

Модели распределений плотности вероятности при расчете неопределенности измерений

Н. А. РУБИЧЕВ

Московский государственный университет путей сообщения (МИИТ),
Москва, Россия, e-mail: vechibur@mail.ru

Обсуждаются модели распределений плотности вероятности при пересчете стандартной неопределенности измерений в расширенную неопределенность, а также модели распределений входных величин, используемые при расчете их стандартной неопределенности. Даются оценки сверху для используемых в обеих задачах коэффициентов при отсутствии информации о виде распределений.

Ключевые слова: неопределенность измерений, входные величины, распределение плотности вероятности.

The models of distributions during the standard measurements uncertainty recalculation into the expanded uncertainty as well as the models input quantities distributions used for the calculation of their standard uncertainty are considered. The upper estimates for both direct conversion tasks at the absence of information on the distributions form are given.

Key words: measurements uncertainty, input quantities, probability density distribution.

В соответствии с [1, 2] «стандартная неопределенность u определена как неопределенность результата измерений, выраженная в виде среднего квадратического отклонения (СКО) S ». (Здесь и далее в кавычки взят текст, заимствованный из [1, 2].) Аналогично, «суммарная стандартная неопределенность u_c определяется как стандартная неопределенность результата измерений, полученного через значения других величин, равная положительному квадратному корню суммы членов, причем члены являются дисперсиями или ковариациями этих других величин, взвешенными в соответствии с тем, как результат измерений изменяется при изменении этих величин». Введенные таким образом характеристики рассеяния значений, которые могли бы быть обоснованно приписаны измеряемой величине, не содержат каких-либо предположений о законе распределения приписываемых значений. Однако ситуация оказывается иной при расчете расширенной неопределенности U , являющейся «величиной, определяющей интервал вокруг результата измерения y , в пределах которого, как можно ожидать, находится большая часть распределения значений, которые с достаточным основанием могли бы быть приписаны измеряемой величине».

Для расчета расширенной неопределенности используется соотношение

$$U = ku_c, \quad (1)$$

где k — коэффициент охвата.

Отметим, что с учетом определения u_c и соотношения (1) U — это половина длины интервала, симметричного относительно результата измерения y , в пределах которого с вероятностью p находятся значения, которые с достаточным основанием могли бы быть приписаны измеряемой величине Y .

В [1, 2] отмечено, что для расчета k необходимо знать распределения этих значений. Однако при расчете неопределенности по типу А практически рекомендуется единственный метод расчета k с использованием формулы

$$k = t_p(v_{\text{эфф}}),$$

где $t_p(v_{\text{эфф}})$ — квантиль t -распределения (распределения Стьюдента) с эффективным числом степеней свободы $v_{\text{эфф}}$ и доверительной вероятностью (уровнем доверия) p . В соответствии с определением квантиля в данном случае его расчитывают для вероятности $(1+p)/2$.

Это однозначно предполагает, что распределения значений, приписываемых влияющим величинам, и значений, которые могли бы быть приписаны измеряемой величине, являются нормальными. В [3] отмечается неизбежность ошибок при использовании вероятностно-статистических формул, если не выполнены предположения, при которых они получены. В частности, это относится к предположению о нормальности распределений. Если число влияющих факторов велико, то распределение приписываемых значений можно считать нормальным. Однако распределение оценки дисперсии все равно может существенно отличаться от распределения χ -квадрат, что необходимо при правильном применении распределения Стьюдента.

После дополнительных исследований можно получить информацию, позволяющую обосновать нормальность распределения или уточнить действительное распределение. Не располагая этой информацией, зная только СКО, оценку сверху для k находят на основе неравенства Чебышева [4]. В соответствии с этим неравенством вероятность того, что абсолютное значение отклонения случайной величины ξ с СКО S от математического ожидания μ превысит заданное положительное число A , удовлетворяет условию

$$P[|\xi - \mu| \geq A] \leq (S/A)^2. \quad (2)$$

Тогда с учетом (1) и определения U из (2) получаем

$$k \leq 1/\sqrt{1-p}. \quad (3)$$

Оценка (3) применима для стандартной неопределенности как по типу А, так и по типу В. Она справедлива и для суммарной неопределенности. Соотношение (3) дает существенно завышенное значение k . Например, при $p = 0,95$ $k \leq 4,72$, при $p = 0,99$ $k \leq 10$. Для нормального распределения, к которому стремится t -распределение при неограниченном увеличении числа степеней свободы, имеем существенно меньшие значения: $k = 1,96$ и $k = 2,58$, соответственно. С уменьшением числа степеней свободы k , соответствующий распределению Стьюдента, несколько увеличивается, но все равно остается существенно меньше верхней границы, определяемой (3). Это вполне естественно, поскольку знак равенства в (2) достигается для дискретной случайной величины ξ , распределение плотности вероятности $w_\xi(x)$ которой может быть представлено в виде суммы трех δ -функций

$$w_\xi(x) = p\delta(x - \mu) + 0,5(1 - p)\delta(x - \mu - U) + 0,5(1 - p)\delta(x - \mu + U). \quad (4)$$

