УДК 539.16+519.224

Метод идентификации форм распределений малых выборок

М. В. Федоров

МАКСИМ ВАЛЕРИЕВИЧ ФЕДОРОВ — младший научный сотрудник лаборатории физической биохимии Института теоретической и экспериментальной биофизики РАН (ИТЭБ РАН). Область научных интересов: статистическая физика, закономерности в реальных стохастических процессах, непараметрическая статистика, вейвлет-анализ и его приложения.

142290 Пущино, Московская область, Институтская ул., д. 3, ИТЭБ РАН,
E-mail max@pbc.iteb.serpukhov.su

Предложен метод непараметрической гистограммной оценки формы выборочного распределения, основанный на вычислении оптимальных параметров сдвига и масштаба путем поиска экстремума ACF-функционала. Показано, что предлагаемый метод более эффективен в случае небольшого объема выборки, чем различные неоптимизированные гистограммные оценки плотности.

С помощью этой оценки сделано обобщение метода топографической классификации форм распределений на случай малого объема выборки. Метод позволил обнаружить квазипериодические колебания характеристик тонкой структуры распределений результатов измерений α-активности образцов ^{239}Pu.

Введение

Как было показано в работе [1], тонкая структура распределения амплитуд флуктуаций результатов измерений процессов разной природы неслучайна. В связи с этим актуальна задача идентификации и сравнения форм соответствующих гистограмм. Традиционно применяемые методы анализа распределений оказались для этой цели мало пригодными [1]. Однако прогресс в решении этой задачи может быть достигнут при сочетании нескольких статистических критериев.

Предлагаемый метод основан на классификации выборочных распределений по оценкам двух параметров — коэффициента контрэксцесса χ и энтропийного коэффициента K_ε [2—4]. Каждому выборочному распределению в этом случае соответствует определенная точка в (χ, K_ε)-плоскости. Для классификации форм гистограмм оказалось достаточным разбиение этой плоскости на девять областей. Типичные «базовые» формы гистограмм, соответствующие разным областям, показаны на рис. 1. Решение о сходстве (различии) гистограмм определяется в соответствии с принадлежностью их к тем или иным базовым формам. Для классификации форм гистограмм существенна предлагаемая нами оптимизированная оценка параметра K_ε. Такая оценка более эффективна в случае малых выборок, чем обычно применяемые [2—4].

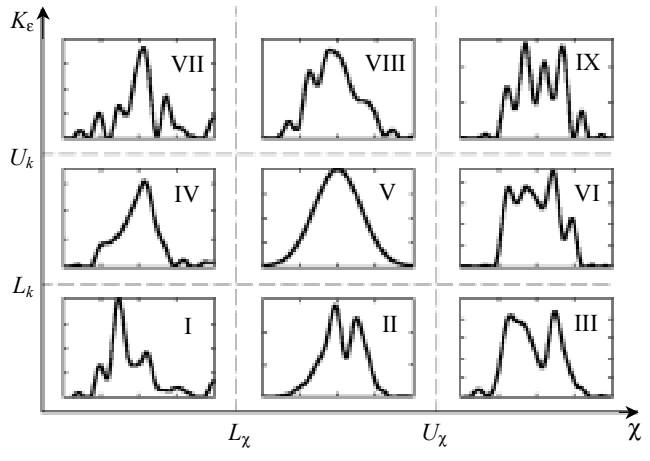

Рис. 1. Зависимость сглаженных форм выборочных распределений из пуассоновского процесса от их расположения на (χ, K_ε^{opt})-плоскости.
Пояснения см. в тексте.

Определение контрэксцесса

Контрэксцесс определяется как [3]:

$$\chi = 1/\sqrt{\eta} \qquad (1)$$

где η — параметр эксцесса, определяемый как

$$\eta = \mu_4/\sigma^4 \qquad (2)$$

где σ — среднеквадратичное отклонение; μ_4 — выборочный четвертый центральный момент [3, 5, 6].

Параметр χ изменяется от 0 (распределение Коши) до 1 (дискретное двузначное распределение). Для нормального распределения $\chi = 1/\sqrt{3} \approx 0{,}577$.

