

Идентификация распределения плотности вероятности случайной величины

О. А. Микус¹, Э. И. Цветков²

Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В. И. Ульянова (Ленина)

¹sadalphamiks@gmail.com ²er-cvetkov@mail.ru

Аннотация. Исследуется зависимость точности идентификации распределения плотности вероятности случайной величины от числа разрядов (интервалов) гистограммы и объема используемой выборки. Оцениваются смещение и среднее квадратическое отклонение погрешности с использованием имитационного моделирования. Количественные результаты получены для равномерного распределения вероятности и распределения Симпсона. Показано, что с ростом интервала гистограммы погрешность смещения монотонно возрастает, а СКО погрешности имеет экстремум

Ключевые слова: идентификация; случайная величина; систематическая погрешность; плотность распределения вероятности; метрологический анализ; достоверность; гистограмма

Метрическая идентификация зависимостей [1] занимает промежуточное положение между измерениями величин и идентификацией объектов и отношений и базируется на случайной величине λ , свойства которой представляются числовыми и функциональными вероятностными характеристиками, определяемыми на множестве ее возможных значений Λ , т. е. случайная величина $\lambda \in \Lambda$.

При идентификации зависимости результат y , зависящий от аргумента x , представляется в виде однозначной функции, позволяющей установить ее значение при любом возможном значении x_i , что формализовано представляется следующим кортежем:

$$x = x_i \rightarrow y(x) = y(x_i).$$

Задача идентификации зависимости $y(x)$ имеет два варианта постановки:

1. формирование с помощью результатов совместных измерений $\{x_{ij}^*, y_{ij}^*\}_{i=1}^I$ функции $y^*(x)$ в виде, позволяющем устанавливать значение y по известному значению x ;
2. установление принадлежности указанной зависимости к определенному классу, определяемому соответствующими свойствами.

Идентификация вида сигналов по результатам совместных измерений аргумента и функции

$\left(\{x_{ij}^*, y_{ij}^*\}_{i=1}^I\right)$ – аппроксимация зависимости на интервале, содержащем указанные отсчеты. Область существования ограничивается интервалом $[x_{\min}, x_{\max}]$ и считается априори известной.

Процесс идентификации зависимости предполагает решение проблемы, связанной с выбором вида аппроксимирующей функции $f(\cdot)$ и последующим необходимым метрологическим анализом получаемых результатов.

В качестве примера полагается рассмотреть процедуру принятия решения о соответствии измерительных средств требованиям [2], где для составляющих погрешность необходима информация об их распределении. Первоначально необходимо определить какой метод аппроксимации использовать для зависимостей, полученных эмпирическим путем. Использование сплайнов предполагает проведение аппроксимирующей функции через отсчеты; аппроксимирующая функция принятого вида с установлением значений параметров, которые ее характеризуют, а также обеспечивают экстремум принятого критерия отличия значений аппроксимирующей функции от полученных отсчетов. При этом выбирается первый метод, если априорная информация о возможной форме практически отсутствует, и второй, если имеется информация о форме и требуется лишь конкретизировать ее вид.

Результатом идентификации зависимости является оценка, основанная на области возможных значений аргумента и известных закономерностях:

$$y_j^*(x) = f\left(\{x_{ij}^*, y_{ij}^*\}_{i=1}^I, x\right),$$

После получения оценки следует анализ полученных результатов, который заключается в исследовании характеристик погрешности

$$\Delta y_j^*(x) = y_j^*(x) - y_j(x),$$

где $\Delta y_j^*(x)$ – оценка погрешности, которая также зависит от каждого значения x_i в j -ом эксперименте.

Появление погрешности результата идентификации зависимости $y_j(x)$ определяется двумя факторами:

- отличии формы аппроксимирующей функции от формы аппроксимируемой
- погрешностями результатов совместных измерений.