Реальные распределения непрерывных случайных величин имеют более гладкую форму, хотя их аналитические выражения могут оказаться более сложными. Для получения информации о действительных распределениях необходимы дополнительные исследования, цель которых — найти обоснованное значение расширенной неопределенности. Проведение этих исследований требует определенных затрат. Однако при отсутствии обоснования вида распределения достоверной является только оценка сверху (3). Поэтому для обеспечения требований к достоверности и повышению точности измерений с целью уменьшения стандартной неопределенности затраты могут быть не меньшими.

Все сказанное выше полностью справедливо для расчета доверительного интервала по известному СКО погрешности. Это является еще одним примером того, что расчет неопределенности и анализ погрешностей несмотря на различие исходных позиций алгоритмически имеют много общего, что многократно отмечается в [1, 2] и отражено в [5].

Другая ситуация при расчете неопределенностей, когда делаются определенные предположения о законе распределения, возникает при анализе стандартной неопределенности по типу В и расчете суммарной неопределенности. В этом случае необходимо знать СКО всех влияющих (входных) величин $X_1, \dots, X_m$, входящих в определение измеряемой величины Y . Однако неопределенности этих величин могут характеризоваться другими показателями, чаще всего длиной полуинтервала a_j . Здесь и ниже в соответствии с [1] $a = (a_+ - a_-)/2$, где a_+ — верхняя граница или верхний предел величины X ; a_- — нижняя граница или нижний предел этой величины.

В этом случае a_j должны быть пересчитаны в СКО S_j величин X_j . Этот пересчет возможен только при определенных предположениях о распределении значений, приписываемых (или принимаемых) X_j . Это распределение должно быть финитным, т. е. оно может быть отличным от нуля только на интервале $x \pm a$.

Нормальное распределение не является финитным. Поэтому для пересчета можно использовать только усеченное нормальное распределение, заданное на интервале $x \pm z_p S$, где z_p — квантиль нормального распределения для доверительной вероятности или уровня доверия p (в зависимости от способа описания величины X). Квантиль тракту-

ется в том же смысле, как и в (2), тогда $S = a/z_p$. Если a получено в результате расчета расширенной неопределенности, то в отчете по расчету должен быть указан уровень доверия p .

Следует отметить, что усеченное нормальное распределение является практически нереализуемой моделью действительного распределения. Поэтому, если для описания погрешности используется усеченное нормальное распределение, то это эквивалентно замене предела абсолютной погрешности на половину длины доверительного интервала. В этом случае p представляет собой доверительную вероятность и следует принять $z_p = 2$ в предположении $p = 0,95$, $z_p = 2,5$ при $p = 0,99$ или $z_p = 3$ при $p = 0,997$.

В [1, 2] в качестве возможных финитных распределений рассматриваются прямоугольное (равномерное), треугольное и трапецидальное распределения. Для симметричного трапецидального распределения справедливо выражение

$$S = a\sqrt{(\beta^2 + 1)/6}, \quad (5)$$

где β — отношение верхнего и нижнего оснований трапеции.

Для финитных распределений при расчете расширенной неопределенности, как и при расчете длины доверительного интервала, отличие p от единицы практически не сказывается на коэффициенте k . Поэтому в дальнейшем для финитных распределений будем предполагать $p = 1$.

Прямоугольное и треугольное распределения являются крайними случаями трапецидального распределения. Поэтому для прямоугольного распределения при $\beta = 1$ в соответствии с (5) имеем $S = a/\sqrt{3} \approx 0,58a$, а для треугольного

распределения при $\beta = 0$ из (5) получаем $S = a/\sqrt{6} \approx 0,41a$. Треугольное распределение ближе к нормальному, чем прямоугольное, и для него значение S/a оказалось более близким к отношению для усеченного нормального распределения.

Очевидно, что трапецидальное распределение не охватывает всего множества практически встречающихся финитных распределений. Например, иногда используется арккосинусное распределение (распределение синусоиды со случайной фазой)

$$w(x) = 1/\sqrt{a^2 - (x - \mu)^2}.$$

Это распределение встречается в радиоизмерениях при рассогласовании каналов. Ближе к нему распределение показателей, которые в процессе регулировки «вгоняются» в интервал допускаемых значений и остаются вблизи его границ. Арккосинусное распределение больше отличается от нормального, чем прямоугольное, и при заданном a оно

имеет большее СКО $S = a/\sqrt{2} \approx 0,71a$.

