УДК 519.2

Ю.Г. Дмитриев

ОБ ОЦЕНКАХ ВЕРОЯТНОСТЕЙ ПРИ НАЛИЧИИ ДАННЫХ С ПРОПУСКАМИ

Рассматривается задача статистической вероятностей событий на основе комплектных и некомплектных наблюдений. Предлагаются оценки с привлечением дополнительной информации, содержащейся в некомплектных наблюдениях, а также имеющейся априори, исследуются свойства оценок.

Ключевые слова: *комплектные и некомплектные наблюдения*, *дополнительная информация*, *статистическая оценка*, *таблица сопряженности.*

Объекты наблюдений в социологических [1], экономических и маркетинговых исследованиях [2] характеризуются многомерным вектором признаков, которые могут быть как непрерывными, так и дискретными переменными. В процессе наблюдения объектов случаются пропуски в компонентах вектора признаков, что приводит к некомплектным наблюдениям и ставит вопрос об их использовании в анализе данных. Довольно часто их просто исключают из рассмотрения. В других случаях пытаются заполнить пропуски, используя различные приемы, и увеличить число комплектных наблюдений. Главной задачей выборочного метода является статистическое оценивание долей объектов с заданными значениями признаков и анализ соотношения этих долей в соответствии с целями исследования. В этой связи проблема оценки долей при наличии пропусков представляет важную научную и практическую задачу. В статистической практике известны методы статистического анализа данных с пропусками [3,4]. Кроме некомплектных наблюдений исследователь дополнительно может располагать априорной информацией о долях объектов в генеральной совокупности с заданными значениями признаков.

В связи с этим представляет интерес разработка методов статистического анализа данных и построения оценок с одновременным использованием всей имеющейся информации, как априорной, так и эмпирической, содержащейся в некомплектных наблюдениях. Рассмотрение этой задачи на примере оценивания вероятности событий по наблюдениям многомерного вектора категориальных признаков приводится в данной работе.

Практическое применение указанных оценок возникает в выборочных обследованиях некоторых совокупностей, когда требуется оценить долю объектов с заданным значением некоторого признака в случае известной доли объектов с заданным значением другого признака. Так, например, при выявлении предпочтений избирателей некоторой территории к тому или иному кандидату или партии проводятся выборочные опросы людей и оцениваются доли избирателей, которые будут голосовать за конкретного кандидата или партию. При этом о населении территории всегда имеется разнообразная статистическая информация (половая, возрастная, национальная, образовательная структура населения и т.д.), которую можно использовать в оценивании долей с целью повышения точности оценок или сокращения объема наблюдений при заданной точности оценивания.

1. Постановка задачи

Пусть объекты характеризуются r-мерным вектором $(X_1,\ldots,X_r)$, компоненты которого принимают конечное число значений. Из генеральной совокупности методом случайной выборки отобраны объекты, составлена матрица данных и результаты измерений сведены в таблицу сопряженности признаков. В некоторых компонентах вектора часть измерений отсутствует. Будем считать эти пропуски случайными. Наблюдения вектора признаков, в которых пропусков нет, назовем комплектными, в противном случае – некомплектными. Компоненты X_l принимают значения $a_{lm_l}, l = 1,...,r; m_l = 1,...,s_r$ с вероятностями $P(A_{lm_l}) = P\{X_l = a_{lm_l}\}$. Нас будут интересовать как вероятности событий $A_{lm_l} = \{X_l = a_{lm_l}\}$, так и других всевозможных событий, связанных с ними. В работе интересующее нас событие будем обозначать через A, опуская для простоты сопутствующие индексы. Полную группу событий также будем обозначать единообразно $H = (H_1,\ldots,H_k)$. Разбиения множеств могут быть различными как по составу событий, так и по их числу. В частности, это может быть разбиение, связанное с конкретным признаком, например X_r, тогда $H_j = A_{rj}, j = 1,...,s_r, P(H_j) = P(A_{rj}),\ k = s_r$. Разбиения могу быть по паре признаков и т.д. Эмпирическими вероятностями (относительными частотами) событий являются

$$P_n(A_{lm_l}) = \frac{1}{n}\sum_{i=1}^{n} I_i(A_{lm_l}),$$

где $I(\cdot)$ – индикаторная функция соответствующего события**,** n – объем выборки.

Рассмотрим задачу оценивания $P(A)$, используя наряду с комплектными и некомплектные наблюдения с целью повышения точности оценки.