Соответственно полная погрешность может быть представлена в виде суммы двух компонент $\Delta y_j^*(x) = \Delta_f y_j^*(x) + \Delta_\Delta y_j^*(x)$,

где $\Delta_f y_j^*(x) = f(\{x_{ij}, y_{ij}\}_{i=1}^I, x) - y_j(x)$ – компонента, обусловленная отличием аппроксимирующей функции $f(\cdot)$ от аппроксимируемой;

$\Delta_\Delta y_j^*(x) = f(\{x_{ij}^*, y_{ij}^*\}_{i=1}^I, x) - f(\{x_{ij}, y_{ij}\}_{i=1}^I, x)$ – компонента, обусловленная погрешностями результатов совместных измерений.

Уравнение идентификации многомерной функции в операторной форме можно представить следующим образом:

$$y_j^*\left(\{x_r\}_{r=1}^p\right) = R_f \left(\left\{ \left\{ R_{r mx} \dots R_{r lx} \gamma_{r ijx}(t) \right\}_{r=1}^p, \left\{ R_{my} \dots R_{ly} \gamma_{ijy}(t) \right\}_{i=1}^I, \{x_r\}_{r=1}^p \right), \right)$$

где R_f – оператор, представляющий преобразование результатов совместных измерений $\left\{ \left\{ x_{ri}^* \right\}_{r=1}^p, y_i^* \right\}_{i=1}^I$ и аргумента x в результат идентификации; R_{rx} и R_{sy} – операторы, преобразований, выполняемых при измерениях x_r и y соответственно; $\gamma_{rijx}(t)$ и $\gamma_{ijy}(t)$ – входные воздействия при выполнении i -х совместных измерений x_r и y соответственно.

Погрешность результата метрической идентификации определяется как зависимость разности между результатом идентификации и истинным значением функции $y(x)$, т. е.

$$\begin{aligned} \Delta y_j^*\left(\{x_r\}_{r=1}^p\right) &= y_j^*\left(\{x_r\}_{r=1}^p\right) - y\left(\{x_r\}_{r=1}^p\right) = \\ &= R_f \left(\left\{ \left\{ R_{r mx} \dots R_{r lx} \gamma_{r ijx}(t) \right\}_{r=1}^p, \left\{ R_{my} \dots R_{ly} \gamma_{ijy}(t) \right\}_{i=1}^I, \{x_r\}_{r=1}^p \right) - \right. \\ &\left. - R_f^\Gamma \left(\left\{ \left\{ R_{r mx}^\Gamma \dots R_{r lx}^\Gamma \gamma_{r ijx}(t) \right\}_{r=1}^p, \left\{ R_{my}^\Gamma \dots R_{ly}^\Gamma \gamma_{ijy}(t) \right\}_{i=1}^I, \{x_r\}_{r=1}^p \right) \right). \end{aligned}$$

Данные определения показывают, что $\Delta y_j^*\left(\{x_r\}_{r=1}^p\right)$ обусловлена двумя факторами – погрешностями результатов измерений $\{x_r\}_{r=1}^p$ и y , а также погрешностями аппроксимации, т.е. видом принятой $f(\cdot)$.

Идентификация плотности распределения вероятности случайной величины обобщенно включает в себя установление вероятностной характеристики $\Psi(\lambda) = \{ \{ (\lambda_{ij}^* \in \lambda_{s-1}, \lambda_s) = 1, \{ (\lambda_{ij}^* \notin \lambda_{s-1}, \lambda_s) = 1, \}_{i=1}^I \}_{s=1}^S$ для каждого s -го интервала существования случайной величины и установление оценки плотности распределения $w_j^*(\lambda)$.

Соответствующая последовательность отображений имеет вид:

$$\begin{aligned} &\{ \lambda_{ij}^* \}_{i=1}^I \rightarrow \\ &\{ \{ (\lambda_{ij}^* \in \lambda_{s-1}, \lambda_s) = 1, \{ (\lambda_{ij}^* \notin \lambda_{s-1}, \lambda_s) = 1, \}_{i=1}^I \}_{s=1}^S \\ &w_j^*(\lambda) = \{ \{ n_s \Delta_s^{-1} I_s^{-1} / (\lambda_{ij}^* \in \lambda_{s-1}, \lambda_s) \}_{i=1}^I \}_{s=1}^S \end{aligned}$$

Здесь введены следующие обозначения: I – число используемых отсчетов в j -ом идентификационном эксперименте; S – число интервалов; $\Delta_s = \lambda_s - \lambda_{s-1}$ – протяженность i -го интервала; I_s – число отсчетов, для которых справедливо $\lambda_{ij}^* \in (\lambda_{s-1}, \lambda_s)$; n_s – число отсчетов, попавших в s -й интервал.

