МЕТОДЫ РАСПОЗНАВАНИЯ ЗАКОНА РАСПРЕДЕЛЕНИЯ ПО ВЫБОРКАМ МАЛОГО ОБЪЕМА И ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ИНТЕРВАЛЬНЫХ ОЦЕНОК

Юрий Долгов
Государственный университет им. Т.Г.Шевченко
dolax@mail333.com

Abstract. *It are offered some methods for definition of law distributions and increase estimate effectiveness of small size samples.*

Ключевые слова: *выборка малого объема, центральные моменты, точечные и интервальные оценки.*

I. Введение

При производстве кристаллов ИМС пооперационный выборочный контроль производится только в пяти (десяти) тестовых ячейках на каждой пластине, где располагается 400-5000 рабочих кристаллов. Такое соотношение величин (контрольная выборка составляет около 0,1% от общего объёма) не подходит ни к одному плану гостированного статистического выборочного контроля и поэтому в производстве используется так называемый граничный метод, недостатком которого является слишком большая доля ложной приемки и ложной браковки. Проделанный анализ показал, что эта доля зависит от объективных и субъективных причин [1], причем последние являются следствием нашего незнания законов распределения контролируемых параметров и достаточно грубых (с большой ошибкой) расчетов параметров выборки малого объема. Уменьшение ошибки вычисления параметров выборок малого объема возможно с помощью метода точечных распределений (МТР) [2]. В настоящей статье остановимся на методах определения законов распределения и параметров выборок малого объема и на методах уменьшения ошибок при расчете этих параметров.

II. Определение точечных оценок центральных методов

Согласно классической теории вероятностей вид закона распределения по выборкам большого объема может быть найден графоаналитическим методом через асимметрию (r_3) и эксцесс (r_4) гистограммы [3] по следующей цепочке формул

$$r_3 = \frac{m_3}{m_2^{3/2}}; \qquad r_4 = \frac{m_4}{m_2^2}\ , \tag{1}$$

где

$$m_h = \frac{1}{n-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^h\ ,\ h = 2, 3,4, \tag{2}$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{N} X_i \tag{3}$$

и номограмме рисунка 1.

Однако, как упоминалось выше, при выборках малого объема вычисления их параметров имеют слишком большую ошибку. Так, например, по данным [4] относительная ошибка выборочного среднеквадратического отклонения $s(S)/s$ равна $\sqrt{1/(2n-1{,}4)}$ (при n=5 составляет 0,341 или 34,1%, а при n=10 – 0,232 или23,2%).

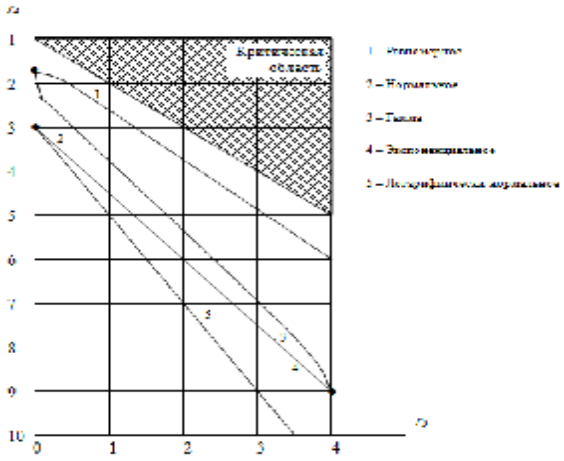


Рисунок 1 – Номограмма определения закона выборочного распределения

Для уменьшения этой ошибки оценки параметров выборок малого объема следует искать по методу точечных распределений (МТР) [2]. Аналогичная цепочка формул с учетом рекомендаций [3] может быть представлена как

$$m_X^* = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} X_j \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]}{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]}; \quad (4)$$

$$m_2^* = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} X_j^2 \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]}{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]} - \left(m_X^*\right)^2 \quad (5)$$

$$m_3^* = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} X_j^3 \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]}{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]} - 3m_2^* m_X^* - \left(m_X^*\right)^3 \quad (6)$$

$$m_4^* = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} X_j^4 \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]}{\sum_{j=1}^{k}\sum_{i=1}^{n} p_{ji} \exp\left[-4{,}5\left(\frac{X_j - X_i}{r}\right)^2\right]} - 4m_3^* m_X^* - 6m_2^*\left(m_X^*\right)^2 - \left(m_X^*\right)^4 \quad (7)$$

где X_i – экспериментальные данные; X_j – середина j-го разряда ($j = \overline{1,k}$), полученного путём деления отрезка (b-a) на k частей (a – начало, b – конец отрезка эквивалентной выборки); p_{ji}=1 при условии накрывания величиной X_j ядра на основе X_i; ρ – половина ширины ядра.

