

Об оценивании параметров TSP-распределения на основе экспериментальных данных

А. В. Степанов

ВНИИМ им. Д.И. Менделеева

stepanov17@yandex.ru

А. Г. Чуновкина

ВНИИМ им. Д.И. Менделеева

A.G.Chunovkina@vniim.ru

Аннотация. Для семейства двусторонних степенных распределений рассматриваются различные методы оценивания параметров по экспериментальным данным: метод максимального правдоподобия, метод моментов, метод выбора наиболее подходящего распределения из семейства (на основе обратного отображения). Оцениваются смещения оценок параметров, полученных с помощью данных методов; для оценок максимального правдоподобия получены корректирующие коэффициенты. Также сравнивается эффективность различных оценок. Рассматриваются некоторые иные типовые задачи: оценивание среднеквадратичного отклонения случайной величины, принадлежащей семейству TSP (включая получение корректирующих коэффициентов для получения несмещенных оценок), оценивание параметра положения. Накладываются ограничения на возможные значения параметров распределения, подлежащие определению, исходя из имеющихся выборочных данных.

Ключевые слова: случайная величина; оценка параметров распределения; метод максимального правдоподобия; эффективность оценок; метод Монте-Карло

I. ВВЕДЕНИЕ

В метрологической практике, при анализе результатов измерений и оценке их точности, традиционно доминирует подход, основанный на предположении о нормальном распределении погрешностей (неопределенностей). Это обусловлено, во-первых, удобством математического аппарата, связанного с нормальным законом, и, во-вторых, эмпирическим наблюдением, что распределение погрешностей во многих случаях приблизительно соответствует нормальному. В то же время, современные возможности математического моделирования (широкая доступность вычислительной техники и специальных программ) и статистического анализа открывают новые перспективы, позволяя использовать гораздо более широкий спектр вероятностных распределений для адекватного описания погрешностей измерений в тех случаях, когда традиционный подход, постулирующий нормальность распределений, может оказаться недостаточно точным, что может привести к существенному искажению результатов анализа. В качестве тривиальных примеров можно упомянуть существенную асимметрию фактических распределений или естественные ограничения на множество значений (в то время как носитель нормального распределения неограничен). Проблема выбора подходящего закона распределения становится особенно актуальной при небольшом объеме экспериментальных данных, когда

классические статистические критерии проверки гипотез о виде распределения оказываются малоэффективными. Возникает необходимость в разработке общих принципов выбора семейства вероятностных распределений для обработки результатов измерений. При этом важно сохранить преемственность со сложившейся метрологической практикой, где преобладает использование нормального и, в отдельных случаях, равномерного распределений. Это означает, что выбранное семейство распределений должно содержать в себе хорошие аппроксимации как нормального, так и равномерного законов, обеспечивая совместимость с уже существующими методиками обработки данных. В то же время, оно должно быть достаточно широким, чтобы охватить и другие типы распределений, в частности, асимметричные, более точно отражающие реальную природу погрешностей для конкретных измерительных задач. Важно также учитывать физическую природу погрешностей и использовать доступную априорную информацию об их возможных источниках, для сужения круга возможных распределений. Например, если известно, что погрешности имеют ограниченный диапазон значений, то равномерное, усеченное нормальное распределение или распределение с ограниченным носителем может быть более подходящим выбором, чем нормальное.

Ниже в качестве такого семейства будет рассмотрено семейство двусторонних степенных (Two-Sided Power, TSP) распределений [1, 2]. В работе [1] с помощью диаграммы моментов показано, что данное семейство является достаточно широким, несмотря на крайнюю простоту описания (которая существенным образом облегчает работу с ним).

II. TSP-РАСПРЕДЕЛЕНИЕ

Двусторонним степенным распределением (Two-Sided Power, TSP) называется непрерывное распределение, плотность которого имеет следующий вид [1]:

$$\begin{aligned} f(x|a,m,b,p) &= c ((x-a)/(m-a))^{p-1}, \quad a < x < m, \\ f(x|a,m,b,p) &= c ((b-x)/(b-m))^{p-1}, \quad m \leq x < b, \quad (1) \\ c &= p/(b-a). \end{aligned}$$

Параметр p положителен и характеризует форму распределения, при $p \geq 1$ плотность (1) определена на отрезке $[a, b]$ (равна нулю в его граничных точках), при $p < 1$ – на интервале (a, b) . Распределение является двухмодальным при $p < 1$, равномерным при $p = 1$ и одномодальным при $p > 1$ (при $p = 2$ – треугольным), при

этом мода распределения совпадает с параметром m . Факт принадлежности случайной величины x семейству (1) будем обозначать следующим образом: $x \in \square TSP(a, m, b, p)$. Ниже, в основном, будут рассматриваться следующие частные случаи распределения (1): симметричное TSP-распределение $TSP(x_0, r, p)$, для которого $a = x_0 - r$, $m = x_0$, $b = x_0 + r$, $r > 0$, и стандартное TSP-распределение (Standard TSP, STSP) [2], для которого $a = 0$, $b = 1$.

III. ОЦЕНИВАНИЕ ПАРАМЕТРОВ TSP-РАСПРЕДЕЛЕНИЯ

Рассмотрим вопрос об оценивании параметров распределения (1), на примере STSP-распределения. В этом случае оценке подлежат параметры m , p . Далее будем считать, что имеется выборка x_i ($1 \leq i \leq n$) длины n из распределения (1). Обозначим $x_{(1)} \leq \dots \leq x_{(n)}$ соответствующий вариационный ряд.

A. Метод максимального правдоподобия

Оценки m_{ML} , p_{ML} максимального правдоподобия параметров STSP-распределения имеют вид [2]:

$$m_{ML} = x_k, p_{ML} = -n / \log P_j, j = \arg \max_i P_i, \\ P_i = a_{1,i} \cdot \dots \cdot a_{i-1,i} \cdot b_{i-1,i} \cdot \dots \cdot b_{n,i},$$

где $a_{k,i} = x_{(k)} / x_{(i)}$, $b_{k,i} = (1 - x_{(k)}) / (1 - x_{(i)})$.

B. Метод моментов

Используя метод моментов, можно установить следующие соотношения на оценки параметров m , p :

$$p = (1 - m - M_1) / (M_1 - m), \\ m^2 p / (p + 2) + 2(1 - m)(1 + m p) / ((p + 1)(p + 2)) = M_2,$$

здесь $M_{1,2}$ – выборочные моменты порядков 1 и 2, соответственно. Решая данную систему относительно m , p , получим значения оценок m_M , p_M .

C. Метод подбора наиболее подходящего распределения на основе обратного отображения

Данный метод был предложен в работе [3] и заключается в следующем. Из конечного множества «претендентов», характеризующихся наборами параметров $s = (m, p)$, выбирается такой, который доставляет минимум функционалу

$$d^2 = \sum (F_s^{-1}((i - 1/2) / n) - x_{(i)})^2,$$

где F_s – функция распределения семейства, отвечающая конкретному набору параметров s . В случае STSP-распределения обратная к ней функция F^{-1} имеет вид:

$$F^{-1}(y) = (y m^{p-1})^{1/p}, 0 \leq y \leq m, \\ F^{-1}(y) = 1 - ((1 - y)(1 - m)^{p-1})^{1/p}, m < y \leq 1.$$

При этом при применении метода подбора наиболее подходящего распределения использовался метод последовательного измельчения сетки значений параметров, когда последующие распределения-«претенденты» рассматривались в окрестности точки сетки, выбранной на предыдущем шаге.

D. Сравнение методов оценки параметров распределения

Проведем сравнение предложенных выше методов получения оценок: максимального правдоподобия (ММП), моментов (ММ), подбора наиболее подходящего распределения (в табл. I обозначен как «подбор», далее для краткости иногда будем называть его «методом подбора») и комбинированного метода («ММП + коррекция»), когда начальные оценки параметров получены с помощью метода максимального правдоподобия, а затем уточняются путем поиска наиболее подходящего распределения семейства путем варьирования параметров в окрестности начальных значений. Предположим, что фактическое значение параметра m равно 0.5 (рассматриваемое распределение является симметричным), в таблице ниже приведены математические ожидания и среднеквадратические отклонения для распределений оценок, полученных методом Монте-Карло (здесь и далее число симуляций для каждого примера составляло 10^6).

ТАБЛИЦА I.

n	метод	p	свойства оценок			
			E(m)	$\sigma(m)$	E(p)	$\sigma(p)$
20	ММП	1.5	0.50	0.26	1.74	0.39
		2.0	0.50	0.15	2.24	0.53
		3.0	0.50	0.07	3.31	0.79
	ММ	1.5	0.50	0.21	1.73	0.42
		2.0	0.50	0.13	2.24	0.55
		3.0	0.50	0.07	3.32	0.82
	подбор	1.5	0.50	0.23	1.70	0.42
		2.0	0.50	0.13	2.21	0.53
		3.0	0.50	0.07	3.23	0.73
	ММП + корр.	1.5	0.50	0.25	1.67	0.39
		2.0	0.50	0.14	2.18	0.53
		3.0	0.50	0.07	3.24	0.77
50	ММП	1.5	0.50	0.18	1.59	0.22
		2.0	0.50	0.09	2.09	0.30
		3.0	0.50	0.04	3.11	0.45
	ММ	1.5	0.50	0.16	1.59	0.23
		2.0	0.50	0.09	2.09	0.31
		3.0	0.50	0.05	3.12	0.46
	подбор	1.5	0.50	0.16	1.57	0.24
		2.0	0.50	0.08	2.07	0.31
		3.0	0.50	0.04	3.08	0.46
	ММП + корр.	1.5	0.50	0.17	1.55	0.23
		2.0	0.50	0.08	2.06	0.30
		3.0	0.50	0.04	3.09	0.45

Легко заметить, что в рассмотренных случаях эффективности оценок сопоставимы. Оценка параметра p смещена вверх (завышена), при этом метод подбора (в том числе, используемый для коррекции результатов, полученных ММП), приводит к несколько меньшему смещению оценки параметра p (отчасти это объясняется априорным ограничением, наложенным на значения параметра: при проведении моделирования предполагалось, что значения параметра p не превосходят пяти, что отсекает значительные выбросы). Оценки параметра m для рассмотренных примеров

являются несмещенными. Тем не менее, в общем случае, это не так. Дальнейшие численные эксперименты показывают [4], что, в случае асимметричного распределения ($m \neq 0.5$) оценка параметра m смещена в сторону более отдаленного конца отрезка $[0, 1]$ тем сильнее, чем ближе значение m к границам данного отрезка. При этом относительное смещение оценки параметра m не слишком велико, по сравнению с относительным смещением оценки p (величины относительных смещений различаются примерно на порядок). Наибольших значений это смещение достигает при значениях p вблизи единицы (убывая с ростом p), что легко объяснимо: распределение при этом имеет относительно «пологий» вид, близкий к равномерному распределению, и его мода слабо различима, даже при значительном объеме выборки, в отличие от случая относительно «островерхового» распределения, отвечающего значениям параметра p , начиная от 2–3 и выше. В последнем случае, особенно при достаточно большом объеме выборки (несколько десятков), смещение оценки параметра m , с точки зрения практических приложений, представляется пренебрежимо малым.

Е. Введение поправочных коэффициентов

Метод Монте-Карло позволяет получить корректирующие коэффициенты для оценок параметров распределения. Рассмотрим еще раз случай симметричного STSP-распределения. В табл. II ниже даны поправочные коэффициенты для оценки p_{ML} параметра p , полученной методом максимального правдоподобия. Умножение оценки p_{ML} на данный коэффициент устраняет ее смещение.

ТАБЛИЦА II.

p_{ML}	n	
	20	50
1.5	0.858	0.956
1.6	0.880	0.963
1.7	0.897	0.969
1.8	0.910	0.972
1.9	0.919	0.974
2.0	0.927	0.976
2.5	0.945	0.981
3.0	0.953	0.983
3.5	0.957	0.984
4.0	0.960	0.985
4.5	0.961	0.986
5.0	0.963	0.986
5.5	0.964	0.987

Аналогичные коэффициенты могут быть вычислены и для других методов, при этом следует иметь в виду, что для метода подбора следует заранее оговорить его вычислительные параметры (шаг сетки и ограничения на множества рассматриваемых значений параметров), так как результат применения метода будет зависеть, в том числе, и от них.

Метод подбора распределения может быть предложен и в случае большего числа неизвестных параметров (в частности, для симметричного TSP-

распределения $TSP(x_0, r, p)$ [5]), и его эффективность при этом продолжает оставаться достаточно высокой.

IV. ОЦЕНИВАНИЕ СРЕДНЕКВАДРАТИЧЕСКОГО ОТКЛОНЕНИЯ ПО ВЫБОРОЧНЫМ СКО И РАЗМАХУ

При оценивании неопределенности по типу А зачастую возникает задача об оценивании СКО результатов измерений. Как правило, при этом пользуются такими характеристиками разброса данных, как выборочное среднеквадратическое отклонение (СКО) и размах.

Несмещенная оценка дисперсии распределения для выборки длиной n имеет вид:

$$S^2 = (n - 1)^{-1} \sum (x_i - x_m)^2,$$

где x_m – среднее выборочное значение. При этом величина S в качестве оценки среднеквадратического отклонения является смещенной [6], вследствие операции взятия корня из S^2 . Рассмотрим оценку СКО σ симметричного распределения (1) вида $\sigma = k_S S$ и оценим значение корректирующего коэффициента k_S . В качестве альтернативы рассмотрим оценку СКО вида $\sigma = k_R R$, где R – выборочный размах. В табл. III ниже приведены значения данных коэффициентов, а также СКО соответствующих оценок σ .

ТАБЛИЦА III.

n	p	k_S	$V \geq \epsilon (k_S S)$	k_R	$V \geq \epsilon (k_R R)$
20	0.8	1.005	0.099	0.335	0.052
	1.0	1.006	0.108	0.319	0.069
	1.5	1.008	0.126	0.295	0.106
	2.0	1.010	0.140	0.282	0.133
	2.5	1.011	0.150	0.273	0.154
	3.0	1.013	0.159	0.268	0.171
	4.0	1.015	0.173	0.261	0.195
	5.0	1.016	0.182	0.257	0.211
	N	1.013	0.163	0.268	0.195
50	0.8	1.002	0.059	0.321	0.017
	1.0	1.002	0.065	0.300	0.028
	1.5	1.003	0.077	0.267	0.055
	2.0	1.004	0.085	0.248	0.077
	2.5	1.004	0.092	0.236	0.096
	3.0	1.005	0.098	0.227	0.110
	4.0	1.006	0.107	0.217	0.132
	5.0	1.006	0.113	0.210	0.148
	N	1.005	0.101	0.222	0.145

Заметим, что при значениях p , меньших 2 (напомним, значение $p = 2$ отвечает треугольному распределению), оценка СКО распределения с использованием выборочного размаха R является более эффективной, по сравнению с оценкой, полученной посредством выборочного СКО S (т. е. СКО соответствующей оценки меньше).

V. ОГРАНИЧЕНИЯ НА ЗНАЧЕНИЯ ПАРАМЕТРОВ TSP-РАСПРЕДЕЛЕНИЯ

При применении метода подбора наиболее подходящего распределения из семейства TSP, очевидно,

желательно ограничить число перебираемых наборов параметров. Данные ограничения можно ввести, основываясь на разнообразных выборочных статистиках.

А. Оценивание величины r для симметричного TSP-распределения по выборочному размаху

Пусть имеется выборка длины n из симметричного TSP-распределения. Рассмотрим величину $c_R = 2r/R$. Очевидно, что при неограниченном росте длины выборки n величина c_R стремится к единице. В табл. IV ниже приведены средние значения $c_{R,0}$ данной величины, а также вероятностно-симметричные интервалы охвата для уровня вероятности 0.95 (обозначим их $[c_{R,1}, c_{R,2}]$). При использовании метода выбора наиболее подходящего распределения приведенные данные могут быть использованы для ограничения множества различных значений-кандидатов параметра r , т. е. для каждого фиксированного p имеет смысл рассматривать для r отрезок значений $[c_{R,1} R, c_{R,2} R]$, а в качестве начального приближения при построении сетки можно порекомендовать выбор $c_{R,0}$.

ТАБЛИЦА IV.

n	20		50	
p	$c_{R,0}$	$[c_{R,1}, c_{R,2}]$	$c_{R,0}$	$[c_{R,1}, c_{R,2}]$
0.8	1.07	[1.00, 1.23]	1.02	[1.00, 1.07]
1.0	1.11	[1.01, 1.33]	1.04	[1.00, 1.12]
1.5	1.25	[1.05, 1.61]	1.12	[1.03, 1.27]
2.0	1.41	[1.12, 1.91]	1.22	[1.07, 1.45]
2.5	1.57	[1.19, 2.22]	1.33	[1.12, 1.64]
3.0	1.75	[1.27, 2.54]	1.46	[1.19, 1.84]
4.0	2.11	[1.45, 3.17]	1.71	[1.33, 2.24]
5.0	2.47	[1.64, 3.82]	1.97	[1.48, 2.65]

В. Оценивание параметра положения по выборочному среднему

Пусть снова имеется выборка длины n из симметричного TSP-распределения. Рассмотрим отношение отклонения среднего выборочного x_a от математического ожидания распределения x_0 , к выборочному СКО S и оценим критические значения c_S этого отклонения для уровня вероятности 0.95:

$$|x_a - x_0| / S \leq c_S.$$

Заметим, что данные критические значения достаточно слабо зависят от p для рассматриваемых значений n , что дает возможность использовать общее значение параметра c_S для разных p (и это весьма полезное свойство). Например, для $n = 20, 50$ они составляют 0.014 и 0.009, соответственно (рассматривали $0.5 \leq p \leq 10$).

С. Ограничения на множество возможных значений параметра p

Очевидные предположения о диапазоне возможных значений параметра p могут быть сделаны, исходя из

вида гистограммы выборки (при достаточной ее длине). Если гистограмма выглядит достаточно «плоской» или имеет достаточно высокие частоты вблизи крайних значений, то следует рассматривать значения p вблизи единицы (возможно, даже меньшие единицы). В случае распределения с явно выраженным пиком следует рассматривать значения p , большие единицы.

Для симметричного распределения более обоснованные ограничения на параметр p могут быть наложены на основе величины отношения R/S (выборочного размаха к выборочному СКО, зависит только от p). Были вычислены средние и критические значения распределения данной статистики. С ростом p эффективность данного метода падает.

VI. ЗАКЛЮЧЕНИЕ

Для обоснованного выбора закона распределения погрешностей при обработке результатов измерений необходимо использовать гибкий подход, основанный на применении широкого семейства распределений и учете как экспериментальных данных, так и априорной информации о характере погрешностей. Такой подход позволит повысить точность и достоверность результатов измерений. Рассмотренное в данной работе семейство TSP-распределений достаточно широко и, в то же время, имеет простое математическое описание, допускающее простую и эффективную имплементацию метода Монте-Карло.

Основываясь на проведенных исследованиях, авторами, в частности, продемонстрирована эффективность метода выбора распределения, на основе выборочных данных, по сравнению с иными методами оценки параметров распределения, которая позволяет рекомендовать его в качестве одного из основных методов оценки параметров TSP-распределения.

СПИСОК ЛИТЕРАТУРЫ

- [1] Kotz S, Van Dorp J.R. Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications. World Scientific Publishing, Singapore, 2004.
- [2] Kotz S, Van Dorp J.R. The Standard Two-Sided Power Distribution and its Properties // The American Statistician, 2002, 56 (2). P. 90-99.
- [3] Тырсин А.Н., Метод подбора наилучшего закона распределения непрерывной случайной величины на основе обратного отображения // Вестник ЮУрГУ, Серия Математика. Механика. Физика. 2017. Т. 9, вып. 1. С. 31-38.
- [4] Об одном методе подбора закона распределения из семейства TSP и его свойствах / Степанов А.В., Чуновкина А.Г. // Материалы международной Воронежской зимней математической школы, Воронеж, ВГУ, 2024. С. 244-246.
- [5] Об оценке параметров асимметричного TSP распределения и его применении / Степанов А.В., Чуновкина А.Г. // Вероятностные методы в дискретной математике: тезисы докладов XI международной Петрозаводской конференции, Петрозаводск, КарНЦ РАН, 2024. С. 106-108.
- [6] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М., МГУ, 1987.