Определение энтропийного коэффициента

Вторым параметром, характеризующим форму распределения, является энтропийный коэффициент K_ε [2—4]

$$K_\varepsilon = \Delta_\varepsilon/\sigma \qquad (3)$$

где Δ_ε — так называемое энтропийное значение погрешности [2—4]

$$\Delta_\varepsilon = \frac{1}{2} e^{H(X/X_п)} \qquad (4)$$

$H(X/X_п)$ — изменение энтропии (неопределенности), вносимое измерением случайной величины X. Эта величина определяется как [3, 7]:

$$H(X/X_п) = -\int_{+\infty}^{-\infty} p(X)\ln(p(X))dX \qquad (5)$$

где $p(X)$ — плотность вероятности случайной величины.

Так как интеграл (5) определен для любых распределений, следовательно, и величину энтропийного значения погрешности можно определить для любого закона распределения случайной величины.

Было показано [7], что максимальным значением $K_e = \sqrt{2\pi e}/2 \approx 2{,}066$ обладает нормальное распределение. Для распределения Коши и дискретного двузначного распределения $K_\varepsilon = 0$, для равномерного распределения $K_\varepsilon = \sqrt{3} \approx 1{,}73$.

Для расчета энтропийного коэффициента по конечной выборке необходимо сделать оценку функции плотности вероятности наблюдаемой случайной величины. Если этой оценкой служит гистограмма (кусочно-постоянная функция на m интервалах группировки, каждый из которых имеет ширину w), то выборочный энтропийный коэффициент по гистограмме определяется [3, 4] как

$$K_\varepsilon = \frac{wN}{2\sigma} 10^{-\frac{1}{N}\sum_{i=1}^{m} n_i \lg(n_i)} \qquad (6)$$

где w — размер интервала группировки (размер бина гистограммы); N — объем выборки; m — число интервалов группировки (число бинов гистограммы); σ — выборочное стандартное отклонение; n_i — число измерений, попавших в i-й интервал группировки соответствующей гистограммы.

В различных руководствах по статистической обработке результатов измерений [2—6, 8, 9] приводятся различные рекомендации по выбору как числа интервалов группировки m, так и их размера w. Одной из наиболее распространенных оценок оптимального числа интервалов является [3, 5, 6, 8]

$$m = \sqrt{N} \qquad (7)$$

В работе [3] рекомендуется выбирать число интервалов m из диапазона:

$$0{,}55N^{0,4} < m < 1{,}25N^{0,4} \qquad (8)$$

При этом m должно быть нечетным [3].

В работах [10, 11] для распределений, близких по форме к нормальному, рекомендуется выбирать w как

$$w = 3{,}5\sigma N^{-1/3} \qquad (9)$$

В большинстве работ ничего не говорится о выборе начального значения первого интервала группировки — параметра сдвига s. Неопределенность в выборе этого параметра может привести к значительному смещению оценок K_ε. В работе [3] рекомендуют выбирать s таким образом, чтобы середина центрального интервала гистограммы совпадала с рассчитанным центром распределения. Основание для этих рекомендаций вызывает сомнения, поскольку s начинает в этом случае зависеть от флуктуирующего параметра центра распределения и неопределенным образом задаваемого числа интервалов. По-видимому, для каждой конкретной выборки существуют оптимальные значения параметров w и s (см., например, [8, 9]). Следовательно, можно попытаться найти некоторый функционал $F(w,s)$, экстремальному значению которого будут соответствовать оптимальные значения w и s. В работах [10, 11] обсуждаются проблемы, связанные с выбором такого функционала. Исходя из результатов этих работ и проведенного нами исследования, мы используем в качестве $F(w,s)$ следующую функциональную зависимость:

$$\mathrm{ACF}(w, s) = \frac{\sum_{i=0}^{m} [n_{w,s}(i) n_{w,s}(i+1)]}{\sum_{i=1}^{m} [n_{w,s}(i)^2]} \qquad (10)$$

где $n_{w,s}(i)$ — число результатов измерений, попавших в i-й интервал группировки гистограммы.

Оптимальным значениям w и s соответствует $\max(\mathrm{ACF}(w,s))$. В случае дискретных распределений расчет w и s следует проводить следующим образом: для каждой конкретной выборки задаются некоторые начальные значения w и s, в соответствии с формулами (7)—(9). После чего производится поиск максимума функционала $\mathrm{ACF}(w,s)$ на двумерной сетке (w,s) с шагом, равным единице. Далее по соответствующим максимуму значениям (w_{opt}, s_{opt}) строится выборочная гистограмма и по формуле (6) рассчитывается $K_\varepsilon^{\mathrm{opt}}$.

