УДК 519.24.(075.8)

Ю.А. ДОЛГОВ

Приднестровский государственный университет им. Т.Г. Шевченко, Тирасполь, Приднестровье, Молдова

ОПРЕДЕЛЕНИЕ ОБЪЕМА ЭКВИВАЛЕНТНОЙ ВЫБОРКИ В МЕТОДЕ ТОЧЕЧНЫХ РАСПРЕДЕЛЕНИЙ

Предлагается метод определения объёма эквивалентной (виртуальной) выборки, полученной методом точечных распределений, из исходной выборки малого объёма (n=3× 20 элементов) при равенстве количества информации в обеих выборках. Найдены прямые зависимостип$_э$= f(n) для различных законов распределения, которые свидетельствуют об увеличении n$_э$ в 2,5-3,2 раза по сравнению с объёмом исходной выборки. Это позволяет уменьшить ошибку вычисления параметров выборки в 1,6-2,5 раза, а размах интервальной оценки – примерно в 3 раза для среднего арифметического и в 9 раз для выборочной дисперсии.

Ключевые слова: *выборка малого объема, эквивалентная выборка, статистика Колмогорова.*

Постановка задачи

Классические методы статистической обработки любой выборки большого объёма, как правило, основаны на идее группировки данных (гистограмма и др.) [1]. Однако теоретическими трудами доказано, что группировка данных вызывает уменьшение информации, которая извлекается из выборок. Если при применении методов обработки, основанной на группировке данных, можно получить необходимую точность при заданной достоверности, то выборка содержит избыточную информацию и является достаточной. Выборку следует считать малой, если при её обработке теми же методами нельзя достичь заданных точности и достоверности [2].

Установлено, что все выборки чётко делятся на три диапазона: выборки малого объёма (n=3ч20 элементов) [3], выборки среднего объёма (n=20ч100 элементов) [4] и выборки большого объёма (свыше 100 элементов) [1]. Выборки каждого диапазона имеют свои особенности при обработке, которые позволяют уменьшить потери информации и тем самым повысить точность и достоверность рассчитываемых параметров. Так для уменьшения потерь информации при обработке выборок малого объема необходимо отказаться от группировки данных и прейти к методу, основанному на использовании каждой отдельной реализации (измерения, числового значения), для чего считать каждое измерение центром распределения с известным законом – метод точечных распределений (МТР) [5]. Это позволяет существенно уменьшить интервал неопределённости выборочных оценок [6], что, в свою очередь, позволяет успешно решать ряд практических задач, например, значительно снизить объёмы контрольных выборок и применить известные статистические методы для разбраковки продукции по ходу технологического процесса там, где ранее это было принципиально невозможно [7].

В настоящей статье представлен алгоритм определения объёма эквивалентной выборки в МТР и даны количественные характеристики уменьшения относительной ошибки вычисления оценок параметров выборок при различных законах распределения.

Методы исследования

Вопрос о величине объёма выборки, эквивалентной в смысле полученной информации той, с которой мы имели бы дело при классических методах определения оценок параметров выборки (например, средних арифметических, эмпирических дисперсий и др.), можно для нормального закона распределения решить через D-статистику Колмогорова [8] и аналогичную ей D_n-статистику, вычисленную в условиях выборок малого объёма n (рис. 1).

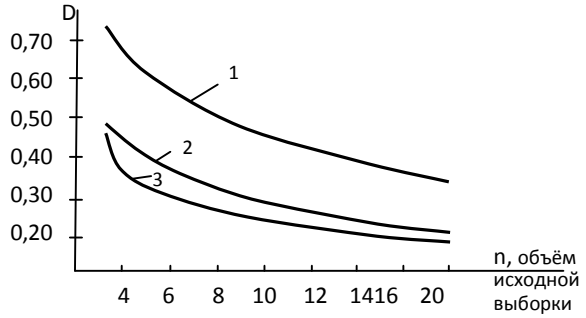

Рис. 1. Различные D-статистики при $P_{дов}$= 0,99: 1 – D-статистика Колмогорова; 2 – D_n-статистика при нормальном распределении; 3 – D_n-статистика при распределении Вейбулла

Считая, что количество информации в малой и эквивалентной (виртуальной) выборках одинаковы, можно написать равенство для одной и той же доверительной вероятности в

$$D(\beta)\sqrt{n} = D_n(\beta)\sqrt{n_э}, \quad (1)$$

решая которое найдем

$$n_э = n\left[D(\beta)/D_n(\beta)\right]^2, \quad (2)$$

где n – объём исходной малой выборки; $n_э$ – объём равной ей по информации эквивалентной выборки при условии нахождения оценок её параметров по классическим формулам математической статистики вместо специальных формул МТР [6].

Результаты расчетов представлены в табл. 1.