Для рассмотренных финитных распределений коэффициент пропорциональности при пересчете a в S изменяется от 0,41 (треугольное распределение) до 0,71 (арккосинусное распределение). При этом обоснование того или иного распределения требует дополнительных исследований, которые не всегда возможны, например, если величина a получена из документации. При обработке результатов исследований довольно громоздкой задачей будет определение

параметра β трапецеидального распределения, значение которого изменяет коэффициент пропорциональности почти в полтора раза.

Аппроксимацию реального распределения прямоугольным часто обосновывают тем, что оно имеет максимальную неопределенность в смысле энтропии Шеннона среди всех финитных распределений, заданных на интервале одной и той же длины. Это обстоятельство, безусловно, существенно для теории информации и примыкающей к ней теории связи. Однако для метрологии это не имеет никакого значения. У арккосинусного распределения энтропия меньше, чем у прямоугольного, однако, как было показано выше, оно имеет большую стандартную неопределенность.

Задача пересчета a в S существенно упрощается, если при известном a использовать оценку S сверху

$$S \leq a. \quad (6)$$

Этой оценке соответствует распределение плотности вероятности

$$w_{\xi}(x) = 0,5[\delta(x - \mu - a) + \delta(x - \mu + a)]. \quad (7)$$

Такое распределение, как и (4), имеет дискретная случайная величина, что мало реально. Однако в отличие от (4) можно привести пример, когда имеется распределение, близкое к распределению (7). При наличии в механизме люфта величиной $\pm d$ погрешность из-за гистерезиса, обусловленного этим люфтом, будет с равными вероятностями принимать дискретные значения $\pm d$ или $\pm d/2$ в зависимости от используемого алгоритма.

Оценка (6) завышена для арккосинусного и прямоугольного распределений примерно в полтора раза, для треугольного и усеченного нормального — более существенно, примерно в два с половиной раза. Однако даже такое завышение может быть оправданным за счет исключения дорогостоящих и громоздких исследований по обоснованию вида распределения.

Все рассмотренные выше распределения симметричны относительно математического ожидания. В [1, п. F.2.4.4] говорится о возможной асимметрии распределения при односторонних ограничениях на входную величину и на конкретном примере показана возможность использования экспоненциального распределения, которое равно нулю для отрицательных значений аргумента, а для положительных значений

$$w(x) = \exp(-x/m)/m.$$

Для этого распределения $\mu = m$, $S = m$, $k = -\ln(1 - p)$. В частности, $k = 3,0$ для $p = 0,95$, $k = 4,6$ для $p = 0,99$.

Применение экспоненциального распределения достаточно хорошо обосновано в теориях массового обслуживания, надежности и в некоторых других случаях. Иногда используется одностороннее нормальное распределение, физически менее обоснованное, но более привычное в связи с широким применением двухстороннего. Одностороннее нормальное распределение, как и экспоненциальное, равно нулю для отрицательных значений аргумента, а для положительных значений

$$w(x) = 2 \left[\exp(-x^2/2\sigma^2) \right] / (\sqrt{2\pi}\sigma).$$

Для этого распределения $\mu = 2\sigma/\sqrt{2\pi}$, $S = \sigma\sqrt{1-2/\pi} \approx 0,6\sigma$, $k = z_p/0,6$, где z_p соответствует упомянутому выше значению, рассчитанному для двухстороннего нормального распределения. При $p = 0,95$ имеем $k = 3,3$, а при $p = 0,99$ — $k = 4,2$, что довольно близко значениям, соответствующим экспоненциальному распределению.

Экспоненциальное и одностороннее нормальные распределения не являются финитными. Поэтому они не применимы, если величина X ограничена сверху значением a_+ . В этом случае $a_- = 0$, $a = a_+/2$. Аналогично двухстороннему ограничению для пересчета a_+ в S можно использовать усеченные экспоненциальное или одностороннее нормальное распределения, заданные на интервале $[0; k_p S]$. Тогда $S = a_+/k_p$. Если a получено в результате расчета расширенной неопределенности, то в отчете по расчету должны быть указаны использованное распределение и уровень доверия p .

Из финитных распределений, как и при двухсторонних ограничениях, можно использовать трапецеидальное распределение. Его целесообразно взять несимметричным и считать, что плотность вероятности постоянна на интервале $[0; \beta a_+]$ и линейно убывает до нуля на интервале $[\beta a_+; a_+]$. Для такого распределения

$$\mu = a_+ (1 + \beta + \beta^2)/(3(1 + \beta));$$

$$S = a_+ \sqrt{(\beta^4 + 2\beta^3 + 2\beta^2 + 1)/18} / (1 + \beta).$$

Для равномерного распределения при $\beta = 1$ имеем $S = a_+/\sqrt{12} \approx 0,29a_+$. При $\beta = 0$ приходим к треугольному распределению с вершиной при $x = 0$, для которого $S = a_+/\sqrt{18} \approx 0,24a_+$.

