = 1$, $P(AB) = P(B)$, $P(AC) = P(C)$, $P(ABC) = P(BC)$, получим

$$\sigma_2^2 = P(B)(1 - P(B)) - \frac{[P(BC) - P(B)P(C)]^2}{P(C)(1 - P(C))}, \quad \delta_1 = \frac{[P(BC) - P(B)P(C)]^2}{P(C)(1 - P(C))}; \quad (11)$$

$$V_1 = 1 - \frac{[P(BC) - P(B)P(C)]^2}{P(C)(1 - P(C))P(B)(1 - P(B))}. \quad (12)$$

Данные соотношения применяются для анализа точности оценок вероятностей и других событий.

Аналогичным образом анализируются свойства статистической оценки для ёмкости детерминации

$$G = C(A \rightarrow B) = P(A|B) = \frac{P(AB)}{P(B)}.$$

Статистическая оценка без учета дополнительной информации дается формулой (3), а с учетом информации формулой

$$\hat{G}_n = G_n(\hat{P}_n(AB), \hat{P}_n(B)) = \frac{\hat{P}_n(AB)}{\hat{P}_n(B)} = \frac{\frac{P_n(ABC)}{P_n(C)}P(C) + \frac{P_n(AB\bar{C})}{P_n(\bar{C})}P(\bar{C})}{\frac{P_n(BC)}{P_n(C)}P(C) + \frac{P_n(B\bar{C})}{P_n(\bar{C})}P(\bar{C})}. \quad (13)$$

Рассуждениями, аналогичными вышеприведенным, можно показать, что асимптотические свойства оценок (3) и (13) следующие:

$$\sqrt{n}(G_n - G) \sim N(0, \sigma_1^2), \quad \sqrt{n} \cdot (G_n^* - G) \sim N(0, \sigma_2^2),$$

где асимптотические дисперсии имеют вид

$$\bar{\sigma}_1^2 = \frac{P(AB)}{P^2(B)} \left(1 - \frac{P(AB)}{P(B)} \right), \quad (14)$$

$$\bar{\sigma}_2^2 = \bar{\sigma}_1^2 - \frac{1}{P(C)P(\bar{C})} \left[\frac{P(ABC) - P(AB)P(C)}{P(B)} - \frac{P(AB)(P(BC) - P(B)P(C))}{P^2(B)} \right]^2. \quad (15)$$

Второе слагаемое в (15) характеризует величину, на которую уменьшается асимптотическая дисперсия при учете имеющейся информации по сравнению с (14). Отношение асимптотических дисперсий оценок ёмкости с учетом дополнительной информации и без её учета характеризуется величиной

$$W_2 = \frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2} = 1 - \frac{[P(B)(P(ABC) - P(AB)P(C)) - P(AB)(P(BC) - P(B)P(C))]^2}{P(B)P(C)P(\bar{C})P(AB)(P(B) - P(AB))}. \quad (16)$$

Формула (16) показывает, что чем W_2 ближе к нулю, тем сильнее влияние учета дополнительной информации на точность оценивания ёмкости, а чем ближе к единице, тем слабее это влияние. В случае независимости событий B и C , а также независимости A, B, C , величина $W_2 = 1$ и выигрыша в точности оценивания нет.

2. Анализ данных социологического опроса

Пусть исследователя интересует предпочтение населения к тому или иному кандидату перед предстоящими выборами. Производится опрос населения, и по результатам опроса делаются выводы и заключения. Респондент – это объект наблюдения. Выделим два признака: X – пол респондента (значение $X = a$ – респондент мужчина, $X = \bar{a}$ – респондент женщина), Y – фамилия кандидата. Пусть интерес представляет кандидат b . Детерминация $a \rightarrow b$ – это высказывание «если респондент мужчина, то он проголосует за кандидата b »; данному высказыванию приписывается интенсивность $I(a \rightarrow b)$, отражающая его точность, или истинность. Ёмкость $C(a \rightarrow b)$ измеряет долю случаев реализации голосования за b , которая объясняется высказыванием «из a следует b », тем самым она отражает полноту этой детерминации. Интенсивность и ёмкость выражаются через вероятности $P(A) = P\{X = a\}$, $P(B) = P\{Y = b\}$, $P(AB) = P\{X = a, Y = b\}$ по формуле (2).

