АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ

№ 1 1972

УДК 621.317 + 519.21

ОБ ОЦЕНИВАНИИ ИЗМЕРЯЕМОЙ ВЕЛИЧИНЫ ПО ДВУМ ГРУППАМ НАБЛЮДЕНИЙ

В практике часто возникает задача нахождения оценки неизвестного общего среднего по данным нескольких групп наблюдений. Рассмотрим некоторые вопросы, связанные с этой задачей.

Пусть $x_{1i}\,(i=1,2,\ldots,\ n_1)$ и $x_{2j}\,(j=1,2,\ldots,\ n_2)$ — две группы независимых наблюдений, полученных в результате измерения одной и той же величины μ разными методами. Следовательно,

$$M[x_{1i}]=\mu;\quad i=1,2,\ldots,\ n_1;$$
$$M[x_{2j}]=\mu;\quad j=1,2,\ldots,\ n_2;$$
$$D[x_{1i}]=\sigma_1^2;\quad D[x_{2j}]=\sigma_2^2.$$

Обычно в качестве оценки для μ используют взвешенное среднее

$$\bar{\bar{x}}=\lambda_{01}\bar{x}_1+\lambda_{02}\bar{x}_2, \tag{1}$$

где $\bar{x}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}x_{1i}$; $\bar{x}_2=\frac{1}{n_2}\sum_{j=1}^{n_2}x_{2j}$; $\lambda_{0,1}$, $\lambda_{0,2}$ — веса первой и второй групп наблюдений. Если σ_1^2 и σ_2^2 известны, то веса определяются по формулам:

$$\lambda_{01}=\frac{\frac{n_1}{\sigma_1^2}}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}};\quad \lambda_{02}=\frac{\frac{n_2}{\sigma_2^2}}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}}$$

и наилучшая линейная оценка для μ имеет вид

$$\bar{\bar{x}}_0=\frac{\frac{n_1}{\sigma_1^2}\bar{x}_1+\frac{n_2}{\sigma_2^2}\bar{x}_2}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}}. \tag{2}$$

Но чаще всего в практических задачах сами σ_1^2 и σ_2^2 неизвестны, а вместо них используются стандартные оценки:

$$S_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}; \quad S_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1}.$$

Если в формулу (1) подставить следующие оценки весов:

$$\lambda_1 = \frac{\frac{n_1}{S_1^2}}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}}, \quad \lambda_2 = \frac{\frac{n_2}{S_2^2}}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}},$$

то оценка измеряемой величины μ будет иметь вид

$$\bar{\bar{x}} = \frac{\frac{n_1}{S_1^2}\bar{x}_1 + \frac{n_2}{S_2^2}\bar{x}_2}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}} = \lambda_1 \bar{x}_1 + \lambda_2 \bar{x}_2. \tag{3}$$

Как видно из формулы (3), оценка $\bar{\bar{x}}$ является нелинейной функцией от наблюдений и определение ее характеристик среднего $M[\bar{\bar{x}}]$ и дисперсии $D[\bar{\bar{x}}]$ затруднительно. Поэтому многие авторы [1, 2] рассматривают S_1^2 и S_2^2 как детерминированные (неслучайные) величины и, полагая $S_1^2 \approx \sigma_1^2$, $S_2^2 \approx \sigma_2^2$, приходят к формулам:

$$\begin{aligned} M[\bar{\bar{x}}] &= \mu; \\ D_1[\bar{\bar{x}}] &\approx \frac{1}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}}; \\ \sigma_{1\bar{\bar{x}}} &\approx \sqrt{\frac{1}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}}}. \end{aligned} \tag{4}$$

Воспользовавшись тем, что при нормальном законе распределения наблюдений в группах $\bar{x}_i$ и S_i^2 независимы, вычислим точные значения $M[\bar{\bar{x}}]$ и $D[\bar{\bar{x}}]$ и оценим погрешность δ, возникающую из-за пользования приближенной формулой (4). Очевидно, что значение $\sigma_{1\bar{\bar{x}}}$, вычисленное по формуле (4), будет меньше действительного значения, так как не учитывает рассеивания статистик S_1^2 и S_2^2 около σ_1^2 и σ_2^2 соответственно.