Классификация форм выборочных распределений по их расположению в $(\chi, K_\varepsilon^{\mathrm{opt}})$-плоскости

Для анализа поведения во времени форм выборочных распределений различных стохастических процессов нами предлагается следующая процедура.

1) Выбирается размер исследуемых выборок N. Это число не должно быть слишком большим, чтобы не нивелировались особенности каждой конкретной реализации исследуемого случайного процесса; N следует выбирать из диапазона 30—200.

2) Исходный ряд измерений делится на последовательные неперекрывающиеся отрезки размером N.

3) Для каждого из этих отрезков рассчитываются параметры χ и $K_\varepsilon^{\mathrm{opt}}$.

4) По всем полученным значениям χ и $K_\varepsilon^{\mathrm{opt}}$ рассчитываются доверительные интервалы $[L_\chi, U_\chi]$ и $[L_k, U_k]$ для каждого параметра отдельно. Как правило верхними и нижними границами служат соответственно 10 и 90%-е процентили.

5) В соответствии с расположением в $(\chi, K_\varepsilon^{\mathrm{opt}})$-плоскости и выбранными границами каждой из последовательных выборок присваивается индекс от I до IX. Так, выборке с $\chi < L_\chi$ и $K_\varepsilon^{\mathrm{opt}} < U_k$ соответствует индекс I, выборке с $L_\chi < \chi < U_\chi$ — индекс II и т.д. Рис. 1 иллюстрирует приблизительное соответствие сглаженных форм выборочных распределений из пуассоновского процесса с их расположением на $(\chi, K_\varepsilon^{\mathrm{opt}})$-плоскости.

Эта процедура определения форм распределений малых выборок не вполне однозначна, поэтому фигуры на рис. 1 не являются «эталонными формами», скорее это «типичные представители».

В работах [3, 4] предлагается иная классификация форм распределений по их положению в $(\chi, K_\varepsilon^{\mathrm{opt}})$-плоскости с использованием параметрически задаваемых областей для каждого из известных законов распределений и(или) их комбинаций. Но при небольших объемах выборок эти параметры определяются с большими погрешностями, поэтому мы пользуемся для расчета границ непараметрическими оценками в виде про-

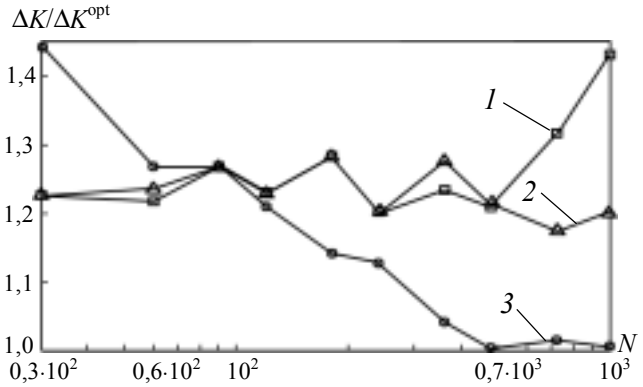

Рис. 2. Зависимость $\Delta K_\varepsilon/\Delta K_\varepsilon^{opt}$ от числа измерений *N* для различных методов расчета параметров *m*, *w* и *s*:

1—3 — расчет по формулам (7), (8) и (9), соответственно

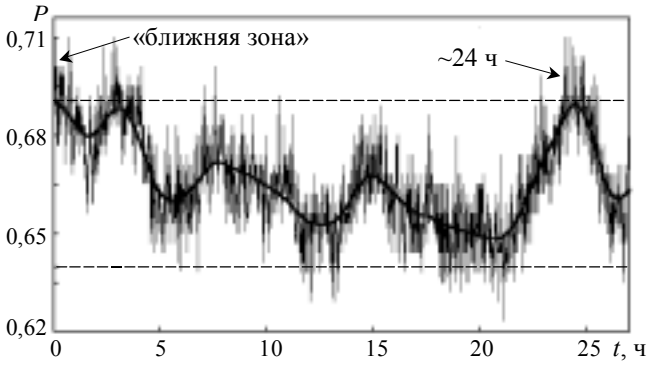

Рис. 3. Зависимость вероятности сходства гистограмм от временного интервала между ними

центилей. Тем не менее вопрос об оптимальном разбиении (χ, K_ε^{opt})-плоскости требует дальнейшего дополнительного исследования.