В качестве примера рассматривается возможность идентификации для задачи бинарного сличения плотности распределения вероятности систематической составляющей погрешности, которая указывается как распределенная по закону Симпсона.

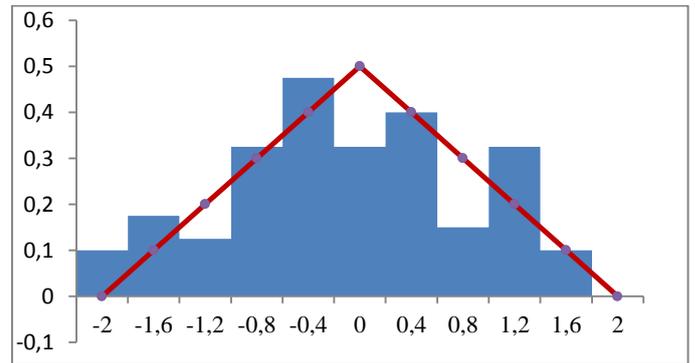


Рис. 1. Экспериментальная и теоретическая плотности распределения вероятности при $N=100$

При планировании метрической идентификации распределения плотности вероятности случайной величины $w(\lambda)$ устанавливаются параметры I, S, N . В общем случае необходимо выбрать критерий, по которому будет определяться точность.

Оценка достоверности [4] получаемых при метрологическом анализе вероятностных характеристик точности идентификации функциональной вероятностной характеристики требует дополнительного усреднения. Это является особенностью метрической идентификации для вероятностных характеристик случайных величин, т.к. их определение предполагает выполнение усреднения по совокупности отсчетов. Указанным особенностям метрологического анализа процедур оценивания вероятностных характеристик не уделялось много внимания, а в то же время практическая потребность в использовании соответствующей информации возрастает. В [5] приводятся варианты оценки достоверности полученных результатов метрологического анализа через ошибки определения, отнесенные к математическому ожиданию, дисперсии или вероятностной характеристике по принадлежности требуемому интервалу результата измерения.

При использовании в качестве критерия точности дисперсии

$$D[\Delta w_j^*(\lambda)] = \frac{1}{N} \sum_{j=1}^N \frac{1}{N_s} \sum_{s=1}^S (w_{js}^*(\lambda_j) - w_s(\lambda))^2,$$

где N – число оценок, используемых при определении

$$D[\Delta w_j^*(\lambda)] \cdot N_s = N \cdot \int_{\lambda_{s-1}}^{\lambda_s} w(\lambda) d\lambda.$$

Результаты, полученные при метрической идентификации зависимости, позволяют определить погрешность идентификации:

$$\Delta w_j^*(\lambda) = w_j^*(\lambda) - w(\lambda)$$

и исследовать ее зависимость от параметров $I, S, \{\Delta_s\}_{s=1}^S$.

Первоначально проходит исследование дисперсии ошибки определения $w_j(\lambda)$ от ширины интервала Δ_s для трех равнозначных выборок в 1000 значений каждая.

Можно видеть, что при увеличении ширины интервала (уменьшении количества интервалов s) дисперсия ошибки возрастает, также появляется различие между абсолютными значениями дисперсии ошибки определения плотности распределения вероятности, что также связано с увеличением ширины интервала Δ_s .