Соотношения эквивалентных и исходных объёмов выборок, представленные в таблице, с помощью метода наименьших квадратов (МНК) можно трансформировать в прямые зависимости $n_э = f(n)$:

для нормального и сопутствующих распределений

$$n_э = 0,8963 + 3,2080n - 0/0628n^2 + 0,0008n^3, \quad (3)$$

для распределения Вейбулла и сопутствующих

$$n_э = \sqrt{-713,7 + 301,2n - 6,907n^2}, \quad (4)$$

которые были выбраны из нескольких десятков возможных формул по двум показателям: самому большому индексу корреляции (I > 0,99) и минимальному среднеквадратическому отклонению (СКО) функции от исходных данных [5] (рис. 2).

Таким образом, следует констатировать, что объёмы эквивалентных выборок виртуально увеличиваются в 2,5-3,2 раза и при расчётах с использованием в МТР любых статистических критериев (Стьюдента, Фишера и т.п.) следует придерживаться именно объёма $n_э$ при вычислении числа степеней свободы.

Таблица 1

Значения объёмов и относительных ошибок СКО
исходной и эквивалентной выборки при различных законах распределения

Закон распределения	n	3	4	5	6	7	8	9	10	11
	y(S)/y, %	46,3	38,9	34,1	30,7	28,2	26,2	24,5	23,2	22,0
Нормальный, экспоненциальный, Парето, Стьюдента и др.	D/D_n	1,817	1,803	1,766	1,727	1,703	1,684	1,680	1,661	1,649
	$n_э$	9,9	13,0	15,6	17,9	20,3	22,7	25,4	27,6	29,9
	$\sigma\left(\sqrt{\mu_2^*}\right)/\sigma$, %	23,3	20,2	18,2	17,0	16,0	15,1	14,2	13,6	13,1
Вейбулла, гамма, хи-квадрат, логнормальный и др.	D/D_n	1,983	2,253	2,218	2,183	2,145	2,098	2,052	2,000	1,954
	$n_э$	11,8	20,3	24,6	28,6	32,2	35,2	37,9	40,0	42,0
	$\sigma\left(\sqrt{\mu_2^*}\right)/\sigma$, %	21,2	16,0	14,5	13,4	12,6	12,0	11,6	11,3	11,0
Закон распределения	n	12	13	14	15	16	17	18	19	20
	y(S)/y, %	21,0	20,2	19,4	18,7	18,1	17,5	17,0	16,5	16,1
Нормальный, экспоненциальный, Парето, Стьюдента и др.	D/D_n	1,633	1,612	1,597	1,581	1,567	1,560	1,544	1,537	1,525
	$n_э$	32,0	33,8	35,7	37,5	39,3	41,4	42,9	44,9	46,5
	$\sigma\left(\sqrt{\mu_2^*}\right)/\sigma$, %	12,6	12,3	12,0	11,7	11,4	11,1	10,9	10,6	10,4
Вейбулла, гамма, хи-квадрат, логнормальный и др.	D/D_n	1,917	1,865	1,824	1,778	1,737	1,694	1,658	1,621	1,594
	$n_э$	44,1	45,2	46,6	47,4	48,3	48,8	49,5	49,9	50,8
	$\sigma\left(\sqrt{\mu_2^*}\right)/\sigma$, %	10,7	10,6	10,4	10,3	10,2	10,2	10,1	10,1	10,0

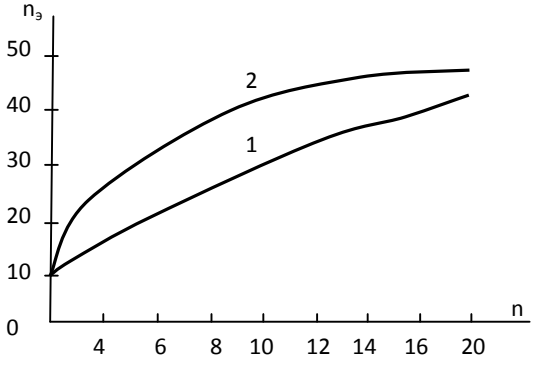

Рис. 2. Зависимости $n_э = f(n)$
при различных законах распределения:
1 – нормальном; 2 – Вейбулла

Следует также обратить внимание на относительную ошибку вычисления параметров выборки. Для СКО выборок малого объёма Я.Б. Шор нашел аппроксимирующую формулу [9]

$$\sigma(S)/\sigma = \sqrt{1/(2n-1,4)}, \quad (5)$$

числовые значения которой приведены в таблице. Если применить формулу (5) к эквивалентным объёмам виртуальных выборок

$$\sigma\left(\sqrt{\mu_2^*}\right)/\sigma = \sqrt{1/(2n_э - 1,4)}, \quad (6)$$

числовые значения которой также приведены в таблице, то видно, что ошибка вычисления оценок параметров уменьшается в разы (1,6–2,5 раза), что су-

щественно сокращает диапазон интервальных оценок и во столько же раз увеличивает точность оценок параметров генеральной совокупности.