Оценка эффективности предлагаемого метода

Для проверки эффективности предлагаемого метода оценки энтропийного коэффициента K_ε^{opt} был проведен численный эксперимент. Из большой совокупности ($>10^6$) односекундных измерений α-радиоактивности, подчиняющейся закону Пуассона (среднее 317), случайным образом выбиралось по 1000 совокупностей объемом *N*. *N* изменяли от 30 до 1000. После чего для каждой выборки рассчитывали параметр K_ε различными методами и K_ε^{opt}. Затем для каждого *N* и для каждого метода было рассчитано расстояние между 75 и 25%-й квантилью как мера разброса ΔK_ε. На рис. 2 изображено отношение $\Delta K_\varepsilon/\Delta K_\varepsilon^{opt}$ для различных методов расчета параметров *m*, *w* и *s*. Как видно из рис. 2, разброс значений K_ε^{opt} во всем диапазоне изменения *N* меньше, чем разброс значений K_ε, вычисленных другими методами (ни одна из линий не пересекает уровень, равный единице). Следовательно, метод расчета K_ε^{opt} является наилучшим во всем диапазоне *N* от 30 до 1000. В то же время оценки по формулам (7) и (8) характеризуются большими смещениями.

Исследование изменения форм выборочных распределений во времени

Иллюстрацией возможностей разработанного метода является представленная на рис. 3 зависимость вероятности сходства (*P*) гистограмм от временного интервала между ними для всех базовых форм гистограмм, построенных по ряду измерений α-активности ^{239}Pu. Как видно, это распределение отличается от случайного, доверительные границы для которого показаны пунктиром. На рис. 3 хорошо видно, что повышенная вероятность сходства наблюдается у гистограмм с небольшим временным интервалом между ними (эффект «ближней зоны», см.[1]) и с приблизительно 24-часовым интервалом (эффект «околосуточной периодичности», см.[1]). Присутствует еще несколько экстремумов, выраженных менее ярко.

Следует отметить, что гистограммы, принадлежащие разным «базовым формам», характеризуются различной зависимостью частоты встречаемости от времени.

Таким образом, предложенный метод позволяет идентифицировать и сравнивать тонкую структуру распределений – формы соответствующих гистограмм более эффективно, чем ранее существующие. Нам представляется вероятным, что различные эффекты макрофлуктуаций, такие как «ближняя зона», «местное время», «околосуточная периодичность» и др. [1], обуславливаются гистограммами различных классов. Мы полагаем, что усовершенствованные методы классификации форм распределений, связанные с представлением гистограмм в (χ, K_ε^{opt})-плоскости, будут полезны в исследованиях различных процессов.

* * *

Представленная работа частично поддержана РФФИ (гранты № 01-04-97032 и № 01-03-32529).

Автор признателен С.Э. Шнолю и М.Н. Кондрашовой за постоянную поддержку, Э.С. Горшкову, С.В. Шаповалову и К.И. Зенченко за помощь в проведении экспериментов, В.Н. Морозову, К.И. Зенченко, Т.А. Зенченко, А.А. Конрадову, В.А. Коломбету и В.В. Стрелкову за полезное обсуждение.

ЛИТЕРАТУРА

1. *Шноль С.Э.* Рос. хим. ж. (Ж. Рос. хим. об-ва им. Д. И. Менделеева), 2002, № 3, с 3.
2. *Новицкий П.В.* Измерительная техника, 1966, № 7, с. 11—14.
3. *Новицкий П.В., Зограф И.А.* Оценка погрешностей результатов измерений. Л.: Энергоатомиздат, 1991.
4. *Алексеева И.У.* Автореф. дис. ... канд. тех. наук. Л.: Ленингр. политехн. ин-т, 1975, 20 с.
5. *Чернецкий В.И.* Математическое моделирование стохастических систем. Петрозаводск: Изд-во Петрозаводск. гос. ун-та, 1994.
6. *Кендалл М., Стьюарт А.* Теория распределений. М.: Наука, 1966.
7. *Шеннон К.* Работы по теории информации и кибернетике. М.: Издатинлит, 1963.
8. *Хальд А.* Математическая статистика с техническими приложениями. М.: Издатинлит, 1956.
9. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.
10. *Renaud O.* Technical Report 2000-34, Statistics Department, Stanford University, 2000.
11. *Renaud O.* Biometrika, 2002, v. 89, № 1, p. 129—143.