Институт механики Академии Наук Союза ССР
Прикладная математика и механика. Том XVI, 1952

РАСЧЕТ ОБЪЕМА ВЫБОРКИ ПРИ ОПРЕДЕЛЕНИИ СРЕДНИХ ЗНАЧЕНИЙ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК РАСЧЕТНЫМ ПУТЕМ

Е. Н. Бухман

(Москва)

Обычно рассматриваемая теорией выборочного метода задача обоснования объема выборки ставится следующим образом. Определяется среднее значение признака или частость (удельный вес) осуществления некоторого события на основе выборочного наблюдения. Требуется определить наименьшее число наблюдений с тем, чтобы определяемая характеристика с достаточно высокой вероятностью обладала ошибкой репрезентативности не свыше определенного, заранее заданного размера. Размер этот задается или в абсолютных числах, или чаще в процентах к определяемой характеристике (соответственно средней или частости). Наиболее общее решение данной задачи дает теорема Ляпунова, на основе которой с сколь угодно большой вероятностью ошибка выборочной характеристики не превзойдет размера, определяемого при неограниченно большом объеме генеральной совокупности как:

$$\Delta = \alpha\sigma = \frac{\alpha\sigma_1}{\sqrt{n}} \qquad \text{(абсолютная ошибка)} \tag{1}$$

$$\beta = \alpha v = \alpha \frac{V_1}{\sqrt{n}} \qquad \text{(относительная ошибка)} \tag{2}$$

Здесь через σ_1 и v_1 обозначены среднее квадратическое отклонение и коэффициент вариации при рассмотрении индивидуальных значений варьирующего признака, через σ и v — соответствующие характеристики для средних значений, основанных на n наблюдениях каждое, через α обозначен параметр, определяющий через интеграл вероятностей

$$P_\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{+\alpha} e^{-1/2 x^2} dx$$

степень надежности формул (1) и (2).

При определении частости величина σ_1 представляет собой $\sqrt{p(1-p)}$, где p — вероятность события, частость которого определяется.

Из формул (1) и (2) легко определяется и минимально необходимое число наблюдений n.

При этом предполагается, что определяемая характеристика наблюдается непосредственно n раз в порядке выборочного учета.

В этой работе рассматривается особый случай обоснования ошибки при выборочном обследовании, когда определяемая характеристика непосредственно не наблюдается, а вычисляется косвенным путем через ряд иных, непосредственно наблюдаемых показателей. При этих условиях надо определить, какой объем наблюдений необходим по каждому из непосредственно наблюдаемых показателей, с тем, чтобы результат с заданной вероятностью обладал ошибкой, не превышающей допускаемых размеров.

В этом случае задача представляется на первый взгляд неопределенной, так как при изменении числа наблюдений по одному из показателей должно измениться необходимое число наблюдений и по всем остальным с тем, чтобы общий результат исчисления обладал заданной точностью (если только при этом ни по одному из остальных показателей это число не является слишком малым).

Однако добавление надлежащим образом обоснованных дополнительных условий делает задачу вполне определенной.

Ошибка расчетной характеристики определяется величиной ее вариации σ_0. Эта ошибка с вероятностью P_α, определяемой интегралом вероятностей, не превышает величины

$$\Delta = \alpha\sigma_0 \tag{3}$$

Величина σ_0 определенным образом зависит от вариации отдельных, непосредственно наблюдаемых исходных показателей.

Если определению подлежит показатель y_1, определяемый расчетным путем через независимые между собою случайные переменные x_1, x_2, x_3, . . . , т. е. $y = f(x_1, x_2, x_3, \ldots)$, то характеристика вариации показателя y_1, лежащая в основе определения объема наблюдений, выражается как

$$\sigma_{y_1}^2 = \left(\frac{\partial f}{\partial x_1}\right)_0^2 \sigma_1^2 + \left(\frac{\partial f}{\partial x_2}\right)_0^2 \sigma_2^2 + \left(\frac{\partial f}{\partial x_3}\right)_0^2 \sigma_3^2 + \cdots \tag{4}$$

Значки 0 указывают, что для исходных переменных x_1, x_2, x_3, . . . принимаются их математические ожидания $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, . . . , отвечающие понятию генеральных средних в процессе выборки.

Применимость (4) обусловлена возможностью пренебрегать в разложении $f(x_1, x_2, x_3, \ldots)$ по строке Тейлора членами, содержащими разности $(x_1 - \bar{x}_1)$, $(x_2 - \bar{x}_2)$, $(x_3 - \bar{x}_3)$, . . . в степени выше первой, так как эти разности выражают отклонения выборочных средних от их математических ожиданий и при достаточно большом объеме выборки стремятся к нулю.

В том случае, если y_1 является алгебраической функцией вида $y_1 = Cx_1^{k_1} x_2^{k_2} x_3^{k_3} \ldots$, применение равенства (4) приводит к очень простому соотношению между коэффициентами вариации всех переменных вида

$$v_{y_1}^2 = \sum k^2 v^2$$

Это определяется путем достаточно элементарных выкладок.

