КОМПЬЮТЕРНАЯ
МАТЕМАТИКА

УДК 303.5

В.А. ПЕПЕЛЯЕВ, Н.А. ГОЛОДНИКОВА, Т.П. ЛЕВАШКО

МЕТОД ПОИСКА ОПТИМАЛЬНЫХ ПЛАНОВ ПРОВЕДЕНИЯ ВЫБОРОЧНОГО ОБСЛЕДОВАНИЯ

Рассматривается задача оптмизации планов проведения стратифицированных выборочных обследований для случая, когда одновременно изучаются два показателя и налагаются ограничения на погрешности оценок каждого из этих показателей как на уровне всей генеральной совокупности, так и на уровне отдельных страт. Предложен численный алгоритм для решения этой задачи. С помощью этого алгоритма найдены численные решения задачи оптимизации выборочной совокупности при обследовании условий жизни домохозяйств по регионам Украины.

Введение. В последнее время целенаправленное исследование общественного мнения относительно тех или иных аспектов социально-экономического развития страны стало неотъемлемой частью экономико-политических процессов в Украине. Традиционно в нашей стране большая часть статистических данных собиралась в рамках всеобщей переписи населения. Всеобщая перепись является трудоемким процессом, требующим больших затрат времени и финансовых ресурсов. В то же время выборочные методы обследований можно проводить в сжатые сроки при значительно меньших финансовых затратах и с обеспечением требуемой точности оценок. Такие результаты достигаются путем оптимизации выборочной совокупности.

В случае, когда значения показателя, который изучается при выборочном обследовании, неоднородно распределено по всей генеральной совокупности, часто применяют стратифицированный отбор. При стратифицированном отборе вся генеральная совокупность делится на меньшие подсовокупности (страты), каждая из которых внутренне однородна, что приводит к уменьшению дисперсии оценки в каждой страте. Тогда стратифицированный отбор определенного количества элементов из каждой страты будет репрезентативным для всей совокупности в целом, и может дать выигрыш в точности при оценивании характеристик генеральной совокупности. Дальнейший выигрыш в точности при оценивании характеристик можно достичь оптимизируя объемы отборов элементов из каждой страты.

Задача оптимизации планов проведения стратифицированных выборочных обследований рассматривалась в работах [1–5]. Формально эта задача состоит в поиске таких оптимальных планов проведения выборочных обследований в каждой страте, чтобы погрешность оценки одного показателя на уровне всей генеральной совокупности была минимальной при фиксированных затратах или были минимальны затраты при фиксированной точности оценки. В монографии А.И. Черняка [1] приведены аналитические выражения для оптимальных планов проведения выборочных обследований в каждой страте для случая линейных и некоторых видов нелинейных функций затрат.

Постановка задачи. Рассматривается генеральная совокупность всех домохозяйств Украины, расположенных в городских населенных пунктах. Эта совокупность состоит из N элементов, пронумерованных от 1 до N: $\{u_1,\ldots,u_k,\ldots,u_N\}$. Для простоты будем идентифицировать k-й элемент с его индексом, k. Обозначим генеральную совокупность следующим образом:

$$U = \{1,\ldots,k,\ldots,N\}. \tag{1}$$

Предполагается, что объем всей генеральной совокупности N известен. Эта генеральная совокупность поделена на H групп (страты) $U_1,\ldots,U_H$. В отдельную группу входят все домохозяйства, расположенные в городских населенных пунктах в пределах одного региона Украины. Общее количество домохозяйств N_h в группе U_h считается известным $(h=1,\ldots,H)$. Таким образом, $N = \sum_{h=1}^{H} N_h$.

Случайная выборка s_h формируется из елементов группы U_h соответственно с планом $p_h(\cdot)(h=1,\ldots,H)$. Вероятностные отборы из разных групп проводятся независимо один от другого. Предполагается, что для каждой группы U_h используется одинаковый план формирования выборки.

В результате формируется суммарная стратифицированная выборка $s = s_1 \cup s_2 \cup \cdots \cup s_H$. Такой план формирования виборки называется стратифицированным случайным отбором.