Таким образом, всюду предполагается теперь, что случайные величины x_{1i} независимы и распределены по нормальному закону $N(\mu, \sigma_1^2)$, а величины x_{2j} также независимы и распределены по нормальному закону $N(\mu, \sigma_2^2)$.

Оценка $\bar{\bar{x}}$ для μ, даваемая формулой (3), несмещенная. В самом деле,

$$M[\bar{\bar{x}}] = M\left\{M\left[\sum_{i=1}^{2} \lambda_i \bar{x}_i \middle| S_1^2,\ S_2^2\right]\right\} = M\left\{\sum_{i=1}^{2} \lambda_i M[\bar{x}_i | S_1^2,\ S_2^2]\right\} =$$

$$= M \sum_{i=1}^{2} \lambda_i M[\bar{x}_i] = \mu \sum_{i=1}^{2} \lambda_i = \mu.$$

Дисперсия оценки $\bar{\bar{x}}$ может быть представлена в виде

$$D[\bar{\bar{x}}] = M[\bar{\bar{x}} - \mu]^2 = M[\lambda_1(\bar{x}_1 - \mu) + \lambda_2(\bar{x}_2 - \mu)]^2 = [\lambda_1(\bar{x}_1 - \mu)]^2 + \\ + M[\lambda_2(\bar{x}_2 - \mu)]^2 + 2M\{\lambda_1\lambda_2(\bar{x}_1 - \mu)(\bar{x}_2 - \mu)\} = \\ = M[\lambda_1(\bar{x}_1 - \mu)]^2 + M[\lambda_2(\bar{x}_2 - \mu)]^2, \text{ так как}$$

$$M\{\lambda_1\lambda_2(\bar{x}_1 - \mu)(\bar{x}_2 - \mu)\} = M\{M[\lambda_1\lambda_2(\bar{x}_1 - \mu)(\bar{x}_2 - \mu)|S_1^2, S_2^2]\} = 0.$$

Учитывая еще независимость S_i^2 от $\bar{x}_i$, будем иметь

$$D[\bar{\bar{x}}] = M[\bar{\bar{x}} - \mu]^2 = m_2(\lambda_1)\frac{\sigma_1^2}{n_1} + m_2(\lambda_2)\frac{\sigma_2^2}{n_2}, \tag{5}$$

где

$$m_2(\lambda_i) = M[\lambda_i^2].$$

Следовательно, для нахождения $D[\bar{\bar{x}}]$ надо знать $m_2(\lambda_1)$ и $m_2(\lambda_2)$. Для этого найдем функцию плотности распределений случайных величин:

$$\lambda_1 = \frac{\frac{n_1}{S_1^2}}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}}; \quad \lambda_2 = \frac{\frac{n_2}{S_2^2}}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}}.$$

Зная плотность распределения случайной величины $\frac{S_i^2}{\sigma_i^2}(n_i - 1)$, можно найти плотность распределения $P_i(y)$ случайной величины $y_i = \frac{n_i}{S_i^2}$:

$$P_i(y) = \frac{A_i^{b_i}}{2^{b_i}\Gamma(b_i)}\left(\frac{1}{y}\right)^{b_i+1} e^{-\frac{A_i}{2y}},$$

где $A_i = \frac{n_i(n_i - 1)}{\sigma_i^2}$; $b_i = \frac{n_i - 1}{2}$; $i = 1, 2$; $\Gamma(b_i)$ — гамма-функция.

