Институт механики Академии Наук Союза ССР
Прикладная математика и механика. Том XI, 1947

О СПОСОБАХ ИСПОЛЬЗОВАНИЯ ВЫБОРОЧНЫХ ОБСЛЕДОВАНИЙ[1]

В. М. Обухов
(1873—1945)

1. Имеем генеральную совокупность, состоящую из m единиц. Некоторое явление x зарегистрировано *сплошным*, но *дефектным* инструментарием, вносящим *систематическую* ошибку в результаты исследования.

Допустим, что в целях выявления и определения размеров этой ошибки организовано дополнительное обследование, основанное на случайной выборке n единиц из m элементов генеральной совокупности, причем n весьма невелико сравнительно с m. Это *выборочное* обследование основано на *точном*, устраняющем систематическую ошибку способе. Таким образом, имеется два ряда из n наблюдений: $x_1, x_2, \ldots, x_n$; $y_1, y_2, \ldots, y_n$.

Сводные признаки этих рядов будут соответственно: M_x и M_y — средние арифметические, σ_x и σ_y — средние квадратические отклонения, $E_x = \sigma_x / \sqrt{n}$ и $E_y = \sigma_y / \sqrt{n}$ — средние квадратические отклонения средних арифметических.

Обозначим символом x_0 *среднюю величину признака генеральной совокупности* ряда x по дефектному обследованию. В результате сплошного, но дефектного обследования и поправки, установленной по выборочному обследованию, можно определить размеры общей суммы признака генеральной совокупности с устранением систематической ошибки A_x. Получим

$$A_x = m(x_0 + M_y - M_x) \tag{1.1}$$

Но можно подойти к определению величины A иначе; отбросив дефектный, но массовый источник x, определить размеры генеральной совокупности исключительно по выборочному точному инструментарию. Получим

$$A_y = mM_y \tag{1.2}$$

Численность генеральной совокупности одна и та же m; несовпадение результатов A_x и A_y зависит от множителей $x_0 + M_y - M_x$ в (1.1) и M_y в (1.2).

Очевидно, предпочтение должно быть дано тому способу, который при одинаковом объеме выборки дает более точные результаты. Математически это выражается следующим образом: если

$$E(M_y - M_x) = \sqrt{E_x^2 + E_y^2 - 2E_yE_xr_{xy}} > E_y, \quad \text{или} \quad r_{xy} < \frac{1}{2}\frac{E_x}{E_y} \tag{1.3}$$

то способ поправок должен быть отвергнут; если

$$E(M_y - M_x) = \sqrt{E_x^2 + E_y^2 - 2E_yE_xr_{xy}} = E_y, \quad \text{или} \quad r_{xy} = \frac{1}{2}\frac{E_x}{E_y} \tag{1.4}$$

[1] Настоящая статья представляет краткое извлечение из большой работы под тем же наименованием, составленное самим В. М. Обуховым и переданное из архива его женой Е. М. Обуховой.

то оба способа равноценны; если

$$E(M_y - M_x) = \sqrt{E_x^2 + E_y^2 - 2E_y E_x r_{xy}} < E_y, \quad \text{или} \quad r_{xy} > \frac{1}{2}\frac{E_x}{E_y} \qquad (1.5)$$

то преимущество на стороне поправок.

Обыкновенно σ_x и σ_y, а следовательно, E_x и E_y практически очень мало отличаются друг от друга, так как ряды x и y являются наблюдениями над одними и теми признаками одних и тех же объектов. Поэтому рассмотрим сначала случай, когда $E_x = E_y$; тогда предыдущие формулы соответственно приобретают вид $r_{xy} < 0.5$, $r_{xy} = 0.5$, $r_{xy} > 0.5$.

Итак, преимущество того или иного способа зависит от того, в какой мере наблюдается параллелизм по выборочному обследованию между дефектным рядом x и рядом y. Если этот параллелизм слаб и выражается коэффициентом корреляции менее 0.5, то пользование способом поправок должно быть отвергнуто; при $r_{xy} = 0.5$ наступает перелом. При $r_{xy} > 0.5$ и дальнейшем увеличении параллелизма между рядами x и y преимущество все больше и больше переходит к способу поправок.