Рассматриваются два признака:

совокупные расходы домохозяйства за месяц (y);

совокупные среднедушовые расходы домохозяйства за месяц (x).

В результате выборочного обследования должны быть оценены (как на уровне всей генеральной совокупности, так и на уровне отдельных регионов) такие показатели:

а) средние совокупные расходы домохозяйства за месяц

б) часть домохозяйств, где совокупные среднедушовые расходы домохозяйства за месяц не превышают некоторую заданную величину L.

Если обозначить

$$z = \begin{cases} 1, \text{ если } x \le L, \\ 0 - \text{ в противном случае,} \end{cases}$$

то математическим выражением для показателя б) будет $E[z]$. Таким образом, во время выборочного обследования собирается информация относительно признаков y и z. Оцениваются характеристики $E[y]$ и $E[z]$.

Пусть y_k – значение переменной y для k-го элемента, z_k – значение переменной z для k-го элемента. Тогда среднее значение признака y в регионе U_h будет

$$E_{U_h}[y] = \bar{y}_{U_h} = \frac{1}{N_h} \sum_{U_h} y_k, \; h = 1, \ldots, H, \tag{2}$$

оценка этого значения

$$\hat{y}_{U_h} = \frac{1}{n_h} \sum_{s_h} y_k, \quad h = 1, \ldots, H, \tag{3}$$

где n_h – объем выборки s_h на уровне региона U_h.

Дисперсия этой оценки

$$V\left(\hat{y}_{U_h}\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S^2_{yU_h}, \quad h = 1, \ldots, H, \tag{4}$$

где

$$S^2_{yU_h} = \frac{1}{N_h - 1} \sum_{U_h} \left(y_k - \bar{y}_{U_h}\right), \; h = 1, \ldots, H \tag{5}$$

является дисперсией признака y в регионе U_h.

Среднее значение признака y по всей генеральной совокупности U (на уровне страны) равно

$$E_U[y] = \bar{y}_U = \frac{1}{N} \sum_U y_k = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{U_h} \tag{6}$$

оценка этого значения

$$\hat{y}_U = \sum_{h=1}^{H} \left[\frac{N_h}{N} \sum_{s_h} \frac{y_k}{n_h}\right] = \sum_{h=1}^{H} \left[\frac{N_h}{N} \hat{y}_{U_h}\right], \tag{7}$$

где N_h – количество домохозяйств в регионе U_h.

Дисперсия этой оценки

$$V\left(\hat{y}_U\right)=\sum_{h=1}^{H}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\left(\frac{N_h}{N}\right)^2 S^2_{yU_h}\ . \tag{8}$$

Аналогичные формулы можно выписать для признака z .
В качестве функции затрат на проведение выборочных обследований рассматривается следующая нелинейная функция:

$$C\left(n_1,\ldots,n_H\right)=\sum_{h=1}^{H}c_h n_h^{\alpha}\ , \tag{9}$$

где $\alpha>0$ и c_h – стоимость одного интервью в регионе U_h .

Требуется найти оптимальное распределение выборочной совокупности по регионам Украины, минимизирующее стоимость работ (9), и удовлетворяющее следующим условиям:

- ограничение на коэффициент вариации оценки характеристики $E[y]$ на государственном уровне

$$\frac{\sqrt{V(\hat{y}_U)}}{\hat{y}_U}=\frac{\sqrt{\sum_{h=1}^{H}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\left(\frac{N_h}{N}\right)^2 S^2_{yU_h}}}{\sum_{h=1}^{H}\left[\frac{N_h}{N}\hat{y}_{U_h}\right]}\le\frac{k_s}{100\%}; \tag{10}$$

- ограничение на коэффициент вариации оценки характеристики $E[z]$ на государственном уровне

$$\frac{\sqrt{V(\hat{z}_U)}}{\hat{z}_U}=\frac{\sqrt{\sum_{h=1}^{H}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\left(\frac{N_h}{N}\right)^2 S^2_{zU_h}}}{\sum_{h=1}^{H}\left[\frac{N_h}{N}\hat{z}_{U_h}\right]}\le\frac{k_s}{100\%}; \tag{11}$$