Тогда коэффициент асимметрии можно подсчитать по формуле

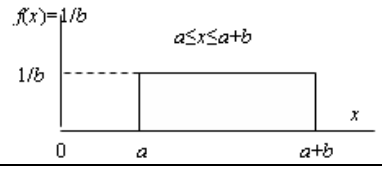
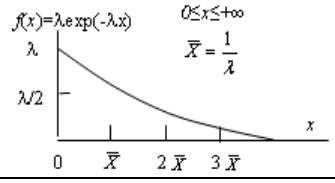
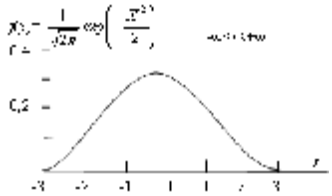
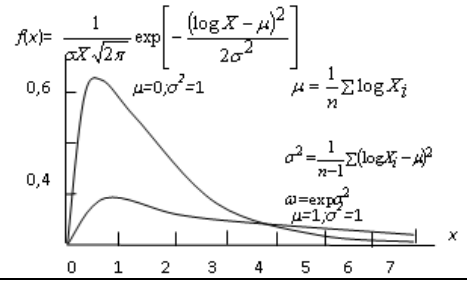
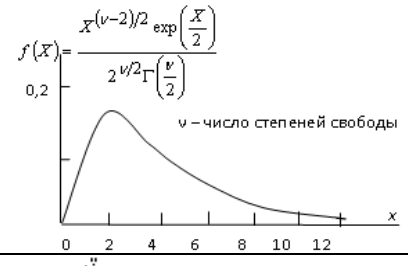
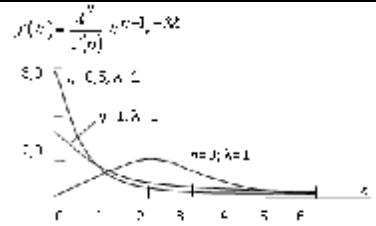
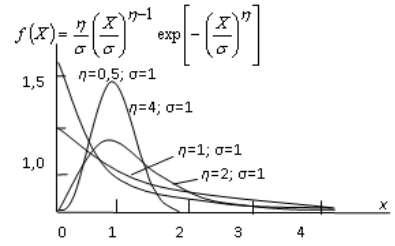
$$r_3^* = \frac{m_3^*}{\left(m_2^*\right)^{3/2}}, \quad (8)$$

а коэффициент эксцесса – по формуле

$$r_4^* = \frac{m_4^*}{\left(m_2^*\right)^2} \ . \quad (9)$$

Поскольку вычисления выборочных моментов по формулам (4)-(9) неизбежно являются приблизительными, то для ориентировки в таблице 1 приведены теоретические значения коэффициентов асимметрии и эксцесса [5].

Таблица 1 – Теоретические значения коэффициентов асимметрии (r_3) и эксцесса (r_4) некоторых распределений

Распределение	Вид и форма	r_3	r_4
Равномерное		0	1,8
Экспоненциальное		2	9
Нормальное		0	3
Логарифмически нормальное		$(w+2)\sqrt{w-1}$	$w^4+2w^3+3w^2-3$
Хи-квадрат		$2\sqrt{\frac{2}{n}}$	$3+\frac{12}{n}$
Гамма		$\frac{2}{\sqrt{h}}$	$\frac{3(h+2)}{h}$
Вейбулла		*	**

$$* \quad r_3 = \frac{\Gamma\left(1+\frac{3}{h}\right) - 3\Gamma\left(1+\frac{2}{h}\right)\Gamma\left(1+\frac{1}{h}\right) + 2\left[\Gamma\left(1+\frac{1}{h}\right)\right]^3}{\left\{\Gamma\left(1+\frac{2}{h}\right) - \left[\Gamma\left(1+\frac{1}{h}\right)\right]^2\right\}^{3/2}};$$

$$** \quad r_4 = \frac{\Gamma\left(1+\frac{4}{h}\right) - 4\Gamma\left(1+\frac{3}{h}\right)\Gamma\left(1+\frac{1}{h}\right) + 6\Gamma\left(1+\frac{2}{h}\right)\left[\Gamma\left(1+\frac{1}{h}\right)\right]^2 - 3\left[\Gamma\left(1+\frac{1}{h}\right)\right]^4}{\left\{\Gamma\left(1+\frac{2}{h}\right) - \left[\Gamma\left(1+\frac{1}{h}\right)\right]^2\right\}^4}.$$

Пример 1. Пусть в результате измерений контрольной выборки в некотором технологическом процессе получены значения (в условных единицах) X: 0, 5, 11, 17, 23, 28, 33, 39.

Определить вид распределения.