С помощью простых преобразований легко установить критерий

$$K = \frac{E(M_y - M_x)}{E_y} = \sqrt{2}\sqrt{1 - r_{xy}}$$

Если $K > 1$, то способ поправок должен быть отвергнут; если $K = 1$ — оба способа равноценны; если $K < 1$, то преимущество на стороне способа поправок. Когда $r > 0.5$, удобнее пользоваться обратной величиной критерия

$$\frac{1}{K} = \frac{E_y}{E(M_y - M_x)}$$

показывающей, насколько способ самостоятельного исчисления менее точен, а следовательно, хуже метода поправок.

Приведем значения $1/K$ для некоторых значений r_{xy}

$r_{xy} =$	0.50	0.60	0.65	0.70	0.75	0.80	0.85	0.875
$1/K^2 =$	1.000	1.250	1.428	1.677	2.000	2.500	3.333	4.000
$1/K =$	1.000	1.118	1.195	1.294	1.414	1.581	1.826	2.000

$r_{xy} =$	0.90	0.95	0.96	0.97	0.98	0.99	1.00
$1/K^2 =$	5.000	10.000	12.500	16.660	25.000	50.000	∞
$1/K =$	2.236	3.162	3.536	4.082	5.000	7.071	∞

Как видно из этой таблички, относительное преимущество способа поправок в сильной степени возрастает по мере увеличения коэффициента корреляции между x и y. Еще в большей степени увеличивается число наблюдений, необходимое при самостоятельном исчислении по выборочному инструментарию, для достижения той же степени точности, которая получается при методе поправок. Так, например, если $r_{xy} = 0.875$, то средняя квадратическая ошибка самостоятельного исчисления в два раза больше ошибки способа поправок. Чтобы добиться того же результата прямым исчислением по выборочному методу, мы должны были бы увеличить объем выборки не в два раза, а в четыре, так как мера точности пропорциональна корню квадратному из числа наблюдений. Значения $1/K^2$ и показывают,

во сколько раз должно быть увеличено число наблюдений, необходимое при самостоятельном использовании выборочного инструментария для получения тех же результатов, как по способу поправок.

2. Перейдем к более общему случаю, когда $\sigma_x \neq \sigma_y$. Обозначим $\sigma_x / \sigma_y = l$; тогда критерий

$$K = \frac{E(M_y - M)_x}{E_y} = \sqrt{l^2 + 1 - 2lr_{xy}} \tag{2.1}$$

Из этой формулы вытекают следствия:

1) при $\sigma_x < \sigma_y$ преимущество способа поправок наступает при $r_{xy} < 0.5$. Например, при $l = 0.8$ инверсия в пользу способа поправок наступает при $r = 0.4$; при $l = 0.6$ она наступает при $r = 0.3$;

2) если $\sigma_x > \sigma_y$, то инверсия запаздывает. Например, при $l = 1.2$ она наступает при $r = 0.6$; если $l = 1.4$, то перелом наступает лишь при $r = 0.7$;

3) в обоих случаях $\sigma_x > \sigma_y$ и $\sigma_x < \sigma_y$ преимущества способа поправок имеют свои предельные значения, наступающие при $r_{xy} = 1$. Эти предельные значения уменьшаются по мере увеличения неравенства между σ_x и σ_y. По обе стороны от $l = 1$ это происходит одинаково.

3. Изложенный выше способ — это способ наложения абсолютных поправок. Второй вариант предусматривает наложение относительных поправок, как они выявились из сопоставления результатов обоих источников по n выбранным единицам $x_0 M_y / M_x$.

Второй более интересный и совершенный способ — это использование n выборочных наблюдений для составления уравнения регрессии $y = a + bx$.

Подставляя в это уравнение среднюю величину x_0 генеральной совокупности сплошного обследования, определяем наиболее вероятное значение интересующего нас явления $y_0 = a + bx_0$.