- ограничение на коэффициент вариации оценки характеристики $E[y]$ на региональном уровне

$$\frac{\sqrt{V\left(\hat{y}_{U_h}\right)}}{\hat{y}_{U_h}}=\sqrt{\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{S^2_{yU_h}}{\hat{y}^2_{U_h}}}\le\frac{k_{рег}}{100\%},\quad h=1,2,\ldots,H; \tag{12}$$

- ограничение на коэффициент вариации оценки характеристики $E[z]$ на региональном уровне

$$\frac{\sqrt{V\left(\hat{z}_{U_h}\right)}}{\hat{z}_{U_h}} = \sqrt{\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{S^2_{zU_h}}{\hat{z}^2_{U_h}}} \le \frac{k_{рег}}{100\%}, \quad h = 1, 2, \ldots, H. \tag{13}$$

Кроме того, объем выборки n_h не может превышать количество домохозяйств N_h в регионе U_h:

$$n_h \le N_h. \tag{14}$$

В этих выражениях $n_h, h = 1, \ldots, H,$ являются неизвестными переменными, которые необходимо найти; $N_h, h = 1, \ldots, H,$ являются известными значениями, а параметры $\hat{y}_{U_h}$, $\hat{z}_{U_h}$, $S^2_{yU_h}$ и $S^2_{zU_h}$, $h = 1, \ldots, H$, – неизвестными значениями, которые определяются по результатам предыдущего обследования условий жизни домохозяйств по регионам Украины.

Ограничения (10)–(13) – нелинейные относительно неизвестных переменных n_h, $h = 1, \ldots, H$. Чтобы сделать эти ограничения линейными, будем рассматривать переменные $v_h = \frac{1}{n_h}$, n_h, $h = 1, \ldots, H$. Тогда задачу оптимизации распределения выборочной совокупности по регионам Украины можно представить в следующем виде:

$$F\left(v_1, v_2, \ldots, v_H\right) = \sum_{h=1}^{H} c_h \left(\frac{1}{v_h}\right)^{\alpha} \to \min \tag{15}$$

при ограничениях

$$\frac{1}{N_h} \le v_h \le \frac{1}{N_h} + \left(\frac{k_{рег}}{100\%}\frac{\hat{y}_{U_h}}{S_{yU_h}}\right)^2, \quad h = 1, 2, \ldots, H, \tag{16}$$

$$\frac{1}{N_h} \le v_h \le \frac{1}{N_h} + \left(\frac{k_{рег}}{100\%}\frac{\hat{z}_{U_h}}{S_{zU_h}}\right)^2, \quad h = 1, 2, \ldots, H, \tag{17}$$

$$\sum_{h=1}^{H} \frac{N_h^2 S^2_{yU_h}}{N^2} v_n \le \sum_{h=1}^{H} \frac{N_h S^2_{yU_h}}{N^2} + \left(\frac{k_s}{100\%} \sum_{h=1}^{H} \left[\frac{N_h \hat{y}_{U_h}}{N}\right]\right)^2, \tag{18}$$

$$\sum_{h=1}^{H} \frac{N_h^2 S^2_{zU_h}}{N} v_n \le \sum_{h=1}^{H} \frac{N_h S^2_{zU_h}}{N} + N\left(\frac{k_s}{100\%} \sum_{h=1}^{H} \left[\frac{N_h \hat{z}_{U_h}}{N}\right]\right)^2. \tag{19}$$

С целью улучшения обусловленности задачи в правую и левую части ограничения (18) введен знаменатель N^2, а в ограничение (19) – знаменатель N. Ограничения (16) и (17) можно объединить в одно ограничение

$$\frac{1}{N_h} \le v_h \le \min\left\{ \frac{1}{N_h} + \left(\frac{k_{рег}}{100\%} \frac{\hat{y}_{U_h}}{S_{yU_h}} \right)^2, \frac{1}{N_h} + \left(\frac{k_{рег}}{100\%} \frac{\hat{z}_{U_h}}{S_{yU_h}} \right)^2 \right\},$$