Приведем пример расчетов для реальных данных. В поселении №1 опрошено 319 респондентов и по результатам опроса составлена таблица сопряженности двух признаков (табл. 1). Кроме того, априори известно общее число избирателей поселения – 6380 человек, среди них мужчин – 3780, а женщин – 2600 человек. Таким образом, доля избирателей мужчин составляет $3780/6380 = 0,5925$, а женщин соответственно $0,4075$. Это знание используем далее в обработке данных.

Таблица 1

Таблица сопряженности 2×2 по переменным X и Y
(объем выборки – 319 человек)

	Y	
$y = \bar{b}$	38	152
$y = b$	90	39
$x = a$	$x = \bar{a}$	X

Из табл. 1 следует, что

$$\begin{aligned} P_n(B) &= \frac{129}{319} \approx 0,4044, \quad P_n(\bar{B}) = \frac{190}{319}, \\ P_n(A) &= \frac{128}{319} \approx 0,4013, \quad P_n(\bar{A}) = \frac{191}{319}, \\ P_n(AB) &= \frac{90}{319} \approx 0,2821; \end{aligned} \quad (17)$$

$$J_n = \frac{P_n(AB)}{P_n(A)} = \frac{90}{128} \approx 0,7029, \quad G_n = \frac{P_n(AB)}{P_n(B)} = \frac{90}{129} \approx 0,6977. \quad (18)$$

Таким образом, за кандидата b проголосовало 40,44 % от общего числа опрошенных респондентов. Эту величину можно уточнить, если учесть имеющуюся информацию о долях избирателей мужчин и женщин в поселении. Если исходить из общих формул, то в нашем случае это означает, что признак Z совпадает с признаком X ($C = A$), $P(A) = 0,5925$, поэтому $P(C) = P(A) = 0,5925$. Данную информацию учтем с помощью формулы (4), полагая в ней $C = A$ и принимая во внимание вычисленные выше оценки вероятностей. Получим уточненную оценку $\hat{P}_n(B) \approx 0,4998 = 49,98\%$, которая отличается от значения $P_n(B)$ на 9,54 % в большую сторону. Заметим, что для оценивания интенсивности знание о долях избирателей по полу ничего не дает, поскольку при $C = A$ из (5) следует, что $\hat{J}_n = J_n$ и $W_1 = 1$.

Для вычисления показателей выигрыша в точности нужно воспользоваться соотношениями (11) и (12). Однако в них не все вероятности известны, что делает невозможным вычисление характеристик точности на практике. В реальности эти характеристики можно оценить по имеющимся данным. Рассмотрим два способа оценивания и сравним их результаты. Первый способ основан на использовании значений (17), что приводит к оценкам характеристик

$$\delta_{n1} = 0,0597 \text{ и } V_{n1} = 0,7517. \quad (19)$$

Второй способ основан на оценках вероятностей, построенных с учетом информации. Вычисления по имеющимся данным приводят к следующим значениям:

$$\hat{P}_n(AB) = \frac{P_n(AB)}{P_n(A)} P(A) \approx 0,4166, \quad \hat{P}_n(A) = P(A) \approx 0,5925, \quad \hat{G}_n \approx 0,8335. \quad (20)$$

$$\hat{\delta}_{n1} = 0,0601, \quad \hat{V}_{n1} = 0,7595, \quad \hat{W}_{n2} = 0,7264. \quad (21)$$

Сравнение результатов (17)–(21) показывает, что значения оценок вероятностей и емкости без учета информации заметно отличаются от значений оценок вероятностей и емкости с учетом информации. Однако при этом значения оценок характеристик точности практически совпадают. Значения величин V_{n1} , $\hat{V}_{n1}$, $\hat{W}_{n2}$ говорят о том, что учет знания долей избирателей по полу при оценивании вероятностей позволяет уменьшить объем выборки на 24–25 % для достижения одинаковой точности с обычными оценками, а при оценивании емкости – на 27 %.