Решение. С помощью метода точечных распределений (МТР) найдём выборочные моменты

$m_X^* = 19{,}62$; $m_2^* = 11{,}3917$; $m_3^* = 4{,}0451$, откуда $r_3^* = 0{,}0027 \to 0$

$m_4^* = 28260{,}1901$, откуда $r_4^* = 1{,}6781 \to 1{,}8$

По номограмме рисунка 1 с большой долей вероятности можно сделать предположение, что выборочное распределение стремится к равномерному виду.

III. Определение интервальных оценок центральных моментов

В отличие от генеральных совокупностей, параметры которых имеют только однозначно неслучайные величины, выборки характеризуются оценками генеральных параметров, которые имеют при некоторой доверительной вероятности минимальные и максимальные значения, внутри которых находятся и сами генеральные параметры, причем в большинстве случаев расположенные несимметрично (интервальные оценки).

Среди всех начальных и центральных методов главенствующую роль по частоте использования имеют средние арифметические (3) и выборочные дисперсии (2 при h= 2).

Проблема интервальных оценок для выборок большого объема хорошо исследована (например, [6])

$$\overline{X} - t_{табл}(q;n)\sqrt{\frac{S^2}{n}} < M[X] < \overline{X} + t_{табл} \qquad (10)$$

$$\frac{nS^2}{c_2^2} < s^2 < \frac{nS^2}{c_1^2} \qquad (11)$$

где n – объем выборки; $t_{табл}(q;\nu)$ – табличный критерий Стьюдента при q – уровне значимости и ν=n-1 – числе степеней свободы; $c_1^2\left(1-\frac{q}{2};n\right)$ и $c_2^2\left(\frac{q}{2};n\right)$ – хи-квадрат распределение при $\frac{q}{2}$ и $1-\frac{q}{2}$ уровнях значимости и ν – числе степеней свободы; $S^2 = m_2$ – выборочная дисперсия.

Проблема интервальных оценок для выборок малого объема исследована не так хорошо и оставляет простор для дискуссий. Понятно, что использую формулы (10) и (11) для выборок малого объема мы получим чрезмерно большие интервалы (пример 2). Для их уменьшения в работе [7] предлагается воспользоваться следующими выражениями, базирующимся на методе точечных распределений:

для нормального и других симметричных распределений

$$m_X^* - U_{\frac{1-b}{2}}\sqrt{m_2^*} < M[X] < m_X^* + U_{\frac{1+b}{2}}\sqrt{m_2^*}; \qquad (12)$$

для экспоненциального и аналогичных ему распределений

$$U_{\frac{1-b}{2}} \cdot m_X^* < M[X] < U_{\frac{1+b}{2}} \cdot m_X^*; \quad (13)$$

для распределения Вейбулла и аналогичных ему

$$U_{\frac{1-b}{2}} \cdot \frac{m_X^*}{0{,}886} < M[X] < U_{\frac{1+b}{2}} \cdot \frac{m_X^*}{0{,}886}, \quad (14)$$

для выборочной дисперсии в случае нормального, экспоненциального и подобных распределений

$$\frac{m_2^*}{U_{\frac{1+b}{2}}} < M[m_2] < \frac{m_2^*}{U_{\frac{1-b}{2}}}; \quad (15)$$

для распределения Вейбулла и аналогичных ему

$$\frac{0{,}214 m_2^*}{U_{\frac{1+b}{2}}} < M[m_2] < \frac{0{,}214 m_2^*}{U_{\frac{1-b}{2}}}, \quad (16)$$

где β - доверительная вероятность, а U_2 – соответствующая квантиль.

Однако, на наш взгляд, существует ещё одна возможность получить интервальные оценки повышенной точности для выборок малого объема. Дело в том, что согласно основополагающей концепции метода точечных распределений, количество информации до и после всех преобразований не должно изменяться, а , следовательно, можно записать равенство

$$D_n(b)\sqrt{n_э} = D(b)\sqrt{n}$$

откуда

$$n_э = n\left[\frac{D(b)}{D_n(b)}\right]^2, \quad (17)$$

где $D(\beta)$ – статистика Колмогорова при доверительной вероятности β [8]; $D_n(\beta)$ – аналогичная ей статистика МТР [7]; n – объем экспериментальной выборки малого объема; $n_э$ – эквивалентный объем выборки, обеспечивающий ту же точность $D(\beta)$ при расчете по классическим формулам.

Некоторые данные расчетов по формуле (17) приведены в таблице 2.