Случайные величины λ_1 и λ_2 функционально связаны со случайными величинами y_1 и y_2:

$$\lambda_1 = \frac{y_1}{y_1 + y_2}; \quad \lambda_2 = \frac{y_2}{y_1 + y_2}.$$

Воспользовавшись известными методами [3] нахождения функции плотности вероятности функционально преобразованных случайных величин, можно вычислить функцию плотности вероятности $r_1(\lambda)$ и $r_2(\lambda)$ случайных величин λ_1 и λ_2:

$$r_1(\lambda) = \frac{1}{B(b_1, b_2)}\left(\frac{A_2}{A_1}\right)^{b_2}(1 - \lambda)^{b_1 - 1}\lambda^{b_2 - 1}\left(1 + \lambda\frac{A_2 - A_1}{A_2}\right)^{-b_1 - b_2};$$

$$r_2(\lambda) = \frac{1}{B(b_1, b_2)}\left(\frac{A_1}{A_2}\right)^{b_1}(1 - \lambda)^{b_2 - 1}\lambda^{b_1 - 1}\left(1 + \lambda\frac{A_1 - A_2}{A_1}\right)^{-b_1 - b_2},$$

где $B(b_1, b_2)$ — бета-функция. Теперь можно вычислить $m_2(\lambda_1)$ и $m_2(\lambda_2)$:

$$m_2(\lambda_1) = \left(\frac{A_2}{A_1}\right)^{b_2}\frac{B(b_1, b_2 + 2)}{B(b_1, b_2)}F_1\left(b_1 + b_2, b_2 + 2, b_1 + b_2 + 2, \frac{A_1 - A_2}{A_1}\right);$$

$$m_2(\lambda_2) = \left(\frac{A_1}{A_2}\right)^{b_1}\frac{B(b_2, b_1 + 2)}{B(b_1, b_2)}F_2\left(b_1 + b_2, b_1 + 2, b_1 + b_2 + 2, \frac{A_2 - A_1}{A_2}\right), \tag{6}$$

где F_1 и F_2 — гипергеометрические функции. Из (5) и (6) выводим

$$D[\bar{x}] = \left(\frac{A_1}{A_2}\right)^{b_1} \frac{B(b_1, b_2+2)}{B(b_1, b_2)} F_1\left(b_1+b_2,\ b_1, b_1+b_2+2,\ 1-\frac{A_1}{A_2}\right)\frac{\sigma_1^2}{n_1} +$$

$$+\left(\frac{A_1}{A_2}\right)^{b_1} \frac{B(b_2, b_1+2)}{B(b_1, b_2)} F_2\left(b_1+b_2,\ b_1+2,\ b_1+b_2+2,\ 1-\frac{A_1}{A_2}\right)\frac{\sigma_2^2}{n_2}.$$

С учетом соотношений для гипергеометрических функций [4] $D[\bar{x}]$ можно представить в следующем виде:

$$D[\bar{x}] = \left[\frac{n_1(n_1-1)}{n_2[n_2-1]}k\right]^{\frac{n_1-1}{2}} \frac{\sigma_1^2}{n_1} \frac{1}{\left(\frac{n_2-3}{2}\right)!\left(\frac{n_1-3}{2}\right)!}\left\{(-1)^{\frac{n_2-1}{2}} \frac{d^{\frac{n_1+n_2}{2}-2}}{dz^{\frac{n_1+n_2}{2}-2}} \times\right.$$

$$\times\left[(1-z)^{\frac{n_2+1}{2}} \frac{d}{dz}\left[-\frac{\ln(1-z)}{z}\right]\right] + k(-1)^{\frac{n_2-5}{2}} \frac{n_1}{n_2} \frac{d^{\frac{n_1+n_2}{2}-2}}{dz^{\frac{n_1+n_2}{2}-2}} \times$$

$$\left.\times\left[(1-z)^{\frac{n_2-3}{2}} \frac{d}{dz}\left[-\frac{\ln(1-z)}{z}\right]\right]\right\}, \qquad (7)$$

где $k = \frac{\sigma_2^2}{\sigma_1^2}$; $\sigma_{\bar{x}} = \sqrt{D[\bar{x}]}$; $z = 1 - \frac{n_1(n_1-1)}{n_2(n_2-1)}k$.