$$h = 1, 2, ..., H. \qquad (20)$$

Функция цели при $v > 0$ является гладкой и имеет непрерывные первые и вторые частные производные. Ограничения (16)–(20) – линейные. Из ограничений (20) находим нижние границы для переменных n_h, $h = 1, \dots, H$:

$$\frac{1}{\min\left\{ \frac{1}{N_h} + \left(\frac{k_{рег}}{100\%} \frac{\hat{z}_{U_h}}{S_{zU_h}} \right)^2, \frac{1}{N_h} + \left(\frac{k_{рег}}{100\%} \frac{\hat{z}_{U_h}}{S_{zU_h}} \right)^2 \right\}},$$

$$h = 1, 2, ..., H. \qquad (21)$$

Алгоритм решения задачи. Для решения оптимизационной задачи (15), (18)–20) предлагается следующий алгоритм.

Шаг 1. В качестве начального приближения возьмем точку $v^0 = = (v_1^0, \dots, v_H^0)$, компоненты которой совпадают с нижними границами в ограничении (20). Легко проверить, что эта точка удовлетворяет также и ограничениям (18)–(19). Положим $k = 0$.

Шаг 2. В окрестности точки v^k функцию цели $F(v_1, \dots, v_H)$ аппроксимируем квадратичной функцией, раскладывая ее в ряд Тейлора и удерживая только линейные и квадратичные члены. Решаем задачу минимизации этой квадратичной функции:

$$\sum_{h=1}^{H} \left[\left\{ -c_h \alpha(\alpha+2) \left(v_h^k\right)^{-(\alpha+1)} \right\} v_h + \left\{ \frac{1}{2} c_h \alpha(\alpha+1) \left(v_h^k\right)^{-(\alpha+2)} \right\} v_h^2 \right] -> \min, \quad (22)$$

при линейных ограничениях (18) – (20). Обозначим решение этой задачи $\tilde{v}^k$ **.**

Шаг 3. Вектор $(\tilde{v}^k - v^k)$ определяет направление убывания функции $F(v_1, \dots, v_H)$. Если длина этого вектора близка к нулю, то точка v^k является решением задачи (15), (18) – (20). В этом случае переходим на шаг 5. В противном случае движемся из точки v^k в направлении вектора $(\tilde{v}^k - v^k)$ с шагом β: $0 < \beta \le 1$ и определяем новую точку $v(\beta)$ по формуле

$$v(\beta) = v^k + \beta(\tilde{v}^k - v^k). \qquad (23)$$

Шаг 4. Поскольку точки v^k и $\tilde{v}^k$ удовлетворяют ограничениям (18) – (20) и допустимое множество является выпуклым, то при любом значении β, таком что $0 < \beta \le 1$, точка (23) также удовлетворяет ограничениям (18) – (20). Выберем такое значение β из интервала (0,1], чтобы максимально уменьшить значение функции $F(v)$, т. е. решаем следующую задачу одномерной оптимизации:

$$\tilde{F}(\alpha) = F\left(v^k + \beta(\tilde{v}^k - v^k)\right) \to \min_{\beta} \tag{24}$$

при ограничениях

$$0 < \beta \le 1 . \tag{25}$$

Пусть β_k – решение задачи (24)–(25). Определяем следующее приближение:

$$v^{k+1} = v^k + \beta_k(\tilde{v}^k - v^k) . \tag{26}$$

Увеличиваем k на единицу ($k := k+1$) и переходим на шаг 2.

Шаг 5. Вычисляем оптимальные значения n_h, $h = 1,\ldots,H$ по формуле

$$n_h = \left[\frac{1}{v_h}\right],\ h = 1,\ldots,H , \tag{27}$$

где $[\cdot]$ – целая часть.

Результаты расчетов. Исходные данные для решения задачи (15), (18)–(20) были предоставлены автором монографии [6]. Расчеты проводились при разных значениях параметров k_s, $k_{рег}$ и α. Результаты расчетов при α = 2, $k_{рег}$ = =12%, k_s = 1%, c_h =1, $h = 1,\ldots,H$ приведены в таблице.