Таблица 2 – Реальные n и эквивалентные $n_э$ объемы выборок

n	3	4	5	6	7	8	9	10
D/D_n	1,863	1,891	1,876	1,821	1,791	1,780	1,741	1,690
$n_э$	10	14	17	20	23	25	27	29
$s(S)/s$, %	46,6	38,9	34,1	30,7	28,2	26,2	24,5	23,2
$s\left(\sqrt{m_2^*}\right)/s$, %	23,2	19,4	17,5	16,1	15,0	14,3	13,8	13,3

Тогда, опираясь на эквивалентные объемы выборок $n_э$, можно использовать для расчета интервальных оценок моментов классические формулы (10) и (11).

$$m_X^* - t_{табл}(q;n)\sqrt{\frac{m_2^*}{n_э}} < M[X] < m_X^* + t_{табл}(q;n)\sqrt{\frac{m_2^*}{n_э}}; \quad (18)$$

$$\frac{n_э m_2^*}{c_2^2} < s^2 < \frac{n_э m_2^*}{c_1^2}, \quad (19)$$

где $n = n_э - 1$; $c_1^2\left(1-\frac{q}{2}; n = n_э - 1\right)$; $c_2^2\left(\frac{q}{2}; n = n_э - 1\right)$.

Произведём сравнение всех методов расчета интервальной оценки выборочной дисперсии с точки зрения их эффективности.

Пример 2. Пусть в результате измерений контрольной выборки в некотором технологическом процессе получены значения величин контролируемого параметра, распределенного по нормальному закону, ***X***: 5,73; 6,40; 7,70; 8,07; 8,47.

Найти интервальные оценки выборочной дисперсии всеми возможными способами.

Решение. Объем выборки n=5, а её параметры вычисленные по формулам (3) и (2), равны

$$\overline{X} = 7{,}274\,;\ S^2 = m_2 = 1{,}3483\,.$$

Те же вычисления МТР по формулам (4) и (5) равны $m_X^* = 7{,}286\,;$ $m_2^* = 1{,}7370\,.$

Тогда согласно формуле (11) и таблицам χ^2[8]

$$\frac{5\cdot 1{,}3483}{10{,}14} < s^2 < \frac{5\cdot 1{,}3483}{0{,}484} \text{ или } 0{,}605 < s^2 < 13{,}929\,,$$

согласно формуле (15) $\dfrac{1{,}737}{1{,}960} < s^2 < \dfrac{1{,}737}{0{,}125}$ или $0{,}886 < s^2 < 13{,}896\,,$

согласно формуле (19) $\dfrac{17\cdot 1{,}737}{28{,}85} < s^2 < \dfrac{17\cdot 1{,}737}{6{,}908}$ или $1{,}024 < s^2 < 4{,}275\,.$

Для наглядности представим полученные результаты в графическом виде (рисунок 2).

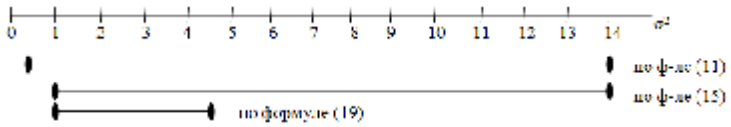


Рисунок 2 – Величины интервальных оценок выборочных дисперсий по разным формулам

IV. Заключение

Разработан метод, позволяющий в случае выборок малого объема (n=3÷16) определять хотя бы приблизительно вид закона распределения генеральной совокупности, из которой взята конкретная выборка, а также в 3-4 раза сокращать размер интервальной оценки выборочной дисперсии, что позволяет в 2-2,5 раз уменьшить долю ложной приемки и ложной браковки за счет уменьшения ошибки вычисления параметров контрольных выборок малого объема.

V. Библиография

1. Dolgov Y.A., Dolgov A.Y. Stochastic Check for Control of Electronic Wares Quality // Trans. of 10-th Intern. Symp. of Apllied Stochastic Models and Data Analysis. – 12-15 June 2001. – v.1/2. – Univer.Technologic de Compiegne. – France. – P.387-390.

2. Долгов Ю.А., Долгов А.Ю., Столяренко Ю.А. Метод повышения точности вычисления параметров выборки малого объема (метод точечных распределений) // Вестник ПГУ-2010. – Юб. вып.– С.232-242.

3. Хан Г., Шапиро С. Статистические модели в инженерных задачах. – М.: Мир, 1969. – 396 с.

4. Шор Я.Б. Статистические методы анализа и контроля качества и надежности. – М.: Сов. радио, 1962. – 553 с.

5. Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. – М.: Статистика, 1980 – 95с.

6. Митропольский А.К. Техника статистических вычислений. – 2-е изд., перераб. и доп. – М.: Наука, 1971. – 576 с.

7. Гаскаров Д.В., Шаповалов В.И. Малая выборка. – М.: Статистика, 1978. – 248 с.

8. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – 3-е изд.– М.: Наука, 1983. – 416 с.