Поскольку формула (7) пригодна для вычисления $D[\bar{x}]$ в случае нечетного числа наблюдений, значения $D[\bar{x}]$ при четных n можно найти, пользуясь интерполяцией. Расчеты по формуле (7) упрощаются, если число наблюдений в группах $n_1 = n_2$. Так,
при $n_1 = n_2 = 3$

$$D[\bar{x}] = \frac{\sigma_1^2}{n_1} 2k\left[\frac{k+1}{(1-k)^2} + 2\frac{k\ln k}{(1-k)^3}\right];$$

при $n_1 = n_2 = 5$

$$D[\bar{x}] = \frac{\sigma_1^2}{5} k^2\left[-17\frac{1}{(1-k)^2} - 42\frac{k}{(1-k)^3} - 24\frac{k^2}{(1-k)^4} - 36\frac{k\ln k}{(1-k)^3} -\right.$$

$$- 54\frac{k^2\ln k}{(1-k)^4} - 24\frac{k^3\ln k}{(1-k)^5} - 6\frac{\ln k}{(1-k)^2} - 24\frac{k}{(1-k)^4} - 6\frac{1}{(1-k)^3} +$$

$$\left.+\frac{1}{k(1-k)^2} - 18\frac{k\ln k}{(1-k)^4} - 24\frac{k^2\ln k}{(1-k)^5}\right];$$

при $n_1 = n_2 = 7$

$$D[\bar{x}] = \frac{\sigma_1^2}{7}\frac{k^3}{4}\left[-620\frac{1}{(1-k)^3} + 1920\frac{k}{(1-k)^4} + 2040\frac{k^2}{(1-k)^5} +\right.$$

$$+ 720\frac{k^3}{(1-k)^6} + 24\frac{1}{k(1-k)^2} + 240\frac{\ln k}{(1-k)^3} + 1440\frac{k\ln k}{(1-k)^4} + 2880\frac{k^2\ln k}{(1-k)^5} +$$

$$+ 2400\frac{k^3\ln k}{(1-k)^6} + 720\frac{k^4\ln k}{(1-k)^7} + 840\frac{k}{(1-k)^5} + 720\frac{k^2}{(1-k)^6} + 120\frac{1}{(1-k)^4} -$$

$$\left.- 20\frac{1}{k(1-k)^3} + 4\frac{1}{k^2(1-k)^2} + 480\frac{k\ln k}{(1-k)^5} + 1200\frac{k^2\ln k}{(1-k)^6} + 720\frac{k^3\ln k}{(1-k)^7}\right].$$

При $k \to 1$ неопределенность выражений для $D[\bar{x}]$ раскрывается с помощью правила Лопиталя.

Оценим теперь относительную погрешность δ, возникающую при использовании приближенной формулы (4) для $\sigma_{\bar{\bar{x}}}$ вместо точной формулы (7):

$$\delta = \frac{\sigma_{1\bar{\bar{x}}} - \sigma_{\bar{\bar{x}}}}{\sigma_{\bar{\bar{x}}}} \cdot 100\%.$$

Значения δ в зависимости от k и числа наблюдений в группах приведены в таблице

k	$n=3$	$n=5$	$n=7$	$n=9$	$n=11$	k	$n=3$	$n=5$	$n=7$	$n=9$	$n=11$
1	13,4	8,7	6,5	5,1	4,3	7,39	19,6	9,3	5,6	3,9	3,0
1,65	13,9	8,8	6,5	5,1	4,2	9,49	20,7	9,2	5,2	3,5	2,6
2,12	14,5	9,0	6,5	5,0	4,1	12,18	21,8	9,0	4,9	3,2	2,3
2,72	15,4	9,1	6,4	4,9	3,9	15,64	22,7	8,6	4,4	2,8	2,0
3,49	16,3	9,3	6,3	4,7	3,7	20,09	23,5	8,2	3,9	2,4	1,7
4,48	17,4	9,4	6,1	4,5	3,5	33,12	25,0	7,1	3,0	1,7	1,2
5,75	18,5	9,4	5,9	4,2	3,2	—	—	—	—	—	—

В [5] показано, что оценка неизвестного среднего μ, вычисленная по формуле (3), целесообразна, когда число наблюдений в каждой из групп превышает 9. При малом числе наблюдений в [6] предлагаются другие оценки измеряемой величины, отличные от $\bar{\bar{x}}$. Одна из них имеет вид

$$\bar{\bar{x}}_* = \frac{\frac{\sqrt{n_1}}{S_1}\bar{x}_1 + \frac{\sqrt{n_2}}{S_2}\bar{x}_2}{\frac{\sqrt{n_1}}{S_1} + \frac{\sqrt{n_2}}{S_2}}. \qquad (8)$$

Там же показано, что оценка $\bar{\bar{x}}_*$ эффективнее оценки $\bar{\bar{x}}$ при $n<10$ и k, близких к единице.