В более сложном случае, когда функция y является суммой независимых случайных переменных, подобных y_1, т. е. когда

$$y = y_1 + y_2 + y_3 + \dots$$

характеристика вариации σ_y определяется условием

$$\sigma_y^2 = \sigma_{y_1}^2 + \sigma_{y_2}^2 + \sigma_{y_3}^3 + \dots$$

Отсюда имеем соотношение между коэффициентами вариации тех же переменных (поскольку $\sigma = v\bar{x}$)

$$v_y^2 \bar{y}^2 = v_{y_1}^2 \bar{y}_1^2 + v_{y_2}^2 \bar{y}_2^2 + v_{y_3}^2 \bar{y}_3^2 + \dots$$

или

$$v_y^2 = \sum_{i=1}^{\omega} \theta_i^2 v_{y_i}^2 \qquad \left(\theta_i = \frac{\bar{y}_i}{\bar{y}}\right) \tag{5}$$

Здесь θ_i выражает удельный вес математических ожиданий $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$, ... в общей их сумме $\bar{y}$.

Через переменные x_1, x_2, x_3, ... в силу (4) v_y^2 выразится так:

$$v_y^2 = \sum_{i=1}^{\omega} \theta_i^2 k_i^2 v_i^2 \tag{6}$$

Здесь знак суммы объединяет все ω переменных x_i, входящих в выражения всех частных функций y_1, y_2, y_3, ... При этом для всех переменных, входящих в состав y_i, значение θ_i остается неизменным и изменяется только при переходе к y_{i+1}.

Коэффициенты вариации v_i характеризуют вариацию средних значений соответствующих переменных, основанных на n_i наблюдениях. Вариация индивидуальных значений тех же переменных V_i в $\sqrt{n_i}$ раз выше, чем v_i, если считать, что выборка производится из совокупностей неограниченного объема. Поэтому через коэффициенты V_i выражение (6) представится в виде

$$v_y^2 = \sum_{i=1}^{\omega} \frac{\theta_i^2 k_i^2 V_i^2}{n_i} \tag{7}$$

Относительная ошибка β_y переменной y в выборочном наблюдении в соответствии с (3) будет

$$\beta_y = \alpha v_y \tag{8}$$

Эта величина ошибки не будет превзойдена с вероятностью, определяемой интегралом вероятностей при соответствующем выборе α. В частности, при $\alpha = 2.6$ эта вероятность равна 0.99; при $\alpha = 3$ она равна 0.997.

Вводя в (7) величину относительной ошибки β_y, имеем, используя (8), зависимость этой ошибки от числа наблюдений по всем переменным x_i:

$$\left(\frac{\beta_y}{\alpha}\right)^2 = \sum_{i=1}^{\omega} \frac{\theta_i^2 k_i^2 V_i^2}{n_i} \tag{9}$$

Это выражение позволяет однозначно определить наименьшее допустимое число наблюдений ν_h по каждому показателю, исходя из условия, что все остальные n_i стремятся к бесконечности и, следовательно, отвечающие им слагаемые (9) стремятся к нулю. Имеем

$$\nu_h = \left(\frac{\theta_h k_h V_h \alpha}{\beta_y}\right)^2 \tag{10}$$

Введя в (9) величины ν_h, являющиеся определенными постоянными числами, это условие представим в виде

$$\sum_{i=1}^{\omega} \frac{\nu_i}{n_i} = 1 \tag{11}$$

Условие (11) определяет значения n_i в виде одного уравнения с ω неизвестными. Задача нахождения оптимальных значений n_i становится определенной лишь при введении дополнительных условий.

Наиболее целесообразным из таких условий является условие, чтобы при заданном размере ошибки затраты труда на выполнение наблюдений были минимальны.

Если на одно наблюдение за переменной x_i требуется время τ_i, то общая сумма времени наблюдения есть

$$T = \sum_{i=1}^{\omega} n_i \tau_i \tag{12}$$

Требуется найти такие значения n_i, удовлетворяющие условию (11), при которых общая сумма времени наблюдений T оказывается наименьшей.

Применяя правило Лагранжа, отыскиваем минимум функции:

$$\Phi = \sum_{i=1}^{\omega} n_i \tau_i + \lambda\left(\sum_{i=1}^{\omega} \frac{\nu_i}{n_i} - 1\right)$$

Для этого приравниваем нулю частные производные:

$$\frac{\partial \Phi}{\partial n_h} = \tau_h - \lambda \frac{\nu_h}{n_h^2} = 0$$

Отсюда получаем

$$\frac{\nu_h^2}{n_h^2} = \frac{1}{\lambda} \tau_h \nu_h, \quad \text{или} \quad \frac{\nu_h}{n_h} = \frac{1}{\sqrt{\lambda}} \sqrt{\tau_h \nu_h} \tag{13}$$

Суммируя все эти выражения при изменении h от 1 до ω (переменную суммирования обозначаем, как обычно, через i) и используя условие (11), имеем

$$\sum_{i=1}^{\omega} \frac{\nu_i}{n_i} = 1 = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{\omega} \sqrt{\tau_i \nu_i}, \quad \text{или} \quad \sqrt{\lambda} = \sum_{i=1}^{\omega} \sqrt{\tau_i \nu_i}$$