Применение симметричного арккосинусного распределения, бесконечно возрастающего на границах интервала при одностороннем ограничении, по мнению автора, нецелесообразно.

Из рассмотренных примеров следует, что при односторонних ограничениях коэффициенты пересчета верхней границы величины в стандартную неопределенность для разных распределений отличаются в меньшей степени, чем при двухсторонних ограничениях. Однако при этом следует иметь в виду, что несимметричные распределения приводят к математическим ожиданиям, различным для разных распределений при одном и том же a_+ .

При одностороннем ограничении можно записать соотношение, аналогичное (6):

$$S \leq a_+/2, \quad (8)$$

которому соответствует распределение, аналогичное (7):

$$w_{\xi}(x) = 0,5[\delta(x) + \delta(x - a_+)]. \quad (9)$$

Соотношение (8) приводит к завышению стандартной неопределенности в полтора—два раза, что несколько меньше, чем для двухсторонних ограничений.

Таким образом, использование распределений, рекомендуемых в [1, 2] для расчета расширенной неопределенности измерений и для расчета стандартной неопределенности входных величин без должного обоснования их применимости, приводит к недостоверности результатов расчета. Поэтому необходимо или проводить дополнительные

исследования по анализу распределений, или ориентироваться на оценки сверху (2), (6) и (8).

Литература

1. **Руководство** по выражению неопределенности измерения / Пер. с англ., под ред. В. А. Слаева. СПб.: ВНИИМ, 1999.
2. **РМГ 43—2001**. ГСИ. Применение Руководства ИСО по выражению неопределенности измерений.

3. **Рубичев Н. А., Рябцев Г. Г.** Типовые ошибки применения статистических методов обработки измерительной информации и способы их устранения // Метрология. 2012. № 6. С. 3—16.

4. **Вентцель Е. С.** Теория вероятностей. М.: ГИФМЛ, 1958.

5. **РМГ 91—2009**. ГСИ. Совместное использование понятий «погрешность измерения» и «неопределенность измерения». Общие принципы.

Дата принятия 24.04.2013 г.

519.24

Оптимальный выбор количества интервалов дискретизации области изменения одномерной случайной величины при оценивании плотности вероятности

А. В. ЛАПКО, В. А. ЛАПКО

Институт вычислительного моделирования СО РАН, Красноярск, Россия,
e-mail: lapko@icm.krasn.ru

Исследованы асимптотические свойства регрессионной оценки плотности вероятности, синтез которой предполагает декомпозицию статистических данных. На этой основе определена зависимость количества интервалов дискретизации от объема исходной информации.

Ключевые слова: плотность вероятности, декомпозиция исходных данных, непараметрическая статистика, регрессионная оценка, асимптотические свойства, количество интервалов дискретизации.

The asymptotic properties of the regression estimator of the probability density the synthesis of which assumes the decomposition of statistical data are examined. On this basis the dependence of number of sampling intervals from the volume of original data is determined.

Key words: probability density, initial data decomposition, nonparametric statistic, regression estimator, asymptotic properties, number of sampling intervals.

Выбор количества интервалов дискретизации области изменения значений случайной величины является одной из задач математической статистики. Рассмотрим ее решение с позиций оптимизации структуры регрессионной оценки плотности вероятности.

Пусть $x^i, i = \overline{1, n}$ — выборка из n независимых наблюдений одномерной случайной величины x с неизвестной плотностью вероятности $p(x)$, которая допускает разложение в ряд Тейлора.

Для условий выборок большого объема n в [1, 2] предложена методика синтеза регрессионной оценки плотности вероятности, которая основана на декомпозиции исходных статистических данных и последующем анализе характеристик получаемых множеств случайных величин. Развитие результатов этих исследований приведено в [3].

Разобьем область определения $p(x)$ на N непересекающихся интервалов длиной 2β и сформируем множества случайных величин $X^j, j = \overline{1, N}$. В качестве характеристик X^j при-

ем частоту P^j попадания случайной величины x в j -й интервал и его центр z^j . На основе полученной информации со-

ставим статистическую выборку $V_1 = (z^j, p^j = P^j / (2\beta), j = \overline{1, N})$,

где центры z введенных интервалов имеют равномерный закон распределения $p_0(z) = (2\beta N)^{-1}$, а объем полученной выборки N может быть значительно меньше объема n исходных статистических данных.

Приближением по эмпирическим данным искомой плотности вероятности $p(x)$ возьмем непараметрическую оценку условного математического ожидания [3]:

$$\tilde{p}(x) = c^{-1} \sum_{j=1}^N P^j \Phi\left((x - z^j)/c\right), \quad (1)$$

где $\Phi(u)$ — ядерная функция, удовлетворяющая условиям [4, 5]: