

Информационный способ установления справедливости гипотез несимметричных распределений

В. Г. ПОЛОСИН, П. П. ПЕРШЕНКОВ

Пензенский государственный университет, Пенза, Россия, e-mail: polosin-v@sura.ru

Рассмотрен способ установления справедливости гипотез несимметричных распределений, основанный на их преобразовании к известным симметричным распределениям и анализе количества информации, содержащейся в результатах эксперимента.

Ключевые слова: статистическое моделирование, гипотетическая реализация, несимметричное распределение, область принятия решения.

The way of an establishment of hypotheses justice of asymmetrical allocations based on their transformation to known symmetric allocations and analysis of information quantity, which the results of experiment contain, is considered.

Key words: statistical modeling, hypothetical realization, asymmetrical distribution, acceptance decisions area.

Задачи статистического моделирования при проведении физических и технических исследований связаны с анализом закономерностей распределений результатов измерений, содержащих наиболее полную информацию о характере и природе предмета наблюдения. Форма принятой гипотетической модели распределения результатов, с одной стороны, указывает на происхождение влияющих факторов и, следовательно, причины разброса результатов, с другой — отражает внутреннее строение самого объекта.

К примеру, если случайная величина формируется под аддитивным воздействием большого числа независимых и соизмеримых факторов, то в качестве статистической модели аппроксимации результатов наиболее приемлемо нормальное распределение [1, 2]. В тех случаях, когда исследуемая величина принимает только положительные значения, при статистическом моделировании предпочтительно использовать несимметричные распределения. На практике в качестве сглаживающей функции несимметричной выборки результатов измерений достаточно часто применяют логарифмическое нормальное распределение, в котором по нормальному закону распределен логарифм отношения результатов случайных измерений x_i к медиане выборки всех полученных результатов $x_{0,5}$ [5]. Особенность образования логнормального распределения состоит в том, что разброс случайной величины образуется под мультипликативным воздействием множества независимых близких по значимости факторов, что приводит к увеличению вероятности появления больших значений регистрируемой величины.

При устойчивом неизменном воздействии, которое значительно превышает другие влияющие факторы, все регистрируемые значения будут превышать некоторое минимальное $x_{\min}$. В этом случае при установлении справедливости гипотез несимметричных распределений (УСГНР) для выборки результатов измерений наиболее приемлемые реализации достигаются, если используется распределение

максимального значения или смещенное трехпараметрическое распределение Вейбулла—Гнеденко. Характерная особенность этих распределений в том, что для них функция плотности распределений в области значений меньших некоторого минимального положительного $x_{\min}$, равна нулю. В силу своей универсальности в современной литературе для описания экспериментальных статистических данных распределение Вейбулла—Гнеденко находит широкое применение [2, 5]:

$$F(x) = 1 - \exp \{ - [(x - m)/a]^c \},$$

где a , c и m — параметры соответственно масштаба, формы и смещения.

Частные случаи этого распределения можно использовать вместо ряда других несимметричных [2, 3]; при $c = 1$ оно превращается в показательное; при $c = 2$ совпадает с распределением Рэлея; при $a > 2$, $m = 0$ хорошо аппроксимирует логарифмическое нормальное.

Правильный выбор гипотетической реализации сглаживающего распределения результатов позволяет установить характер и свойства влияющих на измерения факторов. Так как параметры сглаживающего распределения сильно зависят от метода их оценки, при аппроксимации одной и той же выборки результатов экспериментов образуется множество гипотетических моделей. Выбор наиболее предпочтительной из возможных моделей распределений требует сравнительного анализа статистических критериев.

Способ проверки статистических гипотез. Предлагаемый способ УСГНР основан на информационных свойствах статистических распределений, получивших широкое распространение при идентификации симметричных распределений [5]. Идея способа состоит в приведении несимметричного $F_n(x)$ к симметричному $F_c(z)$ распределению и после-

Преобразование сглаживающих несимметричных распределений в симметричные

Несимметричное распределение	Функция несимметричного распределения	Формула преобразования	Симметричное распределение	Функция симметричного распределения
Логарифмическое нормальное	$F(X) = \Phi[a^{-1} \ln(x/x_{0,5})]$	$z = \ln(x/x_{0,5})$	Нормальное распределение Гаусса—Лапласа	$F_c(z) = \Phi(z/a)$
Классическое двухпараметрическое Вейбулла—Гнеденко	$F(x) = 1 - \exp(-(x/a)^c)$	$y = \pm (x/a)^c$	Симметричное двухстороннее показательное распределение Лапласа	$F_c(z) = \begin{cases} 1 - 0,5e^{-z}, & z = y; \\ 0,5e^{-z}, & z = -y, \\ y > 0 \end{cases}$
Смещенное трехпараметрическое Вейбулла—Гнеденко	$F(x) = 1 - \exp\left\{-[(x-\bar{x})/a]^c\right\}$	$y = \pm [(x-\bar{x})/a]^c$		
Максимального значения	$F(x) = \exp[-e^{-(x-\mu)/\lambda}]$	$y = \pm e^{-(x-\mu)/\lambda}$		
Минимального значения	$F(x) = 1 - \exp[-e^{-(x-\mu)/\lambda}]$	$y = \pm e^{-(x-\mu)/\lambda}$		
Экспоненциальное распределение Грэди	$F(x) = 1 - \exp(-Mx/\mu)$	$y = \pm M_0 x/\mu$	Равномерное распределение	$F_c(z) = R$, где $0 < R < 1$
Произвольное непрерывное распределение	$F(x)$	$z = R = F(x)$		

дующей проверке соответствия полученной статистической выборки результатов гипотезе распределения $F_c(z)$, основанной на совпадении информационных коэффициентов (контрэксцесса χ_z и коэффициента энтропии k_3), рассчитанных для выборки результатов с соответствующими коэффициентами гипотетического симметричного распределения.

В табл. 1 приведены возможные преобразования $z = \xi(x)$ несимметричных сглаживающих распределений выборки результатов измерений в соответствующие симметричные. Если для аппроксимации случайных значений x используются такие распределения, как классическое двухпараметрическое или смещенное трехпараметрическое распределение Вейбулла—Гнеденко, распределения максимального или минимального значений, то допустимые преобразования позволяют сформировать выборку случайных значений y . При верном выборе сглаживающей зависимости $F(x)$ значения y распределены по экспоненциальному закону вида

$$F_0(y) = 1 - e^{-y}.$$

Симметричное отражение положительных значений выборки y относительно нуля на область отрицательных чисел формирует выборку случайных значений z , для которой сглаживающая зависимость примет вид симметричного двухстороннего показательного распределения Лапласа:

$$F(z) = \begin{cases} 1 - 0,5e^{-z} & \text{при } z = y; \\ 0,5e^{-z} & \text{при } z = -y. \end{cases}$$

Среди несимметричных следует выделить линейное экспоненциальное распределение Грэди [6, 7], ориентированное на описание статистики разрушения и широко используемое для задач фрагментации вселенной, задач фрактальной теории. Основанное на статистике Пуассона распределение Грэди так же можно преобразовать в симметричное распределение Лапласа.

Особое место занимает симметричное равномерное распределение. Дело в том, что если случайная величина x имеет непрерывную функцию распределения $F(x)$, то случайная величина $z = F(x)$ имеет равномерное распределение в интервале $0 \leq z \leq 1$ [1, 3]. Это свойство равномерного распределения позволяет выполнить УСГНР случайных значений x на основе соответствия распределения случайных значений z равномерному распределению.

Правомомерность выбора формы или отдельных параметров симметричного распределения случайной величины z , устанавливающих на основе соответствия оценок коэффициента энтропии k_3 , эксцесса ε и контрэксцесса χ_z , рассчитанных для выдвигаемой гипотезы H_1 , оценкам симметричного распределения, используемой в качестве нулевой гипотезы H_0 [5].

Пространство коэффициента энтропии и контрэксцесса. Коэффициент энтропии k_3 , равный отношению энтропийной погрешности к средней квадратической, содержит количественную оценку информативности результатов измерений. Эксцесс распределения ε , равный отношению четвертого центрального момента μ_4 к четвертой степени среднего квадратического отклонения выборки значений z , характеризует островершинность и протяженность спадов распределений [5]. При классификации распределений принято использовать безразмерную величину, называемую контрэксцессом распределения $\chi_z = 1/\sqrt{\varepsilon}$, значения которой находятся в пределах $0-1$, что удобно при построении иллюстраций. Для ряда симметричных распределений k_3 , χ_z имеют индивидуальные значения: для равномерного — 1,73 и 0,745; нормального — 2,066 и 0,577; распределения Лапласа — 1,92 и 0,408, соответственно.

При проведении любых экспериментов все результаты случайны и их количество n ограничено, что обуславливает отклонение оценок k_3 и χ_z гипотетических реализаций от их индивидуальных значений, характерных для нулевой гипотезы. Средние квадратические погрешности рассеяния их

Таблица 2

Параметры симметричных распределений

Распределение	Параметры распределений								
	ε	χ_z	k_3	$S(k_3)$	$S(\chi_z)$	$\Delta k_{0,9}$	$\Delta \chi_{0,9}$	$\Delta k_{0,95}$	$\Delta \chi_{0,95}$
Нормальное	3	0,577	2,066	0,017	0,016	0,027	0,026	0,041	0,04
Лапласа	6	0,408	1,92	0,026	0,034	0,042	0,055	0,064	0,085
Равномерное	1,8	0,745	1,73	0,017	0,008	0,027	0,013	0,041	0,02

оценки можно приближенно аппроксимировать выражениями вида [5]:

$$S(k_3) = 0,9(\chi_z k_3 \sqrt{k_3 n})^{-1}; \quad S(\chi_z) = \chi^4 \sqrt{(\varepsilon^2 - 1)^3} / \sqrt{29n}. \quad (1)$$

Разбросы Δk_3 , $\Delta \chi_z$ оценок k_3 , χ_z зависят от выбора доверительной вероятности оценки, числа n измерений в выборке и пропорциональны их средним квадратическим погрешностям

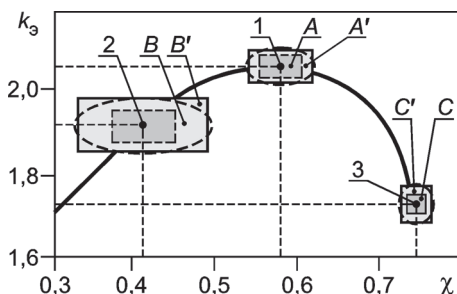
$$\Delta k_3 = t_P S(k_3); \quad \Delta \chi_z = t_P S(\chi_z), \quad (2)$$

где t_P — коэффициент, определяемый доверительной вероятностью P .

Так как эксцесс, контрэксцесс и коэффициент энтропии однозначны для симметричных распределений, соответствующие им разбросы оценок полностью определены выборкой результатов измерений и коэффициентом t_P , что позволяет выделить область принятия гипотезы в пространстве k_3 , χ_z .

Для иллюстрации рассматриваемого материала на рисунке приведена топографическая диаграмма распределений, где точками 1, 2 и 3 отображены положения распределений, соответственно нормального, симметричного показателя Лапласа и равномерного. В непосредственной близости от этих точек выделены области А, В, С принятия гипотез распределений, ограниченные разбросами Δk_3 , $\Delta \chi_z$.

Используемые в работе разбросы оценок рассчитаны согласно (1), (2) для выборки из 1000 значений при доверительной вероятности $P = 0,9$, что позволило выбрать $t_P = 1,6$ независимо от вида гипотезы. Следует отметить, что в этом случае из всех возможных выборок распределений, соответствующих нулевой гипотезе H_0 , 81 % находится в области



Топографическая диаграмма для симметричных распределений: 1—3 — соответственно нормального, симметричного показателя Лапласа, равномерного; А, В, С — области принятия гипотез распределений; А', В', С' — расширенные области

принятия гипотезы и всего 19 % за ее пределами. Результаты расчета основных параметров для симметричных распределений приведены в табл. 2.

Из построенной топографической диаграммы (см. рисунок) следует, что площади областей принятия гипотезы А, В, С при одинаковых доверительной вероятности и количестве результатов уменьшаются с увеличением k_3 . Если для вычисления границ областей принятия гипотезы использовать коэффициент $t_P = 2,33$, что соответствует квантили нормированной функции Лапласа при $P = 0,98$, то

внутри границ соответствующих областей будут попадать более 96 % результатов вычисления оценок k_3 , χ_z для всех выборок, содержащих 1000 результатов измерений. Расширение границ областей принятия гипотез иллюстрируют дополнительные области, обозначенные на топографической диаграмме соответственно А', В' и С'.

Таким образом, если точка, отображающая результаты оценок k_3 , χ_z для исследуемой выборки значений, находится вне пределов области принятия гипотез А', В', С', то выбранную гипотезу распределения следует признать ошибочной. Попадание отображающей точки внутрь областей А, В, С позволяет рассматривать принятую гипотезу в качестве возможной реализации распределения для выборки значений z . В этой области вероятность принадлежности выборки значений какой-либо другой гипотезе распределения будет стремиться к нулю.

Из топографической диаграммы также следует, что вероятность попадания оценок k_3 , χ_z исследуемой выборки в область принятия нулевой гипотезы зависит от площади выделенной области. Для прямоугольной области удельная плотность вероятности распределена неравномерно и имеет минимальное значение в точках, наиболее удаленных от положения нулевой гипотезы. Очевидно, что прямоугольные границы областей можно заменить эллиптическими, что позволяет на границе области допустимых значений гипотезы обеспечить одинаковую функцию плотности вероятности нахождения генерируемых выборок.

Литература

1. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. М.: АСТ, 2003.
2. Харт Х. Введение в измерительную технику. М.: Мир, 1999.
3. Вадзинский Р. Н. Справочник по вероятностным распределениям. СПб.: Наука, 2001.
4. Новицкий П. В., Зограф И. А. Оценка погрешностей результатов измерений. Л.: Энергоатомиздат, 1985.
5. Петрович М. Л., Давидович М. И. Статистическое оценивание и проверка гипотез на ЭВМ. М.: Финансы и статистика, 1989.
6. Grady D. E., Kipp M. E. Geometric statistics and dynamic fragmentation // J. Appl. Phys. 1985. V. 58. N 3. P. 1210—1222.
7. Grady D. E. Particle size statistics in dynamic fragmentation // J. Appl. Phys. 1990. V. 68. N 12. P. 6099—6105.

Дата принятия 04.10.2013 г.