Приближенно ошибка y_0 выражается формулой

$$\frac{\sigma_y}{\sqrt{n}} \sqrt{1 - r_{xy}^2} \tag{3.1}$$

Совершенно ясно, что коэффициент полезного действия этого способа поправок выражается формулой

$$\frac{1}{K'^2} = \frac{1}{1 - r_{xy}^2} \tag{3.2}$$

Приводим значения этого коэффициента для некоторых значений r_{xy}:

$r_{xy} =$	0	0.25	0.50	0.60	0.70	0.75	0.80	0.85	0.90	0.95	0.975
$1/K'^2 =$	1.0	1.07	1.33	1.56	1.96	2.20	2.78	3.60	5.26	10.26	20.24
$1/K^2 =$	—	—	1.00	1.25	1.67	2.00	2.50	3.33	5.00	10.0	20.0

В третьей строке для сравнения приведены значения коэффициентов полезного действия $1 / K^2$ способа наложения абсолютных поправок.

Сравнивая результаты поправок, мы видим преимущества способа использования уравнения регрессии:

во-первых, коэффициент полезного действия в этом случае при всех значениях r больше единицы,

во-вторых, при одинаковых r коэффициент полезного действия способа использования уравнения регрессии выше коэффициента полезного действия способа абсолютных поправок (хотя по мере увеличения r эта разница уменьшается),

в-третьих, он не зависит от соотношения σ_y и σ_x (в этом его большое преимущество).

Формулы поправок могут быть написаны в другом виде:

первый способ наложения абсолютных поправок

$$y = x_0 + M_y - M_x = M_y - (M_x - x_0) \tag{3.3}$$

второй способ использования уравнения регрессии

$$y = M_y - \frac{\sigma_y}{\sigma_x} r_{xy} M_x + \frac{\sigma_y}{\sigma_x} r_{xy} x_0 = M_y - \frac{\sigma_y}{\sigma_x} r_{xy} (M_x - x_0) \tag{3.4}$$

Формула (3.3) показывает, что корректирование сплошного дефектного обследования превращается в корректирование результатов непосредственного использования n точных выборочных наблюдений M_y путем вычитания разности $M_x - x_0$; формула (3.4) показывает, что то же самое происходит и при втором способе корректирования, с той только разницей, что разность $M - x_0$ вычитается из M_y после умножения на множитель $r_{xy}\sigma_y / \sigma_x$.

Остановимся на элементе общности обоих способов, а именно $M_x - x_0$.

Если $M_x < x_0$, то это значит, что выборочная группа по x грешит нарушением репрезентативности в сторону увеличения, а так как ряды y и x коррелятивно связаны между собой, то вероятность, что тот же знак погрешности имеет и M_y, во всяком случае больше 0.5, поэтому для большего приближения M_y к действительности его следует уменьшить на некоторую величину: при первом способе M_y уменьшается на величину $M_x - x_0$, при втором способе — на ту же величину, но умноженную на $r_{xy}\sigma_y / \sigma_x$.

Обратно, если $M_x < x_0$, то нарушение репрезентативности происходит в сторону уменьшения; поэтому величину M_y следует увеличить в первом способе на $x_0 - M_x$, а во втором — на $(x_0 - M_x) r_{xy}\sigma_y / \sigma_x$.

Первый способ, исходя из наличия корреляционной связи, не учитывает степени тесноты этой связи и не использует соотношения σ_y и σ_x; второй способ учитывает тесноту связи, вводя множитель r_{xy}, а при помощи σ_y / σ_x переводит масштаб $M_x - x_0$ в масштаб измерений $M_y - y_0$. Это и обусловливает его бо́льшую точность.

Вариант первого способа – применение относительных поправок — обладает всеми недостатками способа абсолютных поправок. Это видно из преобразованной формулы для поправок

$$\frac{M_y}{M_x} x_0 = M_y - \frac{M_y}{M_x}(M_x - x_0)$$

V. M. OBUKHOV.—APPLICABILITY OF TEST FIGURES

Let a figure x be taken *in toto* but by defective instruments, introducing a constant error in the results. To determine this error, let an additional series of measurements be taken of n-test figures out of the total of m figures, the number n_1 being small as compared with m. These test measurements are carried out by a precise procedure removing the error occasioned by the defective instruments.

The work investigates the method and the conditions by which those results may be attained which will most closely approximate reality, and discusses when the best results may be obtained by means of the defective instrument taking all figures, and when by means of the exact measurement of test figures.