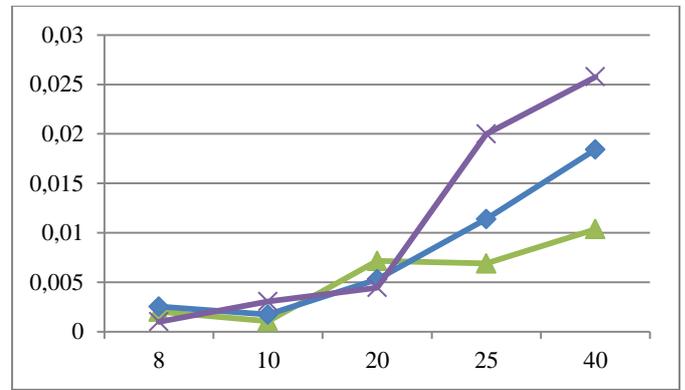


Рис. 2. Величина дисперсии $w_j(\lambda)$ от количества интервалов гистограммы плотности распределения вероятности

Интервал существования случайной величины $[-2;2]$ при делении на $S=10$ составляет ширину $\Delta_s = 0,2$; для $S=8$ $\Delta_s = 0,25$, а при $S=20$ $\Delta_s = 0,1$. Ниже указывается, как объем выборки влияет на $D[\Delta w_j^*(\lambda)]$.

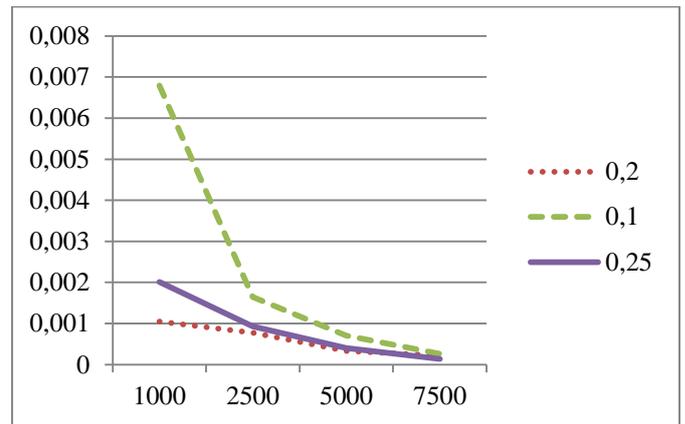


Рис. 3. Влияние объема выборки на величину дисперсии $w_j(\lambda)$ для различных интервалов

Графическое отображение данных полученных при имитационном моделировании позволяет увидеть убывающую зависимость, т.е. уменьшение дисперсии ошибки при увеличении объема выборки значений.

Как можно видеть в рамках нескольких выборок дисперсия при увеличении количества интервалов увеличивается; при увеличении количества значений в рамках одной выборки – наоборот уменьшается.

Появление погрешности также объясняется принятым способом округления и используемым видом аппроксимации. Достоверность результатов определяется неадекватностью используемых математических моделей, конечностью объема используемой при формировании оценки, а также неидеальностью выполняемых преобразований, проявляющихся с учетом особенностей процедур идентификации.

Применение подобного рода метрической идентификации целесообразно в отсутствии априорных знаний о плотности распределения вероятности, которые необходимы для проведения метрологического анализа результатов измерений и имеют специфику, заключающуюся в необходимости учитывать погрешности результатов совместных измерений аргумента и функции, а также погрешности аппроксимации.

СПИСОК ЛИТЕРАТУРЫ

- [1] Цветков Э.И. Установление множества возможных алгоритмов как этап автоматического метрологического синтеза // Мир измерений. 2017. № 2. С. 36-42.
- [2] Suloeva E.S., Tsvetkov E.I., Rzieva M.T. Decisions Based on the Results of Comparison of Standard // Measurement Techniques October 2014, Volume 57, № 7, pp 733-739.
- [3] Сулоева Е.С. Возможности бинарного сличения при наличии сведений о систематических погрешностях // Известия СПбГЭТУ «ЛЭТИ», 2015, №2. С 51-54.
- [4] Tsvetkov E.R., Suloeva E.S. Applying simulation using for the formation of decision rules based on binary comparison results //: Proceedings of the 2016 IEEE North West Russia Section Young Researchers in Electrical and Electronic Engineering Conference, EIConRusNW 2016
- [5] Tsvetkov E.I., Suloeva E.S. Analysis of the parameters that determine the reliability of the results of a verification of measuring instruments // Measurement techniques. V. 61 № 9, 2018. P. 872-877