2. Структура несмещенной оценки

Пусть имеется случайная выборка объема n, по которой необходимо оценить вероятность некоторого события $P(A)$ при условии, что известны вероятности $P(H_j)$, $j = 1,\ldots, k$, где совокупность событий $H = (H_1,\ldots,H_k)$ образует полную группу событий. Данную информацию можно использовать в структуре оценки $P(A)$, применяя формулы полной вероятности и условной вероятности [5]. Рассмотрим следующую оценку:

$$P_n^*(A) = \begin{cases} \displaystyle\sum_{j=1}^{k} \frac{P_n(AH_j)}{P_n(H_j)} P(H_j), \quad \text{если } P_n(H_j) \neq 0,\ j = 1,...,k, \\ \displaystyle\sum_{j=1}^{k-s} \frac{P_n(AH_j)}{P_n(H_j)} \tilde{P}(H_j), \text{ если } P_n(H_j) \neq 0, j = 1,...,k-s, \quad 0 \le s \le k-2, \\ P_n(A), \text{ если } P(H_j) = 0,\ s = k-2. \end{cases} \tag{1}$$

Здесь $P_n(AH_j) = \frac{1}{n}\sum_{i=1}^{n} I_i(AH_j)$, $P_n(A)$, $P_n(H_j)$ – эмпирические вероятности (доли), построенные по исходным данным, s – число событий из H, для которых $P_n(H_j) = 0$, $0 \le s \le k-2$, $\tilde{P}_1, ..., \tilde{P}_{k-s}$ – пересчитанные (нормированные) вероят-

ности после исключения из полной группы событий тех H_j, для которых $P_n(H_j)=0$.

Покажем, что оценка (1) является несмещенной. Для $k=2$ математическое ожидание

$$EP_n^*(A)=E\left\{\frac{P_n(AH_1)}{P_n(H_1)}P(H_1)+\frac{P_n(AH_2)}{P_n(H_2)}P(H_2)\right\}=$$

$$=P(H_1)\sum_{i=0}^{n}E\left\{\frac{P_n(AH_1)}{nP_n(H_1)=i}\,|\,i\right\}P\{nP_n(H_1)=i\}+$$

$$+P(H_2)\sum_{i=0}^{n}E\left\{\frac{P_n(AH_2)}{nP_n(H_2)=i}\,|\,i\right\}P\{nP_n(H_2)=i\}=$$

$$=P(H_1)[P(A)(P\{P_n(H_1)=0\}+P\{P_n(H_1)=1\}+\sum_{i=1}^{n-1}P(A\,|\,H_1)P\{nP_n(H_1)=i\}]+$$

$$+P(H_2)[P(A)(P\{P_n(H_2)=0\}+P\{P_n(H_2)=1\}+\sum_{i=1}^{n-1}P(A\,|\,H_2)P\{nP_n(H_2)=i\}]=$$

$$=P(H_1)P(A)(P\{P_n(H_1)=0\}+P\{P_n(H_1)=1\})+P(AH_1)(1-(P\{P_n(H_1)=0\}+$$

$$+P\{P_n(H_1)=1\}))+P(H_2)P(A)(P\{P_n(H_2)=0\}+P\{P_n(H_2)=1\})+$$

$$+P(AH_2)(1-(P\{P_n(H_2)=0\}+P\{P_n(H_2)=1\}))=P(A). \qquad (2)$$

Равенство в (2) вытекает из того, что

$$P(H_1)+P(H_2)=1 \text{ и } P\{P_n(H_1)=0\}=P\{P_n(H_2)=1\}.$$

Выполняя подобные рассуждения для значений $k>2$ и учитывая (2), устанавливаем несмещенность оценки (1). Данная оценка имеет конечную дисперсию. В зависимости от типа выборки (повторная или бесповторная) дисперсии имеют разные выражения, в силу громоздкости выражений они здесь не приводятся.

3. Асимптотическая нормальность оценок

Рассмотрим асимптотические свойства оценки (1) для повторной выборки. Поскольку эмпирические вероятности с ростом объема выборки стремятся по вероятности к своим истинным значениям, асимптотические свойства оценка будут определяться поведением величины

$$Q_n^*=\sum_{j=1}^{k}\frac{P_n(AH_j)}{P_n(H_j)}P(H_j),\quad P_n(H_j)\neq 0,\ j=1,...,k. \qquad (3)$$

Разложим эту величину в окрестности истинных вероятностей по формуле Тейлора с остаточным членом R_n^* в форме Лагранжа. В результате имеем

$$Q_n^*=P(A)+\sum_{j=1}^{k}[P(AH_j)-P_n(AH_j))-P(A\,|\,H_j)(P(H_j)-P_n(H_j)]+R_n^*. \qquad (4)$$

Главная часть в (4) имеет математическое ожидание равное $P(A)$ и дисперсию σ_A^2/n, где

$$\sigma_A^2 = P(A)(1-P(A)-(\sum_{j=1}^{k} P^2(A \mid H_j)P(H_j) - P^2(A)). \tag{5}$$

На основании теоремы непрерывности (см. [6, гл. 6]) последовательность $\sqrt{n}\ R_n^*$ слабо сходится к нулю при $n \to \infty$. Отсюда, в силу центральной предельной теоремы имеем

$$\lim_{n\to\infty} P\{\sqrt{n}[P_n^*(A) - P(A)] < z\} = N(z,(0,\sigma_A^2)), z \in (-\infty,+\infty),$$

где $N(z,(0,\sigma_A^2))$ – нормальный закон распределения с нулевым математическим ожиданием и дисперсией (5).