После того из (13), подставляя значение $\sqrt{\lambda}$, имеем

$$\frac{\nu_h}{n_h} = \eta_h, \quad \text{или} \quad n_h \equiv \frac{\nu_h}{\eta_h} \tag{14}$$

где

$$\sqrt{\tau_h \nu_h}\left(\sum_{i=1}^{\omega} \sqrt{\tau_i \nu_i}\right)^{-1} = \eta_h \tag{15}$$

Формулы (10) (15), и (14) решают поставленную задачу. Формула (11) очень удобна в целях контроля правильности выполненных расчетов.

Рассмотрим вычисления на примере. В городской телефонии основной характеристикой для определения общей численности соединительных устройств определенного вида, необходимых для установления связи абонентов между собой, является нагрузка, определяемая как общее время занятия всех соединительных устройств

Таблица 1

	θ_h	V_h	τ_h	ν_h	$\nu_h \tau_h$	$\sqrt{\nu_h \tau_h}$	η_h	n_h
$c_и$ . . .	0.15	1.0	0.1	60.8	6.08	2.47	0.0294	2070
$t_и$ . . .	0.15	1.2	1.2	87.6	105.12	10.2	0.1220	720
$c_к$ . . .	0.10	0.8	0.1	17.3	1.73	1.31	0.0156	1110
$t_к$ . . .	0.10	1.2	1.2	38.9	46.68	6.84	0.0813	480
$c_с$ . . .	0.20	0.6	0.1	38.9	3.89	1.97	0.0234	1680
$t_с$ . . .	0.20	1.2	1.2	155.7	186.84	13.67	0.1626	960
$c_у$ . . .	0.55	1.1	0.1	990.0	99.00	9.95	0.1184	8360
$t_у$ . . .	0.55	1.2	1.2	1178.0	1413.60	37.60	0.4473	2640
	1.00					84.06	1.0000	

соответствующего вида со стороны всех абонентов, пользующихся данной группой этих устройств. Если среднее число занятий всех устройств составляет c занятий в час, причем каждое занятие длится в среднем t минут, то произведение ct представляет общее время занятия всех устройств. Но среднее число занятий за 1 час и среднее время занятия различны для абонентов различных категорий. Обычно разграничивают время занятий телефонов квартирных индивидуального пользования $c_и t_и$, квартирных коллективного пользования $c_к t_к$, соединительных линий к коммутаторам учреждений $c_с t_с$, наконец, телефонов учреждений $c_у t_у$. Общая нагрузка y представляется суммой

$$y = c_и t_и + c_к t_к + c_с t_с + c_у t_у$$

Непосредственно наблюдаются 8 величин c и t. На их основе определяется расчетным путем общая нагрузка y, непосредственно не наблюдаемая. Тем не менее именно для величины y ставится условие определить ее с ошибкой, не превышающей β_y.

Соответствующие вычисления приведены в табл. 1.

В этой таблице приведены исходные удельные веса отдельных групп абонентов (указанных выше) в общей нагрузке, т. е. $\theta = ct / y$; исходные коэффициенты вариации V_h и значения затрат труда на производство наблюдений τ_h в мин. Значения η_h и n_h вычислены соответственно по формулам (15) и (14).

Все значения k равны 1; параметр α принят в 2.6. Заметим, что по сравнению с общепринятой гарантией в 3σ сокращение этого параметра до 2.6 позволяет уменьшить объем выборки на 25% при сохранении достаточной гарантии надежности расчета. Наибольший размер допускаемой относительной ошибки β_y при вычисления был принят в 0.05 (т. е. 5%).

Вычисление минимального числа наблюдений ν_h можно провести при помощи зависимости (10) либо, если $k = 1$, по специальным номограммам (см., например, [1] статью, стр. 529 и книгу [2], стр. 159). При применении номограмм значения коэффициентов вариации V_h нужно принимать умноженными на θ_h.

Из приведенной таблицы легко видеть, что число наблюдений оказывается, естественно, более высоким для проще наблюдаемых показателей c, чем для трудоемких t. Вместе с тем число наблюдений тем выше, чем выше удельный вес данного показателя (или группы показателей) в общем итоге θ.

При изменении требований точности число наблюдений по всем показателям изменяется обратно пропорционально квадрату величины допускаемой ошибки (абсолютной или относительной — безразлично). В частности, если требование точности снизить вдвое (приняв, например, $\beta = 0.1$ вместо 0.05), то необходимое число наблюдений равномерно по всем показателям снижается вчетверо.

Поступила 12 VII 1951

ЛИТЕРАТУРА

1. Бухман Е. Н. Номограммы числа наблюдений, необходимого при определении выборочным путем средней или доли (частости). ПММ. 1939. Т. II. Вып. 4.
2. Бухман Е. Н., Подгородецкий П. А., Остроухов Л. И. Статистика связи. Связьиздат. 1950.