ТАБЛИЦА

Регионы (городские населенные пункты), область	Оптимальное распределение выборки				
	при следующих ограничениях на точность оценивания:			согласно формуле (28) для показателей:	
	(20)	(18)–(20)	(18)– (19)	*y*	*z*
1	2	3	4	5	6
АР Крым	199	1423	1425	348	1425
Винницкая	163	748	749	465	749
Волынская	11	316	316	147	316
Днепропетровская	394	3339	3343	1631	3343
Донецкая	277	3745	3750	1582	3750
Житомирская	246	943	944	717	944
Закарпатская	60	454	454	200	454
Запорожская	189	1572	1574	502	1574
Ивано-Франковская.	445	783	784	347	784
Киевская	170	1377	1378	588	1378

Окончание таблицы

1	2	3	4	5	6
Кировоградская	83	755	756	329	756
Луганская	81	1961	1963	811	1963
Львовская	109	1374	1375	799	1375
Николаевская	753	1294	1295	722	1295
Одесская	449	2316	2318	910	2318
Полтавская	150	1106	1108	402	1108
Ровенская	86	597	598	222	598
Сумская	112	969	970	290	970
Тернопольская	92	601	602	244	602
Харьковская	379	2679	2682	959	2682
Херсонская	472	1428	1430	620	1430
Хмельницкая	277	1101	1103	380	1103
Черкасская	291	1044	1045	392	1045
Черновецкая	179	472	472	156	472
Черниговская	239	1017	1019	387	1019
г. Киев	622	2278	2281	2371	2281
г. Севастополь	1041	1041	346	343	346
Функция цели	**3557425**	**69203923**	**68402630**	**17417880**	**68402630**
Коэффициент вариации для показателя *y*	**0.02003**	**0.00802**	**0.00804**	**0.01**	**0.00804**
Коэффициент вариации для показателя *z*	**0.02996**	**0.01**	**0.01**	**0.01529**	**0.01**

Во втором столбце приведено распределение выборки по регионам, полученное по формуле (21). Это распределение является оптимальным для случая, когда учитываются только ограничения на уровне регионов (20) и игнорируются ограничения на уровне всей страны (18)–(19). Значения, приведенные в этом столбце, являются нижними границами для оптимального решения задачи (15), (18)–20). В третьем столбце приведено оптимальное решение задачи (15), (18)–(20).

Четвертый столбец содержит оптимальное решение задачи (15), (18)–(19), не учитывающей ограничения на уровне регионов. Этот вариант расчетов сравнивается с двумя аналитическими решениями рассматриваемой задачи, приведенными в пятом и шестом столбцах. В пятом столбце приведено оптимальное распределение выборки по регионам при фиксированной точности оценки показателя y. Это распределение определяется по формуле [1]

$$n_h = \frac{\left(\dfrac{N_h^2 S_{yU_h}^2}{N^2 c_h}\right)^{\frac{1}{1+\alpha}} \cdot \sum_{h=1}^{H}\left\{ c_h^{\frac{1}{1+\alpha}} \left(\dfrac{N_h^2 S_{yU_h}^2}{N^2}\right)^{\frac{\alpha}{1+\alpha}} \right\}}{\sum_{h=1}^{H} \dfrac{N_h S_{yU_h}^2}{N^2} + \left(\dfrac{k_s}{100\%} \sum_{h=1}^{H}\left[\dfrac{N_h \hat{y}_{U_h}}{N}\right]\right)^2}. \tag{28}$$

Соответственно, шестой столбец содержит оптимальное решение для случая, когда фиксируется точность оценки показателя $\mathbf{z}$. Это решение вычисляется по формуле, аналогичной (28).

В последних двух строках таблицы приведены значения коэффициентов вариации для показателей y и z на уровне страны. Поскольку во втором столбце значения этих коэффициентов для показателей y и z больше 0.01 (k_s =1%), то соответствующее распределение выборки по регионам, полученное по формуле (21), не является допустимым, так как нарушаются ограничения (18)–(19).

Оптимальному решению задачи (15), (18)–(20), представленному в третьем столбце, соответствуют значения коэффициентов вариации для показателей y и z на уровне страны, равные 0.00802 и 0.01соответственно. Поскольку значение коэффициентов вариации для показателей y меньше верхней границы (k_s =1%), то можно сделать вывод, что в точке оптимального решения задачи (15), (18)–(20) ограничение (18) не является активным, и удаление этого ограничения не приведет к изменению решения задачи.

