УДК 311.43

ИННОВАЦИОННАЯ МЕТОДИКА АНАЛИЗА ВЫБОРОЧНОГО НАБЛЮДЕНИЯ

Канд. техн. наук, доц. ШИЛО А. Ф.

Белорусский национальный технический университет

Выборочное наблюдение – несплошное наблюдение, при котором отбор производится в случайном порядке части единиц всей совокупности (генеральной) с целью последующего распространения полученных данных выборки на всю генеральную совокупность.

Расхождения между значениями показателей, полученных по выборке, и соответствующими показателями генеральной совокупности порождаются ошибками (случайными и систематическими) репрезентативности выборки. Величина случайной ошибки репрезентативности зависит от объема выборки, степени вариации изучаемого признака и принятого способа формирования выборки. Различают среднюю и предельную ошибки. Предельная ошибка рассчитывается по средней при заданной вероятности ее появления.

Теория математической статистики устанавливает, что только среднее арифметическое $\overline{x}$, полученное по выборке, является несмещенной, эффективной и состоятельной характеристикой генеральной средней $\overline{x}_г$, т. е. $\overline{x} = \overline{x}_г$. Поэтому с учетом предельной ошибки Δ доверительный интервал, в котором следует ожидать генеральную среднюю при заданной доверительной вероятности γ, определяется неравенством

$$\overline{x} - \Delta \le \overline{x}_г \le x + \Delta.$$

Исходя из него в некоторых источниках по статистике применяется так называемый способ прямого пересчета, по которому другие показатели генеральной совокупности рассчитываются путем его почленного умножения на объем генеральной совокупности N по формуле

$$(\overline{x} - \Delta) N \le \overline{x}N \le (\overline{x} + \Delta) N.$$

При этом наблюдается различное толкование величины $\overline{x}N$: в одних – среднее генеральной совокупности, в других – суммарное количество признака и т. д. Так, в [1] при анализе выборки двадцати предприятий на предмет производительности труда по результатам расчетов $\overline{x} = 21{,}51$ и $\Delta = 3{,}62$ (как ниже будет показано, Δ определено с грубыми ошибками) приводится неравенство $17{,}88 \le \overline{x} \le 25{,}14$. Далее делается заключение: «Пределы, в которых можно ожидать суммарную производительность труда, для ста предприятий составят $17{,}88 \cdot 100 \le \overline{x}N \le 25{,}14 \cdot 100$; $1788 \le \overline{x}N \le 2514$».

Возникает вопрос: какую суммарную производительность отражает последнее неравенство – среднюю или общую? Очевидно, применяемый способ прямого пересчета не является бесспорным, так как не задействован такой важный показатель, как дисперсия; к тому же никак не аргументирован. Как следует из приведенного примера, значимость результатов практически равна нулю.

Между тем очень важно грамотно оценить пределы изменения не только средней генеральной совокупности $\overline{x}_г$, но и всех ее возможных значений $x_г$ (размах признака) при наличии или отсутствии гипотетического объема генеральной совокупности N. Такого рода исследований не обнаруживается.

Автором делается попытка восполнения указанного пробела. В результате математических исследований, подтвержденных расчетами на примерах, получены следующие результаты.

1. Интервал возможных значений признака генеральной совокупности по выборочной

средней $\overline{x}$ и предельной ошибке Δ оценивается правилом трех дельта

$$\overline{x} - 3\Delta \le x_{\text{г}} \le \overline{x} + 3\Delta. \qquad (1)$$

Если распределение симметричное или близко к таковому, то интервалом (1) будут охвачены все возможные значения генеральной совокупности.

При значительной левосторонней асимметрии ($\overline{x}$ больше середины размаха выборки) правый конец расчетного интервала может несколько выходить за рамки возможных значений признака генеральной совокупности, в то время как часть значений может оказаться вне левого конца. Аналогично при правосторонней асимметрии. Однако число возможных значений признака генеральной совокупности, остающихся (если это имеет место) вне расчетного интервала по приведенной формуле, мало; их вероятность в пределах $(1 - \gamma)$.

2. При заданном объеме генеральной совокупности N можно оценить количество ее единиц, обладающих значением признака $x_{\text{к}}$, по формуле

$$N(x_{\text{к}}) = \frac{m_{\text{к}}}{n} N, \qquad (2)$$

где $m_{\text{к}}$ – частота значения $x_{\text{к}}$; n – объем выборки.

Погрешность найденного $N(x_{\text{к}})$ составляет

$$3\Delta\left(1 - \frac{n}{N}\right), \ \%. \qquad (3)$$

Очевидно, погрешность с увеличением объема выборки должна уменьшаться, что имеет место в (3). В самом деле, с возрастанием $n (n \to N)$ сомножитель $\left(1 - \frac{n}{N}\right)$ уменьшается, а следовательно, убывает значение (3). При $n = N$ разность $\left(1 - \frac{n}{N}\right)$ обращается в нуль, т. е. погрешность равна нулю, что тоже соответствует действительности, ибо в этом случае выборка перестает быть таковой – берется вся генеральная совокупность.