Рассмотрим поведение оценки $\bar{\bar{x}}$ при увеличении числа наблюдений в группах. Пусть $n_2/n_1=c$, тогда $n_2=n_1c$.

При $n_1\to\infty$ случайные величины λ_1 и λ_2 сходятся по вероятности соответственно к $\dfrac{\frac{n_1}{\sigma_1^2}}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}}$ и $\dfrac{\frac{n_2}{\sigma_2^2}}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}}$,

так как

$$\frac{\frac{n_1}{\sigma_1^2}}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}} - \frac{\frac{n_1}{S_1^2}}{\frac{n_1}{S_1^2}+\frac{n_2}{S_2^2}} = \frac{c\left(\sigma_2^2 S_1^2 - \sigma_1^2 S_2^2\right)}{\left(\sigma_2^2 + c\sigma_1^2\right)\left(S_2^2 + cS_1^2\right)} \to 0 \quad \text{при } n_1 \to \infty.$$

Распределения случайных величин $\bar{x}_1$ и $\bar{x}_2$ при $n_1\to\infty$ сходятся к нормальным распределениям $N_1\left(\mu, \frac{\sigma_1^2}{n_1}\right)$ и $N_2\left(\mu, \frac{\sigma_2^2}{n_2}\right)$.

Тогда по теореме о сходимости [7] функции распределения случайных величин $\lambda_1\underline{\underline{x}}_1$ и $\lambda_2\underline{\underline{x}}_2$ сходятся по вероятности соответственно к нормальным распределениям $N_1\left(\frac{k}{k+c}\mu, \frac{k^2}{(k+c)^2}\frac{\sigma_1^2}{n_1}\right)$ и $N_2\left(\frac{c}{k+c}\mu, \frac{c^2}{(k+c)^2}\frac{\sigma_2^2}{n_2}\right)$.

Распределение случайной величины x, согласно центральной предельной теореме, стремится к нормальному распределению

$$N\left(\mu, \frac{k^2}{(k+c)^2}\frac{\sigma_1^2}{n_2} + \frac{c^2}{(k+c)^2}\frac{\sigma_2^2}{n_2}\right).$$

ВЫВОДЫ

В статье исследованы относительные погрешности, возникающие при использовании приближенной формулы (4) вместо точной (7).

Составлена таблица значений относительной погрешности при числе наблюдений в группах $n_1=n_2=n=3\div11$ [см. (2)]. Погрешность достигает максимального значения при $n=3$ и больших k. При $n\geqslant5$ относительная погрешность δ практически незначительна и вполне можно использовать приближенную формулу (3) для нахождения $\sigma_{\underline{\underline{x}}}$.

В заключение выражаю глубокую благодарность А. М. Кагану за внимание к работе.

ЛИГЕРАТУРА

1. Н. С. Смирнов, В. В. Дунин—Барковский. Курс теории вероятностей и математической статистики. М., «Наука», 1965.
2. Е. Ф. Долинский. Погрешности измерений и обработка результатов измерений. М., «Машиностроение», 1967.
3. Б. В. Гнеденко. Курс теории вероятностей. М., Физматгиз, 1961.
4. Г. Бейтен и А. Эрдейн. Высшие трансцендентные функции. М., «Наука», 1965.
5. F. Graybill, R. B. Deal. Combining Unbiased Estimators. Biometrics, 1959, v. 15.
6. J. S. Mehta, J. Gurland. On Combining Unbiased Estimators of the Mean.— Trab. estatist y invest. oper, 1969, v. 20, № 2—3.
7. Г. Крамер. Математические методы статистики. М., Изд-во иностр. лит., 1948.

Поступила в редакцию
29 апреля 1971 г.