Аналогичное замечание относится и к решению, представленному в четвертом столбце. Поскольку и в этом случае ограничение (18) не является активным, то это решение является также оптимальным решением задачи с одним ограничением на коэффициент вариации для показателя z на уровне страны. Вот почему это решение полностью совпадает с решением, полученным по аналитической формуле (28) и представленным в шестом столбце Заметим также, что это решение не является допустимым для задачи (15), (18)–(20), поскольку, соответствующий ему объем выборки для г. Севастополя меньше, чем объем выборки для этого же региона, представленного во втором столбце таблицы, который является нижней границей для оптимального решения.

Особо следует остановиться на решении, представленном в пятом столбце. Этому решению соответствует наименьшее значение функции цели из всех значений, приведенных в третьей строке снизу. Тем не менее, это решение не является оптимальным решением задачи (15), (18)–(19), поскольку для него коэффициент вариации для показателя z на уровне страны превышает предельно допустимое значение k_s =1% и, следовательно, это решение нарушает ограничение (19).

Заключение. В данной работе рассматривалась задача оптимизации планов проведения стратифицированных выборочных обследований для случая, когда одновременно изучаются два показателя и налагаются ограничения на погрешности оценок каждого из этих показателей как на уровне всей генеральной совокупности, так и на уровне отдельных страт. Предложен численный алгоритм для решения этой задач. С помощью данного алгоритма найдены численные решения задачи оптимизации выборочной совокупности при обследовании условий жизни домохозяйств по регионам Украины. Полученные численные результаты сравнивались с результатами, полученными по аналитической формуле. Сравнение показало, что в частном случае рассматриваемой задачи, когда применима аналитическая формула, численные результаты совпали с аналитическими.

В.А. Пепеляєв, Н.О. Голоднікова, Т.П. Левашко

МЕТОД ПОШУКУ ОПТИМАЛЬНИХ ПЛАНІВ ПРОВЕДЕННЯ
ВИБІРКОВОГО ОБСТЕЖЕННЯ

Розглядається задача оптимізації планів проведення стратифікованих вибіркових обсте-жень для випадку, коли одночасно вивчаються два показники і обмежуються похибки оцінок кожного з цих показників як на рівні всієї генеральної сукупності, так і на рівні окремих страт. Запропоновано алгоритм для розв’язання цієї задачі.

V.A. Pepelyaev, N.A. Golodnikova, T.P. Levashko

TECHNIQUE FOR SEARCH OF OPTIMAL SAMPLE ALLOCATION

The paper considers the problem of optimization of stratified sample allocation in the case with two variables studied and two restrictions on the coefficients of variation of estimators of each of these variables. We have proposed an algorithm for solving such a problem.

1. *Черняк О.І.* Техніка вибіркових досліджень. – К.: МІВВЦ. 2001. – 248 с.
2. *Särndal C.-E., Swensson B., Wretman J.* Model Assisted Survey Sampling. – New York: Springer-Verlag, 1992. – 532 p.
3. *Cochran W.G.* Sampling technique, 3 ed. – New York: John Wiley &Sons, 1977. – 411 p.
4. *Hansen M.M., Hurwitz W.N., Madow W.G.* Sampling survey. Methods and Theory. – New York: John Wiley & Sons, 1953. – 970 p.
5. *Kish L.* Survey sampling. 2 ed. – New York: John Wiley &Sons, 1976. – 642 p.
6. *Гладун О.М.* Вибіркові обстеження населення: методологія, методика, практика. – Ніжин. «Аспект-Поліграф», 2008. – С. 344.

Получено 16.12.2009

Об авторах:

Пепеляев Владимир Анатольевич,
доктор физико-математических наук, заведущий отделом
Института кибернетики имени В.М. Глушкова НАН Украины,

Голодникова Нина Александровна,
аспирантка Института кибернетики имени В.М. Глушкова НАН Украины,

Левашко Татьяна Петровна,
ведущий математик Института кибернетики имени В.М. Глушкова НАН Украины.