В случае задания признака непрерывной величиной (ряд распределения – интервальный) в формулах (1) и (2) $x_{\text{к}}$ – середины интервалов, на которые разбивается размах признака. Следовательно, $N(x_{\text{к}})$ выражает количество единиц генеральной совокупности интервала, содержащего $x_{\text{к}}$.

Таким образом, в вышеприведенном примере среднее значение производительности труда генеральной совокупности любого предприятия оценивается интервалом $\overline{x}_{\text{г}} = \overline{x} \pm \Delta = 21{,}51 \pm \pm 3{,}62 = (17{,}89;\ 25{,}13)$, а ее возможные значения по формуле (1) – интервалом $x_{\text{г}} = \overline{x} \pm 3\Delta = 21{,}51 \pm 10{,}86 = (10{,}65;\ 32{,}37)$.

Как и должно быть, интервал средних значений производительности труда генеральной совокупности предприятий меньше интервала возможных значений и является симметричным подмножеством множества ее вариации.

При наличии интервального ряда (в [1] он отсутствует), по которому получены $\overline{x}$ и Δ, можно было бы оценить по формуле (2) количество предприятий из ста заявленных, обладающих определенным уровнем производительности труда.

Проиллюстрируем применение разработанной методики распространения результатов выборки на генеральную совокупность примером успеваемости студентов, где итоги можно сопоставить с реальными.

Из 500 студентов, сдавших экзаменационную сессию, произведена случайная повторная выборка объема n = 25, представленная дискретным вариационным рядом:

Оценка	$x_{\text{к}}$	2	3	4	5	6	7	8	9	10
Количество студентов	$m_{\text{к}}$	1	1	5	4	4	3	3	2	2

Требуется:

1. Оценить интервалы средней и возможных оценок экзаменовавшихся (в том числе 500) с доверительной вероятностью γ = 0,99.

2. Оценить количество студентов из 500, получивших соответствующую оценку.

Находим среднее арифметическое x и дисперсию выборки D по формулам:

$$\overline{x} = \frac{\sum x_{\text{к}} m_{\text{к}}}{\sum m_{\text{к}}} = \frac{152}{25} = 6{,}08;$$

$$D = \frac{\sum (x_к - \bar{x})^2 m_к}{\sum m_к} = \frac{113,83}{25} = 4,55.$$

Определяем среднюю ошибку по формуле повторной случайной выборки

$$\mu = \sqrt{\frac{D}{n-1}} = \sqrt{\frac{4,55}{24}} = 0,435.$$

При доверительной вероятности $\gamma = 0,99$ и объеме выборки $n = 25$ коэффициент $t(n; \gamma)$ будет $t(25; 0,99) = 2,8$.

Следует отметить, что в большинстве приведенных источников низкий уровень математической культуры, весьма вольное обращение с формулами. Так, в [1–3] при расчетах μ знаменатель $(n - 1)$ заменяется на n. Это допустимо, если объем выборки $n > 30$, а расчеты ведутся до десятых долей [4, 5].

Также не учитывается объем выборки при определении коэффициента t по доверительной вероятности, что приводит к грубым ошибкам. Так, в том же примере с производительностью труда при бесповторном случайном отборе с $n = 20$ $(20 < 30)$ и заданном объеме генеральной совокупности $N = 100$ средняя ошибка должна определяться по формуле $\mu = \sqrt{\frac{D}{n-1}\left(1 - \frac{n}{N}\right)}$ и составит $\mu = 1,86$ вместо приведенной $\mu = 1,81$. Значение параметра t при объеме выборки $n = 20$ и $\gamma = 0,954$ должно быть $t = 2,18$ [6] вместо приведенного $t = 2,00$. Так что предельная ошибка равна $\Delta = 1,86 \cdot 2,18 = 4,05$ вместо $\Delta = 3,62$, т. е. погрешность ее определения составляет $4,05 - 3,62 = 0,43$ (11 %).

В нашем примере предельная ошибка составит $\Delta = \mu t = 0,435 \cdot 2,8 = 1,22$. Следовательно, средняя генеральная оценок в интервале $\bar{x}_г = 6,08 \pm 1,22 = (4,9; 7,3)$, а их вариация в интервале $x_г = 6,08 \pm 3 \cdot 1,22 = (2,4; 9,7)$.

Таким образом, с вероятностью 0,99 (99 случаев из 100; 495 случаев из 500) можно утверждать, что средняя оценка генеральной совокупности (в том числе 500) будет в интервале (4,9; 7,3), а интервал вариации оценок составит (2,4; 9,7), т. е. расчетный интервал возможных значений оценок генеральной совокупности в данном случае после округления до целых совпадает с интервалом выборки (2; 10).

