УДК 519.253+519.712

Орехов А. В.

Стохастический подход к корректировке выборочных данных

Пусть в однократном статистическом наблюдении фиксируются $s+r$ признаков; поставим им в соответствие упорядоченный набор $(X_1,\ldots,X_s,Y_1,\ldots,Y_r)$, где признаки $X_1,\ldots,X_s$ являются контрольными (законы распределения таких признаков в генеральной совокупности считаются известными), а $Y_1,\ldots,Y_r$ — изучаемыми (их законы распределения в генеральной совокупности неизвестны) и все эти признаки обладают конечными спектрами.

Определим вектор $\mathbf{Z}$, компонентами которого являются индикаторы спектральных значений X_k и Y_l, где $1\leqslant k\leqslant s$ и $1\leqslant l\leqslant r$. Спектральным значениям контрольного признака X_k поставим в соответствие индикаторы $x_{1+\ldots+\alpha_{k-1}},\ldots,x_{\alpha_1+\ldots+\alpha_k}$, а спектральным значениям признака Y_l — индикаторы $y_{1+\ldots+\beta_{l-1}},\ldots,y_{\beta_1+\ldots+\beta_l}$. Пусть $\alpha_1+\ldots+\alpha_s=n$ и $\beta_1+\ldots+\beta_r=m$, тогда вектор $\mathbf{Z}$ можно записать в виде: $\mathbf{Z}=(x_1,x_2,\ldots,x_n,y_1,y_2,\ldots,y_m)$, где все его компоненты имеют распределение Бернулли.

Вектор $\mathbf{Z}$ принадлежит линейному пространству $\mathbb{R}^{n+m}$, в котором $e_1=(1,0,\ldots,0),\ldots,e_{n+m}=(0,0,\ldots,1)$ — единичные базисные векторы.

Представим $\mathbb{R}^{n+m}$ в виде прямой суммы линейных подпространств: $\mathbb{R}^{n+m}=\mathbb{X}\bigoplus\mathbb{Y}$, где подпространство $\mathbb{X}$ — линейная оболочка векторов $e_1,\ldots,e_n$, а дополнительное подпространство $\mathbb{Y}$ — линейная оболочка векторов $e_{n+1},\ldots,e_{n+m}$.

Обозначим через $\mathbf{P}=(p_1,\ldots,p_n,p_{n+1},\ldots,p_{n+m})$ — вектор вероятностей (генеральных долей) событий $x_i=1$ и $y_j=1$. Проекциями $\mathbf{P}$ на $\mathbb{X}$ и $\mathbb{Y}$ будут заданный вектор $\mathbf{P}^X=(p_1,\ldots,p_n)$ и неизвестный вектор $\mathbf{P}^Y=(p_{n+1},\ldots,p_{n+m})$.

Рассмотрим произвольную выборочную совокупность

$$\{\mathbf{Z}_t\}_{t=1}^N=\{\mathbf{Z}_1,\ldots,\mathbf{Z}_N\}.$$

Орехов Андрей Владимирович – старший преподаватель, Санкт-Петербургский государственный университет; e-mail: A_V_Orehov@mail.ru, тел.: +7(921)357-33-23

Обозначим через $\mathbf{W} = (w_1, \ldots, w_n, w_{n+1}, \ldots, w_{n+m})$ вектор, компоненты которого относительные частоты (выборочные доли) событий $x_i = 1$ и $y_j = 1$. Проекциями $\mathbf{W}$ на подпространства $\mathbb{X}$ и $\mathbb{Y}$ будут векторы $\mathbf{W}^X = (w_1, \ldots, w_n)$ и $\mathbf{W}^Y = (w_{n+1}, \ldots, w_{n+m})$, все компоненты которых являются известными.

Одна из основных задач выборочного метода состоит в приближении значений координат неизвестного вектора $\mathbf{P}^Y$ их статистическими оценками. Эти оценки являются компонентами случайного вектора $\mathbf{W}^Y$. Так как эмпирические распределения изучаемых признаков зависят от распределения контрольных признаков, то точность такого приближения тем выше, чем меньше координаты векторов $\mathbf{P}^X$ и $\mathbf{W}^X$ отличаются друг от друга.

Процедура приведения структуры выборки в соответствие со структурой генеральной совокупности называется «корректировкой выборки». Рассмотрим рандомизированный алгоритм корректировки выборочных данных. Итерации этого алгоритма можно описать следующим образом. Сначала выбирается спектральное значение одного из контрольных признаков такое, что модуль разности между соответствующими выборочной и генеральной долями имеет максимальное значение, т. е. по всем k таким, что $1 \leqslant k \leqslant n$, ищется $\varepsilon = \max |w_k - p_k|$. Если таких спектральных значений несколько, то случайным образом выбирается любое из них. Пусть это будет i-тая компонента векторов $\mathbf{W}^X$ и $\mathbf{P}^X$. Затем, если $w_i - p_i > 0$, из выборочной совокупности случайным образом удаляется некоторый вектор $\mathbf{Z_t}$, в котором компонента $x_i = 1$; если же $w_i - p_i < 0$, то в выборочной совокупности случайным образом дублируется вектор $\mathbf{Z_t}$, в котором компонента $x_i = 1$. Алгоритм завершается тогда, когда выполняется неравенство $\varepsilon \leqslant \delta$, где δ некоторое наперед заданное положительное число.

Обоснованием для применения этого алгоритма является тот факт, что при увеличении генеральной совокупности гипергеометрическое распределение все меньше и меньше отличается от биномиального, т. е. при стремлении объема генеральной совокупности к бесконечности нет существенного различия между повторными и бесповторными выборками. Поэтому, когда объем выборки намного меньше объема генеральной совокупности, расчет выборочных долей для бесповторной выборки будет мало отличаться от расчета выборочных долей для повторной выборки.