В заключение отметим, что хотя D-функция Колмогорова и соответствующая ей D_n-функция разработаны только для нормального распределения и распределения Вейбулла, тем не менее с небольшой дополнительной ошибкой выводами и формулами данной статьи можно воспользоваться и для других распределений, разделив их на два класса: те, у которых область существования определена в диапазоне (-∞; +∞) (экспоненциальное и Парето распределения являются исключением, так как тесно связаны с нормальным распределением) и те, у которых область существования определена в диапазоне (0; +∞).

Заключение

Разработан метод определения объёма эквивалентной (виртуальной) выборки, полученной методом точечных распределений, из исходной выборки малого объёма (n = 3ч20 элементов) при равенстве количества информации в обеих выборках. Найдены прямые зависимости $n_э = f(n)$ для различных законов распределения, которые свидетельствуют об увеличении $n_э$ в 2,5-3,2 раза по сравнению с объёмом исходной выборки. Это позволяет уменьшить ошибку вычисления параметров выборки в 1,6 - 2,5 раза, а размах интервальной оценки – примерно в 3 раза для среднего арифметического и в 9 раз для выборочной дисперсии.

Литература

1. Митропольский, А.К. Техника статистических вычислений [Текст] / А.К. Митропольский. – 2-е изд., перераб. и доп. – М.: Наука, 1971. – 576 с.

2. Гаскаров, Д.В. Малая выборка [Текст] / Д.В. Гаскаров, В.И. Шаповалов. – М.: Статистика, 1978. – 248 с.

3. Столяренко, Ю.А. Контроль кристаллов интегральных схем на основе статистического моделирования методом точечных распределений: дис. ... канд. техн. наук [Текст] / Ю.А. Столяренко. – М.: ГУП НПЦ «СПУРТ», 2006. – 191 с.

4. Efron, B. The Jacknife, the Bootstrap and Other Reasamping Plans [Текст] / B. Efron. – Philadelphia, Pa.: SIAM, 1982.

5. Долгов, Ю.А. Статистическое моделирование [Текст]: учеб. для вузов / Ю.А. Долгов. – 2-е изд., доп. – Тирасполь: Полиграфист, 2011. – 352 с.

6. Долгов, Ю.А. Исследование интервальных оценок и моментов высшего порядка для выборок малого объёма [Текст] / Ю.А. Долгов // Радіоелектронні і комп'ютерні системи. – 2012. – №5(57). – С. 165-170.

7. Долгов, А.Ю. Методы повышения эффективности выборочного контроля при производстве кристаллов микросхем [Текст] / А.Ю. Долгов // Радіоелектронні і комп'ютерні системи. – 2012. – № 6 (58). – С. 119-123.

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

9. Шор, Я.Б. Статистические методы анализа и контроля качества и надежности [Текст] / Я.Б. Шор. – М.: Сов. радио, 1962. – 553 с.

Поступила в редакцию 8.02.2013, рассмотрена на редколлегии 6.03.2013

Рецензент: д-р техн. наук, доцент, проф. каф. компьютерных систем и сетей А.В. Горбенко, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.

ВИЗНАЧЕННЯ ОБСЯГУ ЕКВІВАЛЕНТНОЇ ВИБІРКИ У МЕТОДІ ТОЧКОВИХ РОЗПОДІЛІВ

Ю.О. Долгов

Пропонується метод визначення обсягу еквівалентної вибірки, отриманої на основі вихідної вибірки малого об'єму (n = 3 × 20 елементів) при використанні методу точкових розподілів. Знайдено прямі залежності $n_э = f(n)$ для різних законів розподілу, які свідчать про збільшення $n_э$ в 2,5 - 3,2 рази порівняно з обсягом вихідної вибірки. Це дозволяє зменшити помилку обчислення параметрів вибірки в 1,6 - 2,5 рази, а розмах інтервальної оцінки – приблизно в 3 рази для середнього арифметичного і в 9 разів для вибіркової дисперсії.

Ключові слова: вибірка малого обсягу, еквівалентна вибірка, статистика Колмогорова.

DEFINITION OF EQUIVALENT SAMPLE SIZE IN POINT DISTRIBUTION METHOD

Y.A. Dolgov

It is offered the method of definition of equivalent sample size on the basis of small size sample (n = 3 × 20 elements) in point distribution method. There are found the direct dependences of $n_э = f(n)$ for various laws of distribution which testify to increase in $n_э$ by 2,5 - 3,2 times in comparison with the volume of initial selection are found. It allows to reduce an error of calculation of parameters of selection by 1,6 - 2,5 times, and scope of an interval assessment – approximately by 3 times for an arithmetic average and by 9 times for selective dispersion.

Key words: small size sample, equivalent sample, Colmogorov's statistics.

Долгов Юрий Александрович – действ. член РАЕН, д-р техн. наук, проф. кафедры «Информационные технологии и автоматизированное управление производственными процессами» Приднестровского государственного университета им. Т.Г. Шевченко, Тирасполь, Приднестровье, Модлова, e-mail: dolax2012@ yandex.ru.