Если доверительную вероятность взять $\gamma = 0,999$ (999 случаев из 1000), то предельная ошибка составит $\Delta = 0,435 \cdot 3,74 = 1,62$, а интервалы будут $\bar{x}_г = 6,08 \pm 1,62 = (4,5; 7,7)$; $x_г = 6,08 \pm 3 \cdot 1,62 = (1,2; 10,9)$. Расчетный интервал оценок с округлением до целых составит (1; 11), т. е. правый конец содержит невозможную оценку 11, а «за бортом» левого оказалась маловероятная оценка 0, что обусловлено значительной левосторонней асимметрией выборки ($\bar{x} = 6,08$ больше середины выборки $\frac{10-2}{2} = 4$ на 2 единицы).

Оценим количество студентов из 500 получивших соответствующую оценку. Расчеты по формуле (2) представлены третьей строкой расчетной табл. 1, а их абсолютная погрешность по формуле (3) $3,66 \cdot 0,95\ \% = 3,48\ \%$ – четвертой строкой.

Таблица 1

Расчетная таблица

Оценка $x_к$	2	3	4	5	6	7	8	9	10	Сумма
Частота $m_к$	1	1	5	4	4	3	3	2	2	25
Количество студентов $N(x_к)$	20	20	100	80	80	60	60	40	40	500
Абсолютная погрешность	0,7	0,7	3,5	2,8	2,8	2,1	2,1	1,4	1,4	18,0

Итак, например, оценку 5 баллов из 500 получат 80 студентов с погрешностью 3 (округление до целых), т. е. их число от 77 до 83; 8 баллов – от 58 до 62 и т. д. Следует отметить, что только у 18 студентов из 500 по приведенным расчетам оценка может быть иной (она возможна из вошедших в выборку, а также 0 и 1, не попавших в нее).

Как показали расчеты на других примерах, предложенная методика дает возможность объективного и всестороннего анализа любого признака генеральной совокупности по результатам выборки. Она проста в применении и обеспечивает достаточно высокую точность результатов.

С целью удобства и грамотных расчетов в табл. 2 в диапазоне применяемых на практике приводятся значения $t(n; \gamma)$.

Таблица 2

Таблица значений $t(n; \gamma)$

n/γ	0,90	0,95	0,99	0,999
5	1,99	2,78	4,60	8,61
6	1,96	2,57	4,03	6,86
7	1,93	2,45	3,71	5,96
8	1,90	2,37	3,50	5,41
9	1,88	2,31	3,36	5,04
10	1,86	2,26	3,25	4,78
12	1,84	2,20	3,11	4,44
14	1,82	2,16	3,01	4,22
16	1,80	2,13	2,95	4,07
18	1,78	2,11	2,90	3,97
20	1,76	2,09	2,86	3,88
25	1,74	2,06	2,80	3,74
30	1,72	2,04	2,76	3,66
35	1,70	2,03	2,72	3,60
40	1,69	2,02	2,71	3,56
50	1,68	2,01	2,68	3,50
60	1,67	2,00	2,66	3,46
70	1,67	1,99	2,65	3,44
80	1,66	1,99	2,64	3,42
90	1,66	1,98	2,63	3,40
100	1,65	1,98	2,63	3,39
120	1,65	1,97	2,62	3,38

В Ы В О Д Ы

Разработана простая и удобная в применении методика распространения данных выборки на генеральную совокупность. Она позволяет всесторонне проанализировать генеральную совокупность с достаточно высокой степенью надежности:

1) оценить интервал вариации признака генеральной совокупности формулой (правило трех дельта) $\overline{x} - 3\Delta \le x_{\text{г}} \le \overline{x} + 3\Delta$;

2) оценить объем генеральной совокупности N, обладающий значением признака $x_{\text{к}}$, формулой $N\left(x_{\text{к}}\right) = \frac{m_{\text{к}}}{n} N$ с точностью $3\Delta\left(1 - \frac{n}{N}\right)\%$.

Л И Т Е Р А Т У Р А

1. **Статистика** / И. И. Колесникова [и др.]. – М., 2007.

2. **Захаренков, С. Н.** Статистика / С. Н. Захаренков. – Минск: БГУ, 2010.

3. **Общая** теория статистики / под ред. Л. И. Карпенко. – Минск: БГЭУ, 2007.

4. **Статистика** автомобильного транспорта / И. М. Алексеева [и др.]. – М., 2005.

5. **Статистика** / под ред. И. М. Елисеевой. – М., 2009.

6. **Гмурман, В. Е.** Теория вероятностей и математическая статистика / В. Е. Гмурман. – М., 2002.

Поступила 30.03.2012