Пусть исходная выборка объема $n = n_1 + n_2$ состоит из n_1 комплектных наблюдений и n_2 некомплектных, причем событие A наблюдается только в комплектных наблюдениях, а события H_j во всех наблюдениях и вероятности $P(H_j)$ неизвестны. В этом случае оценку для $P(A)$ возьмем в виде

$$\hat{P}_n(A) = \begin{cases} \sum_{j=1}^{k} P_{n_1}(A \mid H_j) P_n(H_j), \quad \text{если } P_{n_1}(H_j) \neq 0,\ j = 1,...,k, \\ \sum_{j=1}^{k-s} P_{n_1}(A \mid H_j)\tilde{P}_n(H_j), \text{ если } P_{n_1}(H_j) \neq 0, j = 1,...,k-s, \quad 0 \le s \le k-2, \\ P_{n_1}(A), \text{ если } P_{n_1}(H_j) = 0,\ s = k-2. \end{cases} \tag{6}$$

Для повторной выборки при $n \to \infty$ асимптотические свойства оценки (6) определяются величиной

$$\hat{Q}_n = \sum_{j=1}^{k} P_{n_1}(A \mid H_j) P_n(H_j), \quad P_{n_1}(H_j) \neq 0,\ \ j = 1,...,k.$$

Разложение этой величины в окрестности истинных вероятностей по формуле Тейлора с остаточным членом R_n в форме Лагранжа приводит к выражению

$$\hat{Q}_n = P(A) + \sum_{j=1}^{k} (P(AH_j) - P_{n_1}(AH_j)) - P(A \mid H_j) \times$$
$$\times [(P(H_j) - P_{n_1}(H_j)) - (P(H_j) - P_n(H_j))] + R_n.$$

В этом выражении главная часть имеет математическое ожидание равное $P(A)$ и дисперсию $\hat{\sigma}_A^2/n_1$, где

$$\hat{\sigma}_A^2 = [P(A)(1-P(A) - \frac{n_2}{n_1+n_2}(\sum_{j=1}^{k} P^2(A \mid H_j)P(H_j) - P^2(A))]. \tag{7}$$

Пусть соотношение между объемом выборки и объемом некомплектных наблюдений задается пропорцией $n_2 = tn$, $0 < t < 1$, которая соблюдается при увеличении n. Тогда отношение $n_2/(n_1+n_2)$ в пределе заменяется в (7) на t. В силу теоремы непрерывности (см. [6, гл. 6]) последовательность $\sqrt{n}R$ слабо сходится к

нулю при $n \to \infty$. Следовательно, на основании центральной предельной теоремы имеем

$$\lim_{n\to\infty} P\{\sqrt{n}[\hat{P}_n(A) - P(A)] < z\} = N(z, (0, \hat{\sigma}_A^2)), z \in (-\infty, +\infty).$$

Сравнение предельных дисперсий (5) и (7) асимптотически нормальных оценок (1) и (6) показывает, что замена вероятностей $P(H_j)$ на оценки $P_n(H_j)$, построенные по выборке, приводит к снижению точности, а величина уменьшения определяется коэффициентом пропорциональности t.

Заключение

Построены оценки для вероятностей событий с использованием априорной информации и с учетом информации, содержащейся в некомплектных наблюдениях категориальных данных в выборочных исследованиях. Установлена асимптотическая нормальность оценок, получены асимптотические дисперсии оценок, которые показываю, как влияет учет дополнительной информации на точность оценок (уменьшение их дисперсий). Эти результаты позволяют строить доверительные интервалы неизвестных вероятностей с меньшей шириной по сравнению с обычными эмпирическими оценками при тех же доверительных вероятностях. Полученные оценки могут применяться в оценивании различных функционалов от распределений (дисперсий оценок, условных распределений) методом подстановки.

ЛИТЕРАТУРА

1. *Ядов В.А.* Стратегия социологического исследования. М.: Омега-Л, 2007. 567 с.
2. *Котлер Ф.* Основы маркетинга: пер. с англ. М.: РосИнтер, 1996. 698 с.
3. *Литтл Дж.А.*, *Рубин Д.Б.* Статистический анализ данных с пропусками. М.: Финансы и статистика, 1991. 430 с.
4. *Чурилова А.А.* Корректировка неответов // Материалы семинара «Несплошные статистические исследования». Нижний Новгород, 2000. С. 27.
5. *Тарима С.С.* Использование дополнительной информации при оценке вероятностей и интерпретации натурного эксперимента: дис. ... канд. техн. наук. Томск: ТГУ, 2001. 149 с.
6. *Боровков А.А.* Математическая статистика. М: Наука, 2007. 704 с.

Дмитриев Юрий Глебович
Томский государственный университет
E-mail: dmit@mail.tsu.ru

Поступила в редакцию 3 июля 2012 г.

Dmitriev Yury G. (Tomsk State University). **On estimates of the probabilities with missing data.**

Keywords: complete and incomplete observations, additional information, statistical estimate, contingency table.

A problem of estimating of the probabilities with missing data is considered under assumption that certain additional information concerning probabilities is available. Different estimators of the probabilities 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.