Пусть признак Z означает возраст респондента и значение c есть возрастной интервал (18–34) лет. Известно, что доля избирателей рассматриваемого поселения в таком возрастном интервале равна 0,35, т.е. число таких избирателей равно 2233 и $P(C) = P\{Z = c\} = 0,35$. Рассмотрим таблицу сопряженности для трех признаков X , Y , Z в следующем виде:

Таблица 2

 Таблица сопряженности 2×2 по переменным X , Y , Z ($n = 319$)

		Z		
c	34	26	16	13
$\bar{c}$	56	12	23	139
ab	$a\bar{b}$	$\bar{a}b$	$\bar{a}\bar{b}$	XY

Применим приведенные выше формулы для расчета оценок вероятностей, емкости и интенсивности с учетом информации о доли избирателей по возрасту (поемим их волной и рядом с оценкой дадим ее точность). Имеем

$$\begin{aligned}\tilde{P}_n(B) &\approx 0,4199 = 41,99\% ; \tilde{V}_{n1} = 0,99918 ; \\ \tilde{P}_n(A) &\approx 0,4281 ; \tilde{V}_{n1} = 0,99919 ; \tilde{P}_n(AB) \approx 0,2919 ; \\ \tilde{J}_n &\approx 0,5105 ; \tilde{W}_{n1} = 0,99999 ; \tilde{G}_n \approx 0,6953 ; \tilde{W}_{n2} = 0,99999 .\end{aligned}\quad (22)$$

Из полученных вычислений следует, что влияние возрастной структуры совокупности на точность оценивания очень слабое, однако интенсивность уменьшилась, а емкость практически не изменилась. Оценки вероятностей изменили свои значения, что дает возможность строить другие доверительные интервалы с одинаковой вероятностью доверия.

Детерминация $c \rightarrow b$ – это высказывание «если возраст респондента в интервале (18–34) лет, то он проголосует за кандидата b ». Интенсивность и емкость этой детерминации по данным табл. 2 равны соответственно

$$\begin{aligned}J_n(c \rightarrow b) &= \frac{P_n(CB)}{P_n(C)} = \frac{50}{89} \approx 0,5618 , \\ G_n(c \rightarrow b) &= \frac{P_n(CB)}{P_n(B)} = \frac{50}{129} \approx 0,3876 .\end{aligned}\quad (23)$$

Учет информации о долях избирателей по возрасту приводит к следующему:

$$\bar{J}_n = \frac{\bar{P}_n(CB)}{\bar{P}_n(C)} = J_n(c \rightarrow b) , \quad \bar{G}_n(c \rightarrow b) = 0,4683 .$$

Таким образом, интенсивность не изменилась, а емкость изменилась в сторону увеличения.

Заключение

В работе показано, каким образом можно привлекать знания (дополнительную информацию) о долях объектов генеральной совокупности по признакам и их значениям для повышения точности оценивания различных характеристик. Рассмотренный пример, основанный на реальных данных, позволил увидеть возможности подхода к построению оценок. Подход допускает обобщение на большее число значений признаков, чем рассмотрено в работе. Таблицы сопряженности могут быть гораздо большей размерности. Показано, что существенное повышение точности не всегда происходит. Формулы, характеризующие точность оценивания, дают возможность подбирать признаки, их комбинацию, конкретные значения признаков, которые приводят к существенному повышению точности при фиксированном объеме наблюдений или сокращению объема выборки при заданной точности. Показано, что сокращение объема данных может быть существенным. Асимптотическая нормальность оценок с учетом информации позволяет строить доверительные интервалы, которые могут быть значительно уже по сравнению с обычными оценками. Детерминационный анализ можно осуществлять по модифицированной таблице сопряженности, заменив обычные эмпирические частоты на их модификацию с учетом информации.

ЛИТЕРАТУРА

1. *Чесноков С.В.* Детерминационный анализ социально-экономических данных. М: Наука, 1982. 214 с.
2. *Тарима С.С.* Использование дополнительной информации при оценке вероятностей и интерпретации натурального эксперимента: дис. ... канд. техн. наук, Томск: ТГУ, 2002. 149 с.
3. *Боровков А.А.* Математическая статистика. М: Наука, 2007. 704 с.

Дмитриев Юрий Глебович

Курицина Светлана Валерьевна

Томский государственный университет

E-mail: dmit@mail.tsu.ru, sniksa1174@vtomske.ru

Поступила в редакцию 4 июня 2012 г.

Dmitriev Yury G., Kuritsina Svetlana V. (Tomsk State University). **On the use of an additional information in statistical estimate parameter of determinate analysis.**

Keywords: determinate analysis, intensity, capacity, additional information, statistical estimate, contingency table.

A problem of estimating of parameter of determinate analysis is considered under assumption that certain additional information concerning features is available. Different estimators of parameter using an additional information are proposed, and their properties are studied. The impact of using of additional information on the accuracy of the estimators is studied, examples of